Environment-sensitive generalization and exploration strategies

Fien Goetmaeckers*,[1,A], Charley M. Wu[2,3,4,5,B], Tom Verguts[1,C], Senne Braem[1,D]

[1]Department of Experimental Psychology, Ghent University, Henri Dunantlaan 2, 9000 Ghent, Belgium

[2]Human and Machine Cognition Lab, University of Tübingen, Maria-von-Linden-Str. 6, 72076 Tübingen, Germany

[3]Department of Computational Neuroscience, Max Planck Institute for Biological Cybernetics, Max-Planck-Ring 8, 72076 Tübingen, Germany

[4]Center for Cognitive Science, TU Darmstadt, Darmstadt, Germany

[5]Hessian.AI, Darmstadt, Germany

[A]fien.goetmaeckers@ugent.be; [B]charley.wu@tu-darmstadt.de; [C]tom.verguts@ugent.be; [D]senne.braem@ugent.be

Word count: 4977

Abstract

Humans are remarkably efficient in exploring vast decision spaces. However, environments vary in how they are structured, raising the question of whether and how people adapt their generalization and exploration strategies accordingly. People might flexibly adapt their behavior within each environment (i.e., local adaptation), or also learn to associate specific strategies with distinct environments (i.e., meta-learning). Using a spatially correlated bandit paradigm and computational modeling, we examined how humans adapt their search strategy across two environments optimized for different levels of generalization. Across three experiments, participants showed more generalization and more random exploration in environments with stronger spatial correlations. Moreover, they meta-learned to associate different degrees of generalization with different environmental cues. These findings illustrate how humans control and adapt search strategies across diverse decision spaces.

Keywords: value-based decision making, cognitive control, meta-learning, contextual multi-armed bandit

Public significance statement

This study shows that people not only change how they search and make decisions depending on the demands of the environment, but can also learn to store, link, and retrieve these search strategies whenever they return to similar environments. This form of meta-learning reflects a flexible learning process that helps us make more adaptive decisions in real-world, information-rich settings.

Acknowledgments

This work was funded by an FWO fellowship awarded to FG (#11E2523N), an FWO project grant awarded to TV and SB (G010319N), and an ERC Starting grant awarded to SB (European Union's Horizon 2020 research and innovation program, Grant agreement 852570).

1 Introduction

Humans can search in different ways depending on how structured their environment is. For instance, shopping in a department store allows for more targeted search because similar items are grouped together, whereas a disorganized thrift shop may demand more exploration and limited generalization. Many models of decision-making suggest that we can adjust our search strategies to different environments, likely based on a learned cost-benefit trade-off (Cogliati Dezza et al., 2019; Otto et al., 2022; Shenhav et al., 2017; Wu et al., 2022). Beyond such local adaptation, humans may also *meta-learn* – that is, associate effective strategies with recurring environmental cues (Abrahamse et al., 2016; Braem et al., 2024; Braem & Egner, 2018; Chiu & Egner, 2019; Verbeke & Verguts, 2024; Wang, 2021). However, few studies have directly tested whether people actually implement this form of meta-learning, particularly in complex, choice-rich environments (but see Wu, Deffner, et al., 2025), and some found similar search patterns across differently structured environments (Wu, Schulz, Speekenbrink, et al., 2018). In this study, we contrast two environments that require different optimal generalization strategies (based on model simulations), investigate people's environment-specific search behavior, and provide a new test of whether people meta-learn to associate different strategies to different environments.

Most paradigms on value-based decision-making test human decisions using just a few (typically two or three) choice options, largely underappreciating the vast decision spaces typically faced by searching agents (Wu, Meder, et al., 2025). In everyday contexts – such as searching through a large store, people must generalize from limited experience to estimate the potential value of undiscovered options. To study how humans generalize and explore in such vast decision spaces, we can use spatially correlated multi-armed bandits (Borji & Itti, 2013; Reverdy et al., 2019; Wu, Schulz, Speekenbrink, et al., 2018), where nearby options yield correlated rewards. This spatial structure enables generalization, allowing learners to infer which unexplored options are promising when exhaustive exploration of all options is not feasible (*Figure 1*).

Previous research has shown that people can flexibly adapt their learning and control strategies to local environmental statistics, such as adapting their learning rate to volatility (Behrens et al., 2007; Goris et al., 2021; Simoens et al., 2024; Wang et al., 2018) or task control to task switch probabilities (Dreisbach & Haider, 2006; Wen et al., 2023; Xu et al., 2024). These forms of local adaptation reflect on-the-fly adjustments to momentary task demands (Krugel et al., 2009), i.e., local strategy adaptations. In contrast, meta-learning implies forming associations between strategies and higher-order statistics of the environment, enabling the reuse of effective strategies when revisiting an environment (Binz et al., 2024; Simoens et al., 2024).

Here, we examine whether people adapt their generalization and exploration strategies across different environments with different reward correlations in a spatially correlated bandit task. Specifically, we compared within-participant strategy differences in environments with stronger reward correlations (smooth grids) to environments with weaker correlations (rough grids). Across three experiments, we test whether participants (1) adjusted their search and generalization parameters to the structure of each environment (Experiments 1-3), and (2) meta-learned to associate these distinct strategies with different environmental cues (Experiment 3).

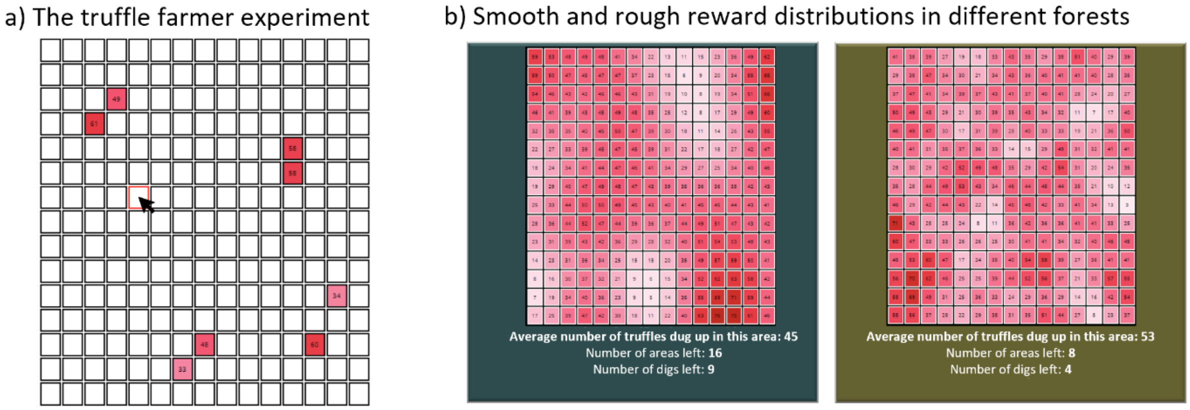
2 Methods

2.1 Participants

Across three online experiments, we recruited 274 participants (Exp 1: N = 90; Exp 2: N = 87; Exp 3: N = 97) from Prolific. To focus on the cognitive mechanisms of healthy, non-aging brains (Salthouse, 2009; Strittmatter et al., 2020), we selected participants between 18 and 35 years old (*M* = 26.6, *SD* = 4.5). Experiments 1, 2, and 3 took 13, 19, and 32 minutes to complete, in return for which participants received a participation fee of £1.50, £4.56, and £5.40 and average reward bonuses of £1.47, £2.29, and £2.23, respectively.

2.2 Behavioral paradigm

Participants conducted the ‘truffle farmer experiment’ (*Figure 1a*; see also, Goetmaeckers et al., 2025) in which they assumed the role of a truffle farmer and guided their truffle pig to the best location in a forest to dig up truffles. The task was an adaptation of the design by Wu et al. (2018). In our version, the goal was to find as many truffles as possible by clicking on locations (cells) in an area (grid of 15 by 15 cells) of a forest (shown in the background, either teal-green or khaki-green; see *Figure 1b*). After selecting a location to dig for truffles, the selected cell showed a reward value and a corresponding color, with darker red indicating higher rewards. The observed rewards were generated following a normal distribution with a standard deviation of 1 to allow for slight variations in the observed rewards of cells that were selected multiple times. The mean of each cell’s normal distribution varied between a minimum of 5 and a maximum randomly generated per round between 65 and 85. The maximum was different per round to avoid participants identifying global optima.


a) The truffle farmer experiment
b) Smooth and rough reward distributions in different forests

Average number of truffles dug up in this area: 45
Number of areas left: 16
Number of digs left: 9

Average number of truffles dug up in this area: 53
Number of areas left: 8
Number of digs left: 4

Figure 1: The behavioral paradigm a) A screenshot of one trial in the truffle experiment. b) Example reward distributions of a smooth (left) and rough (right) grid in their respective environment.

The cell’s mean rewards were spatially correlated, resulting in nearby cells yielding similar rewards. This implied that the location of a cell on the grid was informative of the reward to be expected. These reward distributions over the grids’ cells were created by sampling from a Gaussian process prior parameterized by a radial basis function kernel, which transforms spatial (Euclidian) distance to a reward covariance:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{||\boldsymbol{x}-\boldsymbol{x}'||}{2\,\lambda_{gen}^{2}}\right),$$

with λ_{gen} the length scale of the kernel, parametrizing the rate of the exponential decay of similarity for increasing distance and thereby defining the *smoothness* of the reward distribution. Larger length scales correspond to slower decays, stronger spatial correlations, and smoother reward landscapes. Half of the rounds used *smooth* grids, with length scales $\lambda_{gen} = 16$, and the other half used *rough* grids, with length scales $\lambda_{gen} = 2$ (*Figure 1b*). These values were chosen based on an optimality analysis (see *Appendix A*) showing that the differences in reward correlations differed sufficiently to allow for different optimal generalization strategies. To study environment-sensitive decision-making, grids with the same smoothness λ_{gen} were encountered in the same forest. Participants were told they would travel to two different forests, visiting 8 (Exp 1) or 16 (Exp 2 and 3) different areas in each (i.e., one area per round), and they could dig 30 times per area. Experiments 2 and 3 were double the length of Experiment 1 to study learning effects over rounds. In Experiments 1 and 2, the two different environments were encountered in a block-wise manner of two consecutive blocks of each 8 or 16 rounds, respectively (e.g., AB or AABB). Experiment 2 was a longer version of Experiment 1 to evaluate the robustness of our findings. Experiment 3 differed from the other two experiments in that the environments were experienced in four alternating blocks of 8 rounds each (e.g., ABAB), and ended with additional test rounds at the end of the experiment that allowed us to test meta-learned associations between generalization parameters and environment cues (see *section 2.5*).

2.3 Behavioral exploration measures

We measured exploration using four behavioral measures (see also, Goetmaeckers et al., 2025). First, we counted the number of novel clicks, which we defined as the frequency of selecting a cell that had not been opened before in the round (*Figure 2a*). More novel clicks indicate more exploration, whereas reselecting cells with high rewards indicates exploitation. Thus, we counted the number of high-value clicks, defined as the frequency of selecting an already revealed cell with a visible reward that is at least 90% of the highest current observation in the round (*Figure 2b*). Due to the spatially correlated nature of the rewards, clicking close to a high-value observation can also be considered a less explorative choice than clicking further away. Additionally, travelling longer distances between consecutive clicks can be considered explorative behavior, since the participant actively moves away from exploiting from previous experiences. Therefore, we defined two distance measures, based on the distance between consecutive clicks (*Figure 2c*) and the distance from most nearby high-value cell (*Figure 2d*).

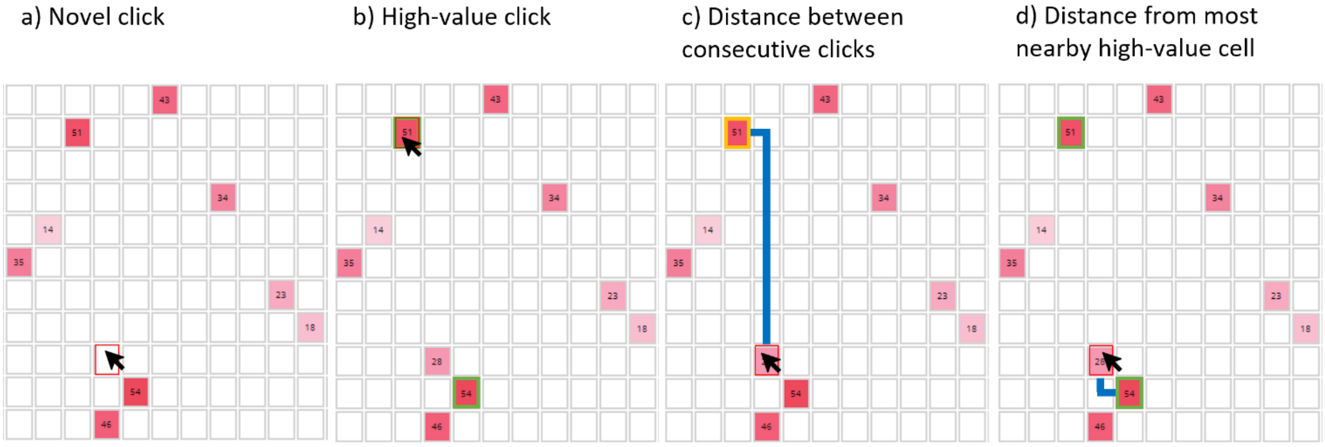

a) Novel click
b) High-value click
c) Distance between consecutive clicks
d) Distance from most nearby high-value cell
43
51
34
14
35
23
18
28
54
46

Figure 2: The four behavioral exploration measures a) Novel clicks are clicks where the participant selects a cell that has not been opened yet in the current round. b) High-value clicks are clicks where the participant selects a cell that has been opened in that round and that has a reward that is at least 90% of the highest observed reward. For this trial, the highest observed reward is 54, so all cells higher than 48.6 (= 90% of 54) are high-value cells. All high-value cells have a green border in this illustration. c) The distance between consecutive clicks is the Manhattan distance between the current and previous (here illustrated as the cell with a yellow border) click. In this illustration, this distance is 8. d) The distance from most nearby high-value cell is 2 in this illustration.

Differences between environments in these four different measures were evaluated using non-parametric Wilcoxon signed rank tests, as our measures did not meet the assumptions for parametric analyses. Finally, we also analyzed the obtained rewards using linear mixed effects models with trial number (within a round), round number (within an environment), and environment (smooth versus rough) as fixed factors, and subject as a random effect, to evaluate whether rewards increased over time (i.e., trials and rounds).

2.4 Computational model

We modeled participants' search and generalization strategies with the Gaussian Process – Upper Confidence Bound (GP-UCB) model (see *Figure 3*), which has been shown to account for a variety of human value-based decision-making behaviours (Giron et al., 2023; Goetmaeckers et al., 2025; Wu, Meder, et al., 2025; Wu, Schulz, Garvert, et al., 2018; Wu, Schulz, Speekenbrink, et al., 2018). The GP provides a Bayesian and kernel-based account of value generalization that bridges classical theories (e.g., Shepard's law of generalization; Shepard, 1987) and modern reinforcement learning mechanisms (e.g., function approximators; Silver et al., 2016). Namely, we use Gaussian Process (GP) Regression, as a Bayesian non-parametric approach to function learning, to make predictive generalizations about the expected reward $m(\boldsymbol{x})$ and uncertainty $s(\boldsymbol{x})$ conditioned on past observations across (unseen) locations **x** (see *Appendix B* for full equations). These predictions are in the form of a normally distributed posterior, where a kernel function defines the covariances. Here, we use the common Radial Basis Function (RBF) kernel,

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

which defines the covariance between cells $\boldsymbol{x}$ and $\boldsymbol{x}'$ as an exponentially decaying function of their Euclidean distance. The length scale parameter λ controls the rate of this decay. We treated λ as a free parameter, where larger values correspond to the participant assuming smooth reward structures and generalizing over larger spatial extents.

In addition to valuing rewards, humans also display an intrinsic motivation to explore uncertain options (Mehlhorn et al., 2015). The GP-UCB model accounts for this by valuing each option $\boldsymbol{x}$ as a weighted sum of the expected reward m($\boldsymbol{x}$) and the uncertainty s($\boldsymbol{x}$);

$$UCB(\boldsymbol{x}) = m(\boldsymbol{x}) + \beta \cdot s(\boldsymbol{x}) \qquad (2)$$

with β quantifying an exploration bonus, determining how strongly the reduction of uncertainty is valued relative to reward expectation. The UCB can be intuitively understood as optimistically inflating the values of expectations with their uncertainty. Since higher values of β lead to sampling in the direction of more uncertain options, this second free parameter regulates the degree to which behavior is guided by an uncertainty-guided exploration strategy.

To test a locality bias, where subsequent choices are likely to stay near one another (Wu, Schulz, Speekenbrink, et al., 2018), we also used a localized version of the model. There, the *UCB* value of a cell $\boldsymbol{x}$ was further weighted by its inverse Manhattan distance from the previous choice $\boldsymbol{x}_{previous}$:

$$UCB_{localized}(\boldsymbol{x}) = \frac{UCB(\boldsymbol{x})}{||\boldsymbol{x}-\boldsymbol{x}_{previous}||},$$

except for the *UCB* value of the previous cell (if $\boldsymbol{x} = \boldsymbol{x}_{previous}$), which remained unaffected. Through model comparison, this alternative variant is compared to the original model.

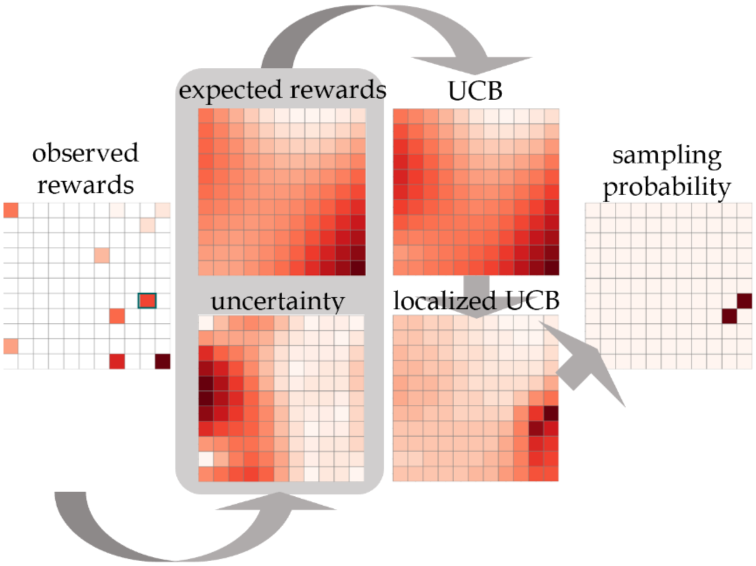


Figure 3: the localized GP-UCB model Through Gaussian Process Regression, the observed rewards generate expected rewards and uncertainty per cell, which are combined in an Upper Confidence Bound (UCB). After adding local bias and SoftMax, the sampling probability guides the next choice. The previous choice is indicated by a blue box around the observed reward.

Finally, to translate the value of each option (*UCB*$UCB_{localized}$) to a sampling probability, a SoftMax choice rule is applied, weighted with a temperature τ;

$$P(\boldsymbol{x}) \sim exp\left(\frac{UCB_{localized}(\boldsymbol{x})}{\tau}\right). \qquad (3)$$

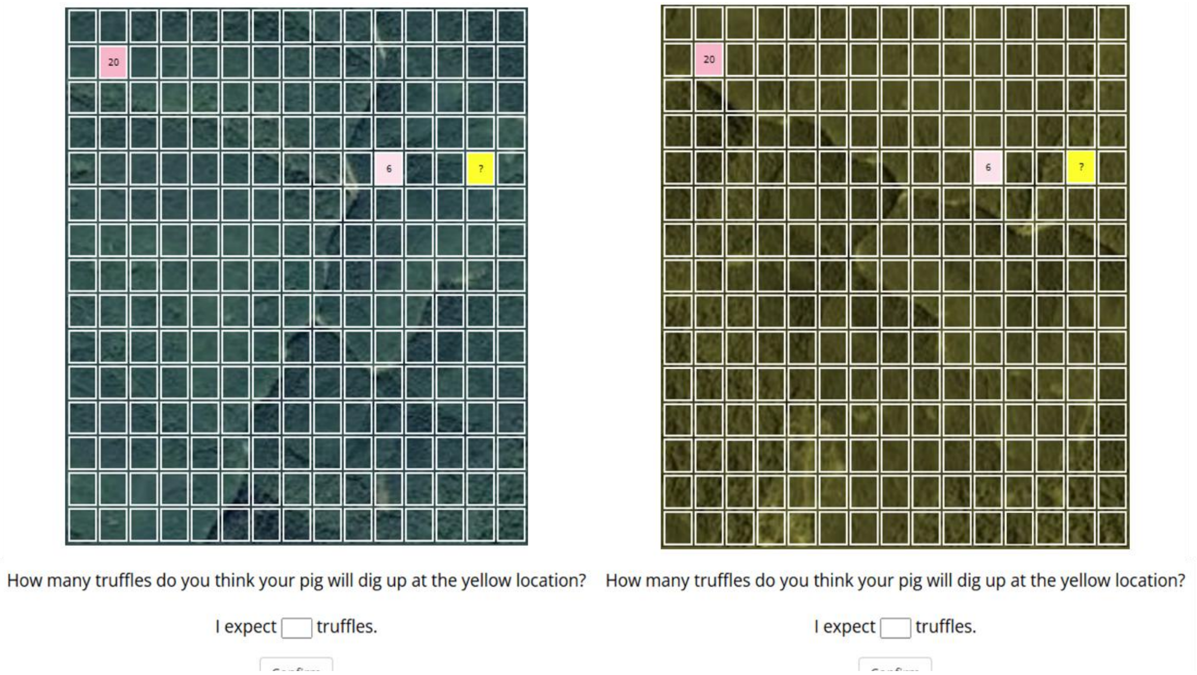
Temperature τ scales the level of noise in sampling, where $\tau \to 0$ refers to purely deterministic sampling. This third free parameter governs a random exploration strategy (in contrast to uncertainty-guided exploration).

We estimated the three model parameters per participant by minimizing the negative log loss (nLL) using leave-one-round-out cross-validation, where we imposed lower and upper bounds (i.e., $\lambda \in [e^{-5}, e^{6}], \beta \in [e^{-5}, e^{3}]$, and $\tau \in [e^{-5}, e^{3}]$).

2.5 Meta-learning test phase

To tease apart the difference between local strategy adaptation (i.e., adapting generalization strategies to the needs of the environment) and meta-learning (i.e., reactivating previously learned strategies based on associated environmental cues), we included an unannounced meta-learning test phase at the end of Experiment 3. Specifically, participants encountered a grid with only two opened cells and were asked to report their reward expectations on a third, highlighted cell (see *Figure 4*). This allowed us to measure how strongly participants generalized from prior observations. Yet, since only two cells were opened, the observed rewards were not informative of the smoothness of the reward distribution. Because of this, participants could not gauge the needs of the grids and potential behavioral differences between the two forests (environmental cues) could not be attributed to local, grid-specific strategy adaptation. Instead, a behavioral difference between environments would prove that participants used the environmental cues to decide which strategy to employ, i.e., meta-learning.

The different meta-learning test phase grids were created using optimality analyses to find maximally diagnostic stimuli where each single-shot decision assessed if different environments triggered different generalization strategies (see *Appendix C*). Participants were presented with 10 grids per environment, in a randomized order. We analyzed these data using three different measures to show differences in meta-learned generalization strategy, which we will introduce in S*ubsection 3.2*.



How many truffles do you think your pig will dig up at the yellow location?

I expect □ truffles.

How many truffles do you think your pig will dig up at the yellow location?

I expect □ truffles.

Figure 4: Meta-learning test phase. At the end of Experiment 3, there was an unannounced meta-learning test phase to test if participants meta-learned associations between the different environmental cues and their generalization strategies. Participants were presented with different trials of minimally revealed grids with just two opened cells and were asked to indicate which reward they expected on a third, highlighted, unopened cell. The background was either the environment (forest) formerly associated with smooth or rough grids, and we investigated whether participants reward expectancy was guided by differentially associated generalization strategies.

3 Results

3.1 Environment-specific strategies

In what follows, we will discuss the results across experiments, as all three experiments were highly similar. Across-experiment analyses also showed that none of our measures (including computational parameters) showed differences across experiments in the effects of environment (all *F* < 1.6, *p* > .191). Results per experiment can be found in *Appendix D*.

Behavioral results

Throughout the experiment, participants gradually improved by opening higher rewards over trials (*b* = 3.74 ± 0.04, *t* = 84.21, *p* < .001) and over rounds (*b* = 0.51 ± 0.05, *t* = 10.58, *p* < .001), with a positive interaction of trial and round (*b* = 0.41 ± 0.04, *t* = 9.25, *p* < .001). These results indicate participants learned from their experiences both within and across rounds.

Next, we examine differences in exploration patterns between smooth and rough environments. We observed a higher number of novel clicks in smooth than in rough grids (Wilcoxon signed rank test: *W* = 13812, *z* = -3.40, *p* < .001; see *Figure 5a*), indicating that participants selected more unique options in the grids with stronger reward correlations. The number of high-value clicks, indicating how often participants reclicked high reward choice options, did not differ between environments (*W* = 19272, *z* = -1.55, *p* = .121; see *Figure 5b*). The distances between consecutive clicks were lower for smooth grids (*W* = 26319, *z* = -5.83, *p* < .001; see *Figure 5c*), as were the distances from most nearby high-value cell (*W* = 32519, *z* = -10.58, *p* < .001; see *Figure 5d*).

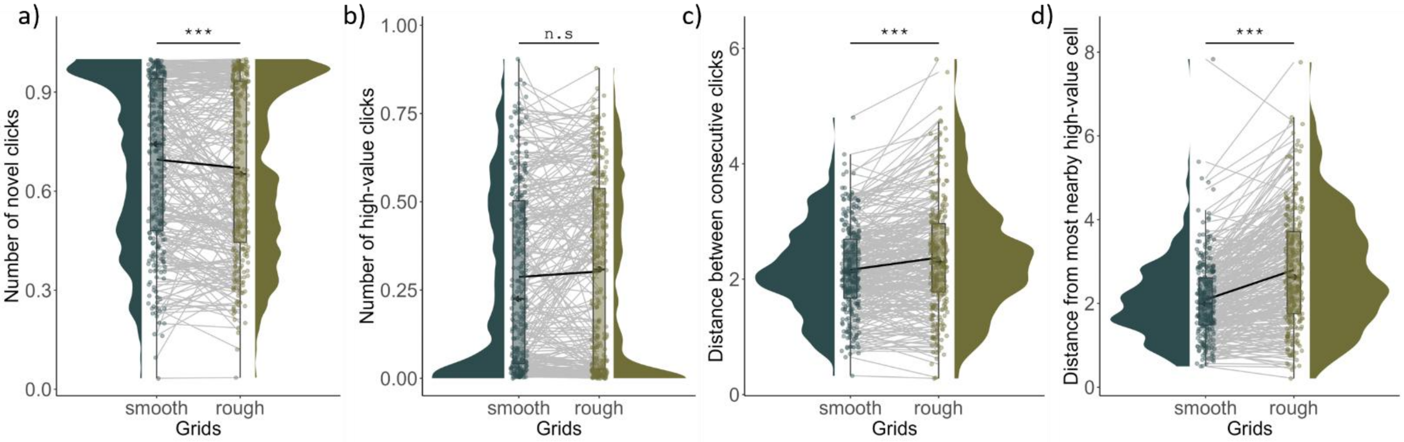


Figure 5: Behavioral results. In smooth grids, we observed more novel clicks (a), lower distances between consecutive clicks (c), and lower distances from most nearby high-value cell (d). The number of high-value clicks was similar between both grids (b). *** *p* value Wilcoxon signed-rank test < .001, n.s. *p* value Wilcoxon signed-rank test > .05.

Computational results

To better try and understand the underlying mental processes driving people to explore, we estimated the model parameters for generalization, uncertainty-guided, and random exploration using the GP-UCB model. The model fits for the localized version of the model (Exp 1: R = 0.376; Exp 2: R = 0.434; Exp 3: R = 0.397) surpassed the original (Exp 1: R = 0.217; Exp 2: R = 0.320; Exp 3: R = 0.312). For that reason, we only discuss the parameter estimates for the localized version of the model.

As expected, we observed higher values for generalization in smooth grids (W = 8645, z = -7.76, p < .001; see *Figure 6a*). This highlights that people (correctly) generalized previous observations over longer distances in grids with stronger spatial correlations between the rewards. Participants' generalization parameters show a bimodal pattern, to which we will return in *section 3.3*. We did not find evidence for different strategies related to uncertainty-guided exploration. (W = 15357, z = -1.52, p = .13; see *Figure 6b*). Additionally, we also saw more random exploration in smooth grids (W = 9662, z = -6.99, p < .001; see *Figure 6c*). This suggests that people sampled less deterministically in grids with stronger reward correlations. Together, these results indicate that the discussed behavioral differences are driven by different generalization and random exploration strategies, but not by a different way in which participants try to reduce the uncertainty of the grid.

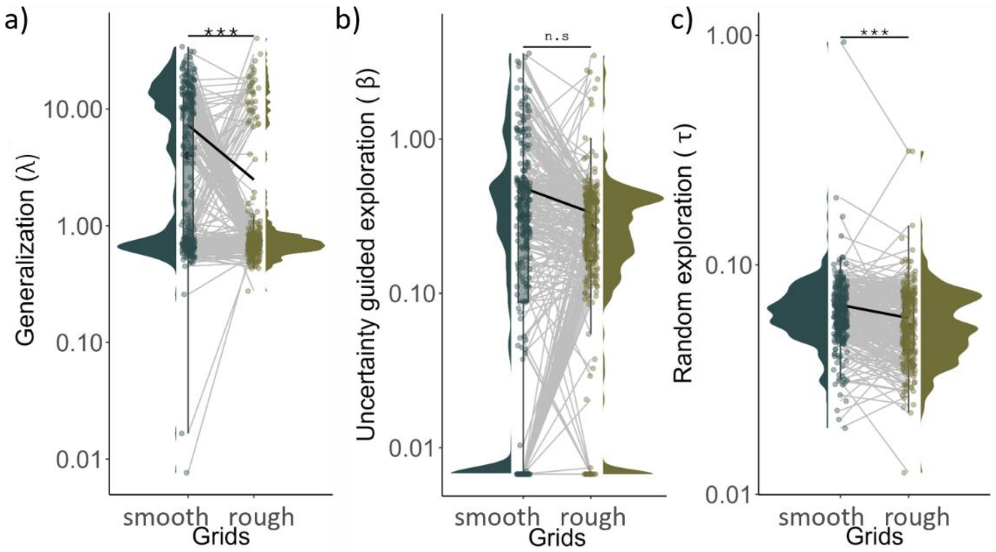


Figure 6: Computational results. In smooth grids, we observed higher values for the generalization (a) and the random exploration (c) parameters. The uncertainty-guided exploration parameter was similar across grids (b). *** *p* value Wilcoxon signed-rank test < .001, n.s. *p* value Wilcoxon signed-rank test > .05.

Interpreting environment-specific strategies

As an interim summary of these first results, we conclude that participants adapted to the stronger spatial reward correlations in the smooth grids compared to the rough grid, by showing more generalization in the smooth environments. The more participants generalize, the more they can use previous rewards to generate expectations about nearby, not yet opened cells. This is consistent with our observation that people also showed an increased choice for novel (unopened) cells and more local sampling (lower distances between clicks and from most nearby high-value cells) in the smooth grids. Concretely, stronger generalization led to the opening of more novel cells but also encouraged more targeted decisions within a smaller spatial range (i.e., local exploration).

These analyses focused on participants' behavior within these different environments with different spatial reward correlations (i.e., one forest always offered smooth grids, and another offered rough grids). While they allow us to conclude that participants adapt different strategies to the statistical structure of the environment, they cannot distinguish between local, on-the-fly strategy adaptation in response to current task demands, versus a meta-learning of associations between these different environments and parameter settings. That is, the observed behavior could still reflect either flexible, grid-specific adjustments or more meta-learned environment-specific strategies. To evaluate whether participants also did the latter, we turned to the data of the meta-learning test phase.

3.2 Meta-learned environment-sensitive generalization strategies

To test whether participants also meta-learned environment-parameter associations, we included a meta-learning test phase in Experiment 3 (see *section 2.5*). Participants were presented with grids in either environment that had only two opened cells and were asked to report their reward expectations on a third, highlighted cell. We performed three different analyses to investigate whether participants generated different expectations based on the environmental cue:

First, we simply evaluated the similarity between the observed reward of the closest open cell and responded reward expectation. A more similar reward expectancy should suggest stronger generalization. Indeed, we observed that the reward expectations were more similar to the most nearby opened cell in rounds with smooth rather than rough environmental cues (b = 4.86 ± 1.90, t = 2.56, p = .011), indicating stronger generalization with smooth environmental cues. Furthermore, this similarity should decrease as the highlighted cell becomes further away from the open cells. Effectively, reward expectancies were more dissimilar with increasing distances (b = 1.06 ± 0.14, t = 7.58, p < .001). Finally, stronger generalization should imply a steeper effect of distance. Indeed, the effect of distance interacted with the environmental cue (b = -0.52 ± 0.21, t = -2.53, p = .011; see *Figure 7a*), showing stronger spatial generalization with smooth environments.

Second, depending on the scenario, stronger generalization could lead to either smaller (as in the scenario in *Figure 4*) or larger reward expectations, depending on whether the most nearby cell is lower or larger than the further away cell. We used simulations from smooth and rough generalization strategies (of which we used the estimated generalization parameters of Experiments 1 and 2) to group scenarios into those where stronger generalization should result in lower versus higher expectations. Next, we operationalized participants' generalization strength by subtracting average expectations in the lower from the higher expectation scenarios. The bigger this difference score, the more participants take the revealed cells' rewards into account, so the stronger their generalization strength. We observed this generalization strength to be higher in grids shown in a smooth environment than in a rough environment (W = 4182, z = -6.22, p < .001; see *Figure 7b*), indicating that participants indeed used environmental cues to flexibly adapt their generalization strategy.

Finally, as a last measure of meta-learned generalization, we also estimated each individual participant's smooth and rough generalization strength based on their performance on the first phase of Experiment 3. Next, we used these estimated model parameters to predict their responses in the meta-learning test phase. Through simulations, we predicted what their expectations should be if they generalized in the same way as they did in smooth and rough training rounds (respectively *smooth generalization expectation* and *rough generalization expectation*). To assess whether their responded expectations were indeed more similar to their smooth generalization expectation (than their rough generalization expectation) if the scenarios were shown with smooth environmental cues (and the other way around for rough environmental cues), we used the following linear mixed effects regression:

*responded expectation ~ (smooth generalization expectation + rough generalization expectation) * environment + (1|subject ID).*

We observed an interaction effect between smooth generalization expectation and environment, where the linear relationship between the responded expectation and smooth generalization expectation was stronger with smooth environmental cues (b = -0.10 ± 0.03, t = -4.06, p < .001; see *Figure 7c*). In other words, the responded expectations followed the predicted model-based smooth generalization expectations more when participants encountered smooth environmental cues. This suggests that the smooth generalization strategy is actively switched on (off) by smooth (rough) environmental cues. We did not see this interaction effect for the rough generalization expectations (b = -0.03 ± 0.04, t = -0.62, p = .539), which could be due to the more limited adaptive value of participants' rough generalization strategies.

Together, these results prove that participants' generalization strategies were actively influenced by the environment in which they are. Since the meta-learning test phase grids did not differ in their statistical structure, this strategy difference could not be attributed to a local grid-specific adaptation. Rather, it proves that participants meta-learned to associate environment-sensitive generalization strategies with the environment they were in.

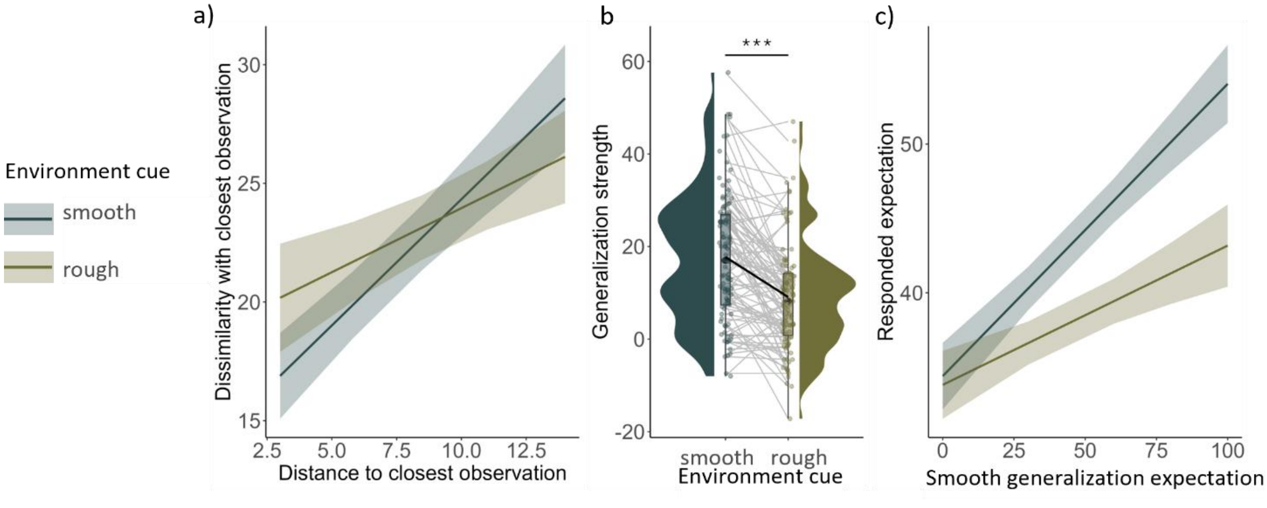


Figure 7: Measures showing meta-learned environment-sensitive generalization in the meta-learning test phase. At the end of Experiment 3, we tested whether participants generalized differently from observed rewards based on the environment they were in. When the observations were presented together with smooth environment cues, a) participants were more affected by spatial distances, b) showed higher generalization strength, and their responses were c) more similar to predicted smooth generalization expectations.

3.3 Individual differences in environment-sensitive generalization strategies

Although we observed a clear positive adaptation of generalization at the group level (i.e., stronger generalization in smooth grids than in rough grids), not all participants showed this effect. In fact, a bimodal distribution appeared visible in the estimated generalization parameters (see *Figure 6a*). Indeed, a Hartigan's dip test for multimodality refuted the unimodality of the generalization parameters distribution (*D* = 0.06, *p* < .001). Interestingly, these clear individual differences appeared to be largely independent of other adaptations, as there were no correlations with the environment-specific differences in uncertainty-guided or random exploration (all *p* > .2, see *Appendix Figure 4*). This suggests that other individual strategy adaptations in both forms of exploration occurred independently. However, these results could also reflect limitations in the reliability of these parameters, as split-half reliability scores between the parameters estimates on odd versus even rounds (Spearman-Brown-corrected Spearman's rho rank-order correlations) showed ρ > .28, *p* < .006 for λ, ρ > .64, *p* < .001 for β, and ρ > .68, *p* < .001 for τ, for smooth or rough environments in Experiments 2 or 3 (where we had sufficient rounds to perform these analyses).

We also evaluated whether the observed environment-specific adaptations led to better reward accumulation (see *Appendix Figure 5*). While the observed trends of more novel clicks and higher random exploration in smooth grids were indeed linked to higher rewards (*r*(272) = .23, *t* = 3.85, *p* < .001 and *r*(272) = .29, *t* = 4.93, *p* < .001, respectively), we also observed that the lower sampling distances and stronger generalization observed in smooth grids actually correlated with lower rewards (distance between consecutive clicks: *r*(272) = .17, *t* = 2.79, *p* = .006; distance from most nearby high-value click: *r*(272) = .36, *t* = 6.32, *p* < .001; generalization: *r*(272) = -.21, *t* = -3.58, *p* < .001). We will return to this observation in the discussion.

4 Discussion

This study investigated whether humans flexibly adjust search strategies based on different environmental needs and meta-learned associations to environmental cues. Across three experiments, we observed different generalization and exploration strategies between two environments with different reward structures that either allowed for more versus less generalization. Our behavioral analyses indicated a strategy adaptation featuring more novel clicks but also more local sampling in smooth environments, which can be summarized as more local exploration. Computational modeling further revealed that participants showed stronger generalization, and engaged in more random exploration, in smooth grids. These findings suggest that when faced with more strongly correlated environments, people generalize more and shift to an exploration strategy governed by less deterministic sampling, but increased novel and local option selecting — adapting their search strategies based on environmental needs. Finally, and critically, we demonstrated that people did not just adjust their generalization strategies to these local environment-specific task demands, but also meta-learned to associate these different generalization strategies to different environmental cues.

Previous work has demonstrated that people can make adaptive changes in decision-making model parameters, such as adjustments of learning rates (Behrens et al., 2007; Simoens et al., 2024; Wang et al., 2018) or caution parameters (Cavanagh et al., 2011; Dunovan & Verstynen, 2019; Held et al., 2024; Ratcliff & Frank, 2012). In contrast, there is only limited evidence that people can also adapt exploration and generalization parameters to environmental demands in large decision spaces. Some studies suggest that exploration strategies may become more prominent in volatile or sparse environments (Gershman, 2018; Wilson et al., 2014), but empirical support for within-participant adaptation of these strategies remained scarce. Similarly, there is limited empirical evidence that people adjust generalization strategies in response to changes in environmental structure, despite theoretical motivation (Schulz et al., 2018; but see Wu, Deffner, et al., 2025). Our findings provide such direct evidence, showing that both generalization and exploration are not fixed but adapt flexibly within individuals, depending on the structure of the environment.

We further demonstrate how these adaptations can reflect both local strategy adaptations (i.e., on-the-fly responses to the local differences in the reward correlations) or meta-learning (i.e., learned environment-sensitive strategies that are flexibly activated when revisiting an environment). Using our meta-learning test phase, we showed that participants exhibited more generalization in environments formerly associated with smooth grids. Our findings indicate that people meta-learn to associate environmental cues to specific generalization strategies, and can flexibly switch between these strategies based on the environment they are in. This learned flexibility aligns with research on meta-learning and cognitive control, which suggests that humans not only learn which actions to take but also how to adjust their decision-making strategies based on context (Botvinick et al., 2019; Collins & Frank, 2013; Nussenbaum & Hartley, 2024; Simoens et al., 2024, 2025; Verbeke & Verguts, 2024; Wang, 2021; Xu et al., 2024). This pattern resonates with evidence from cognitive neuroscience suggesting that the prefrontal cortex maintains hierarchical control policies along its rostro-caudal axis (Badre & Nee, 2018; Fuster, 2001; Hunt & Hayden, 2017; Koechlin et al., 2003; Nee & D'Esposito, 2016). Meta-learned, higher-order representations that encode "how to learn" within a given environment may be represented in more rostral regions, such as the orbitofrontal cortex (Fine & Hayden, 2021; Hattori et al., 2023; Moneta et al., 2024; Simoens et al., 2025), supporting flexible adjustment of generalization and exploration strategies.

Notably, our results suggest this meta-learning can develop quite quickly (here, during a short experiment). This contrasts with other theories and findings (Botvinick et al., 2019; Braem et al., 2024; Xu et al., 2024), who suggested the learning of environment-specific control strategies often develops more gradually and slowly (over multiple sessions and/or days). Future research should compare the different learning mechanisms and trajectories for meta-learned environment-sensitive generalization, as studied here, as opposed to the meta-learning of other decision-making parameters such as learning rates (Simoens et al., 2024, 2025). Another exciting research question is whether environment-specific parameter settings can transfer across different task domains or other related environments (Verbeke et al., 2025), providing insights into the generalizability of meta-learning.

Although we observed distinct within-participant strategy adaptations at the group level, individual differences suggest that not all participants adapted their generalization strategies. Interestingly, some of the observed environment-specific strategy adaptations were even associated with lower reward accumulation. These correlations are difficult to interpret, given the complex interplay between exploration and generalization strategies. One possibility is that people who eventually discovered more adaptive generalization strategies may have initially explored a broader range of strategies, temporarily reducing reward (e.g., (Masís et al., 2023; Masís Obando et al., 2025). We also observed a bimodal distribution of generalization parameters in the smooth environment, which may point to deeper differences in how participants represented the task's structure. One group of participants may have clustered both environments into a single latent context, applying a common generalization strategy across them. In contrast, the other group may have inferred distinct latent states for each environment and adjusted their generalization accordingly. This interpretation aligns with theories of latent state inference in reinforcement learning, which propose that individuals segment their experiences based on perceived contextual structure (Anderson, 1991; Collins & Frank, 2013).

Beyond its theoretical importance for understanding how people adapt their decision-making strategies to the structure of their environment, our gamified design (Allen et al., 2024) can be expanded for developmental (Gopnik, 2020; Nussenbaum & Hartley, 2019) and aging research (Samanez-Larkin & Knutson, 2014), as well as for clinical populations (Browning et al., 2015; Goetmaeckers et al., 2025; Goris et al., 2021; Maia & Frank, 2011). Additionally, because the task formalizes adaptive exploration in structured decision spaces, it may also act as a link between human cognition and machine learning approaches that model meta-learning and hierarchical adaptation (Lake et al., 2017; Wang, 2021).

Our findings extend previous research on generalization and exploration in vast decision spaces (Giron et al., 2023; Wu, Schulz, Speekenbrink, et al., 2018) by confirming that humans can actively modulate their generalization and exploration strategies based on environmental cues. This work complements previous research on environment-sensitive model parameter adaptation (Simoens et al., 2025; Xu et al., 2024) and theories of associative learning of cognitive control (Abrahamse et al., 2016; Braem & Egner, 2018). Together, these results emphasize the flexibility of human decision-making and the human capacity to meta-learn environment-specific strategies in complex environments.

Declarations

Ethics

The study was approved by the Ethics Committee of the Faculty of Psychological and Pedagogical Sciences of Ghent University and all participants gave their informed consent.

Availability of data and materials

The dataset supporting the conclusions of this article is available on GitHub; https://github.com/FienGoetmaeckers/Environment-sensitive/tree/main/Data.

All relevant code is publicly available at https://github.com/FienGoetmaeckers/Environment-sensitive.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was funded by an FWO fellowship awarded to FG (#11E2523N), an FWO project grant awarded to TV and SB (G010319N), and an ERC Starting grant awarded to SB (European Union's Horizon 2020 research and innovation program, Grant agreement 852570).

Authors’ contributions

SB, TV, CW, and FG conceptualized the study, FG conducted the data collection. FG analyzed the data and wrote the first version of the manuscript. SB, TV, and CW supervised the project. All authors contributed to developing the methodology and reviewing and editing the manuscript.

References

Abrahamse, E., Braem, S., Notebaert, W., & Verguts, T. (2016). Grounding cognitive control in associative learning. *Psychological Bulletin*, *142*(7), 693–728. https://doi.org/10.1037/bul0000047

Allen, K., Brändle, F., Botvinick, M., Fan, J. E., Gershman, S. J., Gopnik, A., Griffiths, T. L., Hartshorne, J. K., Hauser, T. U., Ho, M. K., de Leeuw, J. R., Ma, W. J., Murayama, K., Nelson, J. D., van Opheusden, B., Pouncy, T., Rafner, J., Rahwan, I., Rutledge, R. B., … Schulz, E. (2024). Using games to understand the mind. *Nature Human Behaviour*, *8*(6), 1035–1043. https://doi.org/10.1038/s41562-024-01878-9

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Badre, D., & Nee, D. E. (2018). Frontal Cortex and the Hierarchical Control of Behavior. *Trends in Cognitive Sciences*, *22*(2), 170–188. https://doi.org/10.1016/j.tics.2017.11.005

Behrens, T. E. J., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. S. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience*, *10*(9), Article 9. https://doi.org/10.1038/nn1954

Binz, M., Dasgupta, I., Jagadish, A. K., Botvinick, M., Wang, J. X., & Schulz, E. (2024). Meta-learned models of cognition. *Behavioral and Brain Sciences*, *47*, e147. https://doi.org/10.1017/S0140525X23003266

Borji, A., & Itti, L. (2013). Bayesian optimization explains human active search. *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1*, *1*, 55–63.

Botvinick, M., Ritter, S., Wang, J. X., Kurth-Nelson, Z., Blundell, C., & Hassabis, D. (2019). Reinforcement Learning, Fast and Slow. *Trends in Cognitive Sciences*, *23*(5), 408–422. https://doi.org/10.1016/j.tics.2019.02.006

Braem, S., Chai, M., Held, L. K., & Xu, S. (2024). One cannot simply 'be flexible': Regulating control parameters requires learning. *Current Opinion in Behavioral Sciences*, *55*, 101347. https://doi.org/10.1016/j.cobeha.2023.101347

Braem, S., & Egner, T. (2018). Getting a Grip on Cognitive Flexibility. *Current Directions in Psychological*

Science, *27*(6), 470–476. https://doi.org/10.1177/0963721418787475

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, *18*(4), 590–596. https://doi.org/10.1038/nn.3961

Cavanagh, J. F., Wiecki, T. V., Cohen, M. X., Figueroa, C. M., Samanta, J., Sherman, S. J., & Frank, M. J. (2011). Subthalamic nucleus stimulation reverses mediofrontal influence over decision threshold. *Nature Neuroscience*, *14*(11), 1462–1467. https://doi.org/10.1038/nn.2925

Chiu, Y.-C., & Egner, T. (2019). Cortical and subcortical contributions to context-control learning. *Neuroscience and Biobehavioral Reviews*, *99*, 33–41. https://doi.org/10.1016/j.neubiorev.2019.01.019

Cogliati Dezza, I., Cleeremans, A., & Alexander, W. (2019). Should we control? The interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *Journal of Experimental Psychology: General*, *148*(6), 977–993. https://doi.org/10.1037/xge0000546

Collins, A. G. E., & Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, *120*, 190–229. https://doi.org/10.1037/a0030852

Dreisbach, G., & Haider, H. (2006). Preparatory adjustment of cognitive control in the task switching paradigm. *Psychonomic Bulletin & Review*, *13*(2), 334–338. https://doi.org/10.3758/BF03193853

Dunovan, K., & Verstynen, T. (2019). Errors in Action Timing and Inhibition Facilitate Learning by Tuning Distinct Mechanisms in the Underlying Decision Process. *The Journal of Neuroscience*, *39*(12), 2251–2264. https://doi.org/10.1523/JNEUROSCI.1924-18.2019

Fine, J. M., & Hayden, B. Y. (2021). The whole prefrontal cortex is premotor cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1844), 20200524. https://doi.org/10.1098/rstb.2020.0524

Fuster, J. M. (2001). The prefrontal cortex--an update: Time is of the essence. *Neuron*, *30*(2), 319–333. https://doi.org/10.1016/s0896-6273(01)00285-9

Gershman, S. J. (2018). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.

https://doi.org/10.1016/j.cognition.2017.12.014

Giron, A. P., Ciranka, S., Schulz, E., van den Bos, W., Ruggeri, A., Meder, B., & Wu, C. M. (2023). Developmental changes in exploration resemble stochastic optimization. *Nature Human Behaviour*, *7*(11), 1955–1967. https://doi.org/10.1038/s41562-023-01662-1

Goetmaeckers, F., Goris, J., Wiersema, J. R., Verguts, T., & Braem, S. (2025). Different exploration strategies along the autism spectrum: Diverging effects of autism diagnosis and autism traits. *Molecular Autism*, *16*(1), 47. https://doi.org/10.1186/s13229-025-00679-9

Goetmaeckers, F. (2025, November 24). *Environment-sensitive: All scripts for the project "Environment-sensitive generalization and exploration strategies"* [Source code]. GitHub. https://github.com/FienGoetmaeckers/Environment-sensitive

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *375*(1803), 20190502. https://doi.org/10.1098/rstb.2019.0502

Goris, J., Silvetti, M., Verguts, T., Wiersema, R., Brass, M., & Braem, S. (2021). Autistic traits are related to worse performance in a volatile reward learning task despite adaptive learning rates. *AUTISM*, *25*(2), Article 2. https://doi.org/10.1177/1362361320962237

Hattori, R., Hedrick, N. G., Jain, A., Chen, S., You, H., Hattori, M., Choi, J.-H., Lim, B. K., Yasuda, R., & Komiyama, T. (2023). Meta-reinforcement learning via orbitofrontal cortex. *Nature Neuroscience*, *26*(12), 2182–2191. https://doi.org/10.1038/s41593-023-01485-3

Held, L. K., Vermeylen, L., Dignath, D., Notebaert, W., Krebs, R. M., & Braem, S. (2024). Reinforcement learning of adaptive control strategies. *Communications Psychology*, *2*(1), 1–13. https://doi.org/10.1038/s44271-024-00055-y

Hunt, L. T., & Hayden, B. Y. (2017). A distributed, hierarchical and recurrent framework for reward-based choice. *Nature Reviews Neuroscience*, *18*(3), Article 3. https://doi.org/10.1038/nrn.2017.7

Koechlin, E., Ody, C., & Kounelher, F. (2003). The Architecture of Cognitive Control in the Human Prefrontal Cortex. *Science*, *302*(5648), 1181–1185. https://doi.org/10.1126/science.1088545

Krugel, L. K., Biele, G., Mohr, P. N. C., Li, S.-C., & Heekeren, H. R. (2009). Genetic variation in dopaminergic neuromodulation influences the ability to rapidly and flexibly adapt decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(42), 17951–17956. https://doi.org/10.1073/pnas.0905191106

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *The Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Maia, T. V., & Frank, M. J. (2011). From reinforcement learning models to psychiatric and neurological disorders. *Nature Neuroscience*, *14*(2), 154–162. https://doi.org/10.1038/nn.2723

Masís, J., Chapman, T., Rhee, J. Y., Cox, D. D., & Saxe, A. M. (2023). Strategically managing learning during perceptual decision making. *eLife*, *12*, e64978. https://doi.org/10.7554/eLife.64978

Masís, J., Musslick, S., & Cohen, J. D. (2023). Learning expectations shape cognitive control allocation. *PNAS*, *122*, e2416720122. https://doi.org/10.1073/pnas.2416720122

Mehlhorn, K., Newell, B. R., Todd, P. M., Lee, M. D., Morgan, K., Braithwaite, V. A., Hausmann, D., Fiedler, K., & Gonzalez, C. (2015). Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decision*, *2*(3), 191–215. https://doi.org/10.1037/dec0000033

Moneta, N., Grossman, S., & Schuck, N. W. (2024). Representational spaces in orbitofrontal and ventromedial prefrontal cortex: Task states, values, and beyond. *Trends in Neurosciences*, *47*(12), 1055–1069. https://doi.org/10.1016/j.tins.2024.10.005

Nee, D. E., & D'Esposito, M. (2016). The hierarchical organization of the lateral prefrontal cortex. *eLife*, *5*, e12112. https://doi.org/10.7554/eLife.12112

Nussenbaum, K., & Hartley, C. A. (2019). Reinforcement learning across development: What insights can we draw from a decade of research? *Developmental Cognitive Neuroscience*, *40*, 100733. https://doi.org/10.1016/j.dcn.2019.100733

Nussenbaum, K., & Hartley, C. A. (2024). Understanding the development of reward learning through the lens of meta-learning. *Nature Reviews Psychology*, *3*(6), 424–438. https://doi.org/10.1038/s44159-024-00304-1

Otto, A. R., Braem, S., Silvetti, M., & Vassena, E. (2022). Is the juice worth the squeeze? Learning the marginal value of mental effort over time. *Journal of Experimental Psychology: General*, *151*, 2324–2341. https://doi.org/10.1037/xge0001208

Ratcliff, R., & Frank, M. J. (2012). Reinforcement-based decision making in corticostriatal circuits: Mutual constraints by neurocomputational and diffusion models. *Neural Computation*, *24*(5), 1186–1229. https://doi.org/10.1162/NECO_a_00270

Reverdy, P., Srivastava, V., & Leonard, N. E. (2019). *Modeling Human Decision-making in Generalized Gaussian Multi-armed Bandits* (No. arXiv:1307.6134). arXiv. https://doi.org/10.48550/arXiv.1307.6134

Salthouse, T. A. (2009). When does age-related cognitive decline begin? *Neurobiology of Aging*, *30*(4), 507–514. https://doi.org/10.1016/j.neurobiolaging.2008.09.023

Samanez-Larkin, G. R., & Knutson, B. (2014). Reward processing and risky decision making in the aging brain. In *The neuroscience of risky decision making* (pp. 123–142). American Psychological Association. https://doi.org/10.1037/14322-006

Schulz, E., Speekenbrink, M., & Krause, A. (2018). A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, *85*, 1–16. https://doi.org/10.1016/j.jmp.2018.03.001

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (2019). Searching for Rewards Like a Child Means Less Generalization and More Directed Exploration. *Psychological Science*, *30*(11), 1561–1572. https://doi.org/10.1177/0956797619863663

Shenhav, A., Musslick, S., Lieder, F., Kool, W., Griffiths, T. L., Cohen, J. D., & Botvinick, M. M. (2017). Toward a Rational and Mechanistic Account of Mental Effort. *Annual Review of Neuroscience*, *40*(1), 99–124. https://doi.org/10.1146/annurev-neuro-072116-031526

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science*, *237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961

Simoens, J., Braem, S., Verbeke, P., Chen, H., Mattioni, S., Chai, M., Schuck, N. W., & Verguts, T. (2025). Two time scales of adaptation in human learning rates. *eLife*, *14*. https://doi.org/10.7554/eLife.108223.1

Simoens, J., Verguts, T., & Braem, S. (2024). Learning environment-specific learning rates. *PLOS Computational Biology*, *20*(3), e1011978. https://doi.org/10.1371/journal.pcbi.1011978

Strittmatter, A., Sunde, U., & Zegners, D. (2020). Life cycle patterns of cognitive performance over the long run. *Proceedings of the National Academy of Sciences*, *117*(44), 27255–27261. https://doi.org/10.1073/pnas.2006653117

Verbeke, P., De Walsche, M., Maelfait, P., & Verguts, T. (2025). Between structure and flexibility: Testing the limits of human generalization. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001825

Verbeke, P., & Verguts, T. (2024). Humans adaptively select different computational strategies in different learning environments. *Psychological Review*. https://doi.org/10.1037/rev0000474

Wang, J. X. (2021). Meta-learning in natural and artificial intelligence. *Current Opinion in Behavioral Sciences*, *38*, 90–95. https://doi.org/10.1016/j.cobeha.2021.01.002

Wang, J. X., Kurth-Nelson, Z., Kumaran, D., Tirumala, D., Soyer, H., Leibo, J. Z., Hassabis, D., & Botvinick, M. (2018). Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, *21*(6), 860–868. https://doi.org/10.1038/s41593-018-0147-8

Wen, T., Geddert, R. M., Madlon-Kay, S., & Egner, T. (2023). Transfer of Learned Cognitive Flexibility to Novel Stimuli and Task Sets. *Psychological Science*, *34*(4), 435–454.

https://doi.org/10.1177/09567976221141854

Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A., & Cohen, J. D. (2014). Humans Use Directed and Random Exploration to Solve the Explore–Exploit Dilemma. *Journal of Experimental Psychology. General*, *143*(6), 2074–2081. https://doi.org/10.1037/a0038199

Wu, C. M., Deffner, D., Kahl, B., Meder, B., Ho, M. K., & Kurvers, R. H. J. M. (2025). Adaptive mechanisms of social and asocial learning in immersive collective foraging. *Nature Communications*, *16*(1), 3539. https://doi.org/10.1038/s41467-025-58365-6

Wu, C. M., Meder, B., & Schulz, E. (2025). Unifying Principles of Generalization: Past, Present, and Future. *Annual Review of Psychology*, *76*(Volume 76, 2025), 275–302. https://doi.org/10.1146/annurev-psych-021524-110810

Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2018). *Connecting conceptual and spatial search via a model of generalization* (p. 258665). https://doi.org/10.1101/258665

Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2020). Similarities and differences in spatial and non-spatial cognitive maps. *PLOS Computational Biology*, *16*(9), e1008149. https://doi.org/10.1371/journal.pcbi.1008149

Wu, C. M., Schulz, E., & Gershman, S. J. (2021). Inference and Search on Graph-Structured Spaces. *Computational Brain & Behavior*, *4*(2), 125–147. https://doi.org/10.1007/s42113-020-00091-x

Wu, C. M., Schulz, E., Pleskac, T. J., & Speekenbrink, M. (2022). Time pressure changes how people explore and respond to uncertainty. *Scientific Reports*, *12*(1), 4122. https://doi.org/10.1038/s41598-022-07901-1

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924. https://doi.org/10.1038/s41562-018-0467-4

Xu, S., Simoens, J., Verguts, T., & Braem, S. (2024). Learning where to be flexible: Using environmental cues to regulate cognitive control. *Journal of Experimental Psychology: General*, *153*(2), 328–338.

https://doi.org/10.1037/xge0001488

Appendices

A. Optimality analysis for creating environment-sensitive generalization grids

To test if people can use distinct generalization strategies, we needed to design two environments to meet these criteria:

(1) Task performance (i.e., the cumulative rewards) should benefit from adapting the generalization strategy, such that some generalization strengths are more optimal than others.

(2) The optimal generalization strategy should differ between environments so that both environments favour a different, distinct level of generalization.

(3) Using environment-sensitive strategies should lead to higher cumulative rewards than applying a single shared strategy across both environments.

We performed an optimality analysis to systematically identify environment pairs that would maximize the environment-sensitive gain, i.e., the difference in performance between using environment-sensitive strategies versus a shared strategy. The analysis consisted of the following steps:

1. Define task parameters.

   We specified the grid size, the search horizon (number of trials per round), and the strength of the reward correlations (λ_{gen}) for the rough and smooth environment.

2. Simulate average performance as a function of generalization strength.

   We varied the decision-making strategies by randomly generating the three model parameters across a wide range and simulated how many rewards each agent accumulated.

3. Estimate maximal environment-specific performance.

For each environment separately, we identified the generalization parameter that yielded the highest mean reward, i.e., the optimal environment-specific generalization strategy.

4. Estimate maximal shared performance across environments.

   We also simulated performance when using a shared strategy (same three model parameters) across both environments, and identified the best shared strategy.

5. Calculate environment-sensitive gain.

   We computed the difference between the total performance achieved by using environment-specific strategies per environment and the performance achieved by using a shared strategy across environments.

6. Optimize environments.

   We repeated this process for multiple candidate environment pairs (by varying the strength of the reward correlations of the rough and smooth environment), and for various search horizons and grid sizes, to identify which pair yields the largest environment-sensitive gain.

The optimized environments used in the experiments had reward correlations of λ_{gen} = 2 and λ_{gen} = 16 for the rough and smooth environments, respectively, and search horizons of 30 with a grid size of 15x15. These values offered sufficiently different strengths of reward correlations to support distinct optimal environment-specific generalization strategies, while ensuring that adapting to the environment would meaningfully increase cumulative rewards compared to a single shared strategy (see *also Figure A1*).

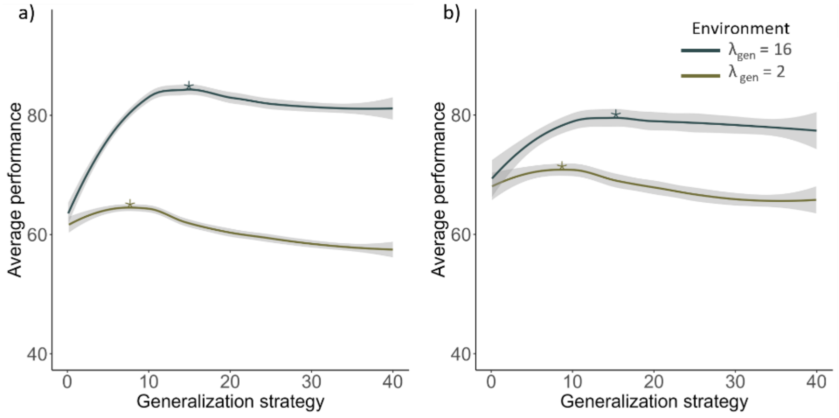


Figure A1: Visualization of optimized environments. The chosen environments, with reward correlations of λ_{gen} = 2 and λ_{gen} = 16 for the rough and smooth environments, respectively, lead to optimal performances (see *) for distinct generalization strategies. The full lines represent the mean across 1000 simulations, while the error bands indicate the standard error. a) shows the performance in function of generation strategy for the non-localized version of the model, b) shows the localized version of the model.

B. Gaussian Process Regression

To model how participants generalized observed rewards across the grid, we used Gaussian Process (GP) regression. This non-parametric Bayesian function learning approach maps each cell location **x** to a real-valued scalar output (representing the expected rewards for that cell location). A GP specifies a prior distribution over functions, formalized as:

$$GP\big(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')\big),$$

where $m(\boldsymbol{x})$ is the prior mean and $k(\boldsymbol{x}, \boldsymbol{x}')$ the kernel function that governs the covariance between pairs of cells. In our case, the prior mean was set to the median reward of the task $(m(\boldsymbol{x}) = 40)$. The kernel was a radial basis function (RBF), which encodes the similarity of two cells based on their Euclidian distance:

$$k_{RBF}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-\frac{\left||\boldsymbol{x}-\boldsymbol{x}'|\right|^2}{2\lambda^2}\right).$$

This kernel ensures that cells closer on the grid are more correlated, with the strength of this correlation decaying exponentially as distance increases. The decay rate is controlled by the length scale parameter λ, for which larger values imply longer-range generalization, more strongly correlated environments, and smoother reward landscapes, whereas smaller values yield weaker correlated environments and rougher landscapes. In the limit as $\lambda \rightarrow 0$, cells become independent, and no generalization occurs.

Formally, at trial *t* (after *t* choices), the dataset $D_t = \{\boldsymbol{X}_t, \boldsymbol{y}_t\}$ contains choice option inputs $\boldsymbol{X}_t = (x_1, \dots, x_t)$ and observed noisy rewards $\boldsymbol{y}_t = \ y_1, \ \dots \ y_t$, where each choice option j's reward $y_j \sim N\left(f\left(\boldsymbol{x}_j\right), \sigma^2\right)$ is sampled from the latent reward function $f\left(\boldsymbol{x}_j\right)$ with Gaussian noise $\sigma^2 = 1$. GP regression combines these observations with the prior to compute a posterior distribution over functions. For a new input $\boldsymbol{x}_*$, the predictive posterior is Gaussian with mean and variance:

$$m(\boldsymbol{x}_*|D_t) = \boldsymbol{K}_{*,t}^{T} \ (\boldsymbol{K}_{\boldsymbol{t},\boldsymbol{t}} + \ \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{y}_t \ ,$$

$$v(\boldsymbol{x}_*|D_t) = \boldsymbol{K}_{*,*} - \boldsymbol{K}_{*,t}^{T} \ (\boldsymbol{K}_{\boldsymbol{t},\boldsymbol{t}} + \ \sigma^2 \boldsymbol{I})^{-1} \boldsymbol{K}_{*,t} \ ,$$

with $\boldsymbol{K}_{*,t} = (k(\boldsymbol{x}_1, \boldsymbol{x}_*), \dots, k(\boldsymbol{x}_t, \boldsymbol{x}_*))^{\mathrm{T}}$ the covariance between each observed input $\boldsymbol{x}$ up to $\boldsymbol{x}_t$ and new input $\boldsymbol{x}_*$, and $\boldsymbol{K}_{\boldsymbol{t},\boldsymbol{t}} = k(\boldsymbol{x}_t, \boldsymbol{x}_t)$ between observed inputs.

In practice, the posterior mean $m(\boldsymbol{x})$ provides the model's estimate of the expected reward at each cell, while the posterior variance $v(\boldsymbol{x})$ yields an uncertainty estimate $s(\boldsymbol{x}) = \sqrt{v(\boldsymbol{x})}$. Together, these characterize how observed rewards are generalized across the grid, enabling us to capture the strength and extent of participants' generalization strategies.

C. Optimality analysis for creating maximally diagnostic generalization test scenarios

In the meta-learning test phase of Experiment 3, our goal was to test whether participants had meta-learned to associate the generalization strength with specific environmental cues. To do so, we designed single-shot test scenarios that maximally differentiated between smooth-like (stronger) and rough-like (weaker) generalization strategies. Each test scenario consisted of two revealed cells and a third highlighted cell for which participants were asked to report their reward expectation. The highlighted cell was optimally selected such that participants' expectations, if prompted by the generalization strategies previously used in the two environments, were maximally different between the environments. To identify the most diagnostic test scenarios, we performed an optimality analysis, containing the following steps:

1. Generate a candidate scenario.

   Similarly to the reward distributions used in the training phases, we created a 15x15 reward grid using a Gaussian Process with a smooth kernel. Two cells were randomly selected and opened.

2. Ensure environment-agnostic observations.

   To prevent the observed rewards from revealing the needs of the grid (i.e., if the reward distribution was smooth or rough), we fitted Gaussian Process Regression with the length scales of the smooth ($\lambda_{gen} = 16$) and rough ($\lambda_{gen} = 2$) reward distributions. We assessed the goodness of fit to verify that both length scales explained the observations equally well.

3. Assess the diagnostic value of each potential third highlighted cell.

   We used participants' estimated generalization parameters of the smooth and rough environments of Experiment 1 (we omitted overlapping generalization parameters from the full smooth and rough distributions) to respectively simulate smooth-like and rough-like reward expectations. For each candidate unopened cell, we computed the mean reward expectation under both smooth-like and rough-like generalization. The difference between these means defines the diagnostic value of the candidate cell.

4. Select the maximally diagnostic cell.

We identified the cell for which the difference between smooth-like and rough-like mean reward expectations was largest. If this difference exceeded a predefined threshold (≥ 15), we accepted the scenario.

This procedure was repeated until we identified 15 unique test scenarios. We accepted scenarios such that half of them (7) showed lower mean expectations for rough-like generalization. 5 test scenarios were uniquely shown in smooth environments, 5 in rough, and 5 were presented in both environments, creating 20 different test trials.

D. Results per experiment

Table 1

Environment-specific strategies: behavioral results per experiment

	Experiment 1			**Experiment 2**			**Experiment 3**		
Behavioral measure	**W**	**z**	**p**	**W**	**z**	**p**	**W**	**z**	**p**
Novel clicks	1788	-0.71	.481	1408	-2.29	.022	1412	-2.96	.003
High value clicks	1601	-0.46	.644	2267	-1.49	.136	2685	-1.70	.088
Distance from previous click	2921	-3.51	<.001	2318	-1.50	.135	3618	-4.96	<.001
Distance from most nearby high value cell	3567	-6.40	<.001	3223	-5.26	<.001	4125	-6.56	<.001

Table 2

Environment-specific strategies: computational results per experiment

	Experiment 1			Experiment 2			Experiment 3		
Behavioral measure	**W**	**Z**	**p**	**W**	**z**	**p**	**W**	**z**	**p**
Generalization	851	-4.81	<.001	1100	-3.57	<.001	933	-5.10	<.001
Uncertainty-guided exploration	1417	-2.25	.025	1566	-1.47	.140	2095	-0.96	.33
Random exploration	1214	-3.35	<.001	355	-6.67	< .001	1858	-1.72	.086

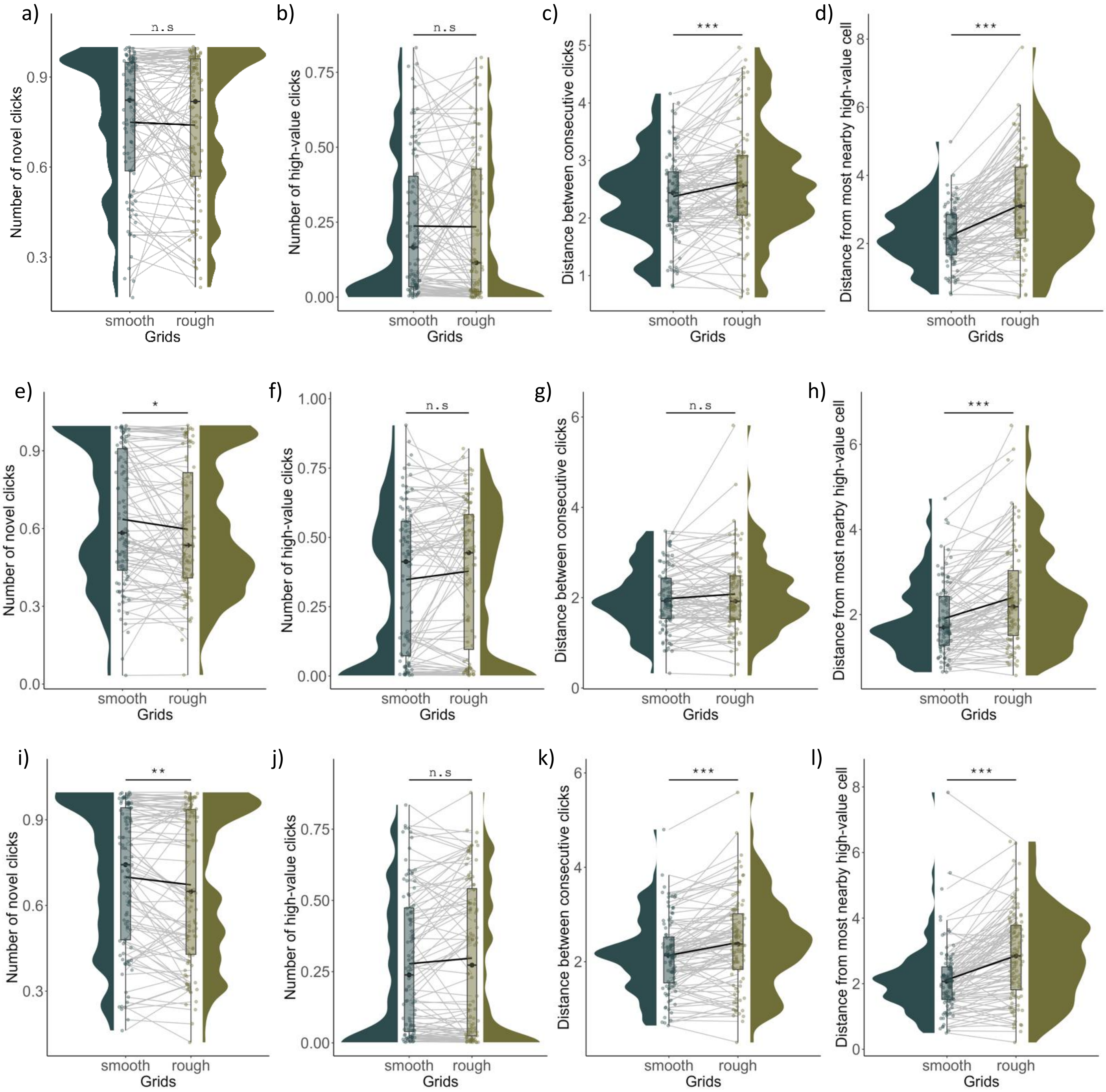


Figure A2: Behavioral results per experiment (Experiment 1 a-d, Experiment 2 e-h, Experiment 3 j-l). *** *p* value Wilcoxon signed-rank test <.001, ** *p* value Wilcoxon signed-rank test ε [.001, .005[, * *p* value Wilcoxon signed-rank test ε [.005, .05[, n.s. *p* value Wilcoxon signed-rank test > .05.

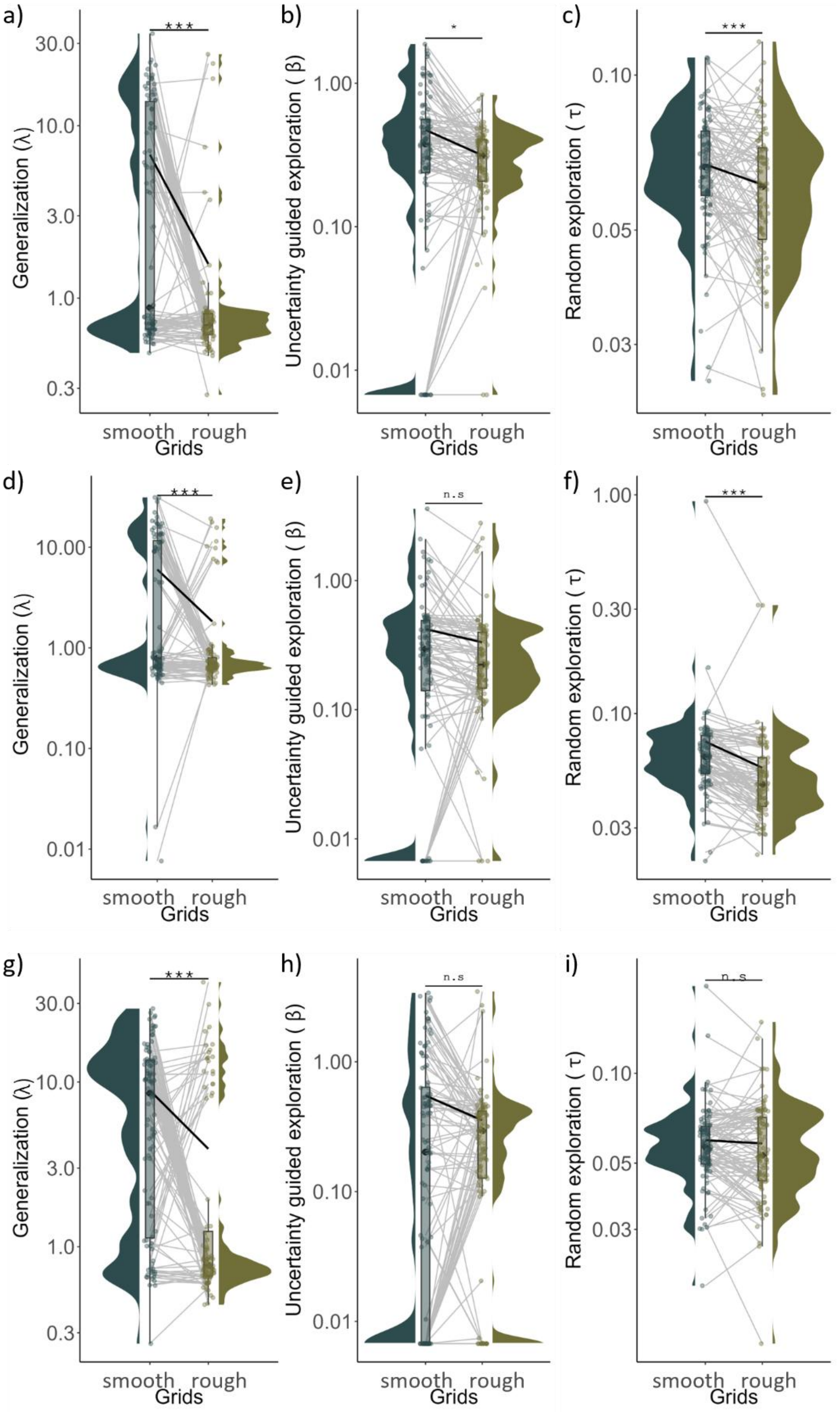


Figure A3: Computational results per experiment (Experiment 1 a-c, Experiment 2 d-f, Experiment 3 g-i). *** *p* value Wilcoxon signed-rank test <.001, * *p* value Wilcoxon signed-rank test ε [.005, .05[, n.s. *p* value Wilcoxon signed-rank test > .05.

E. Individual differences: extra figures

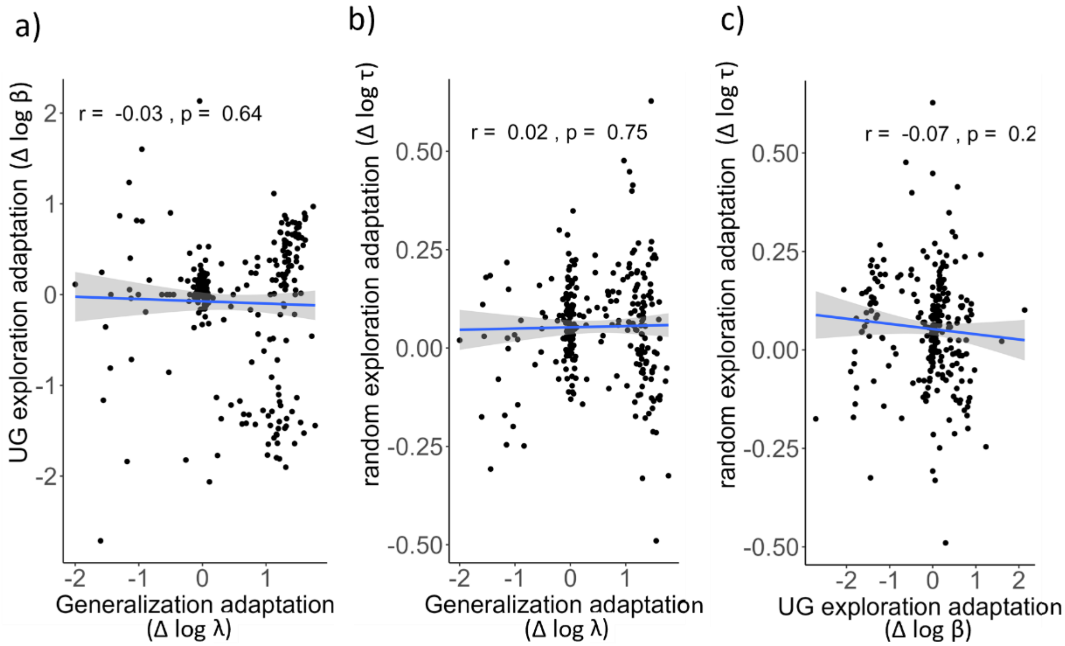


Figure A4: Model parameter adaptations. Adaptations are calculated by subtracting the logged parameters in the rough environments from the logged parameters in the smooth environments. Positive parameter adaptations therefore refer to larger parameters in smooth environments. Pearson's correlations and *p* values are shown in the respective plots. We did not find any pairwise parameter adaptation.

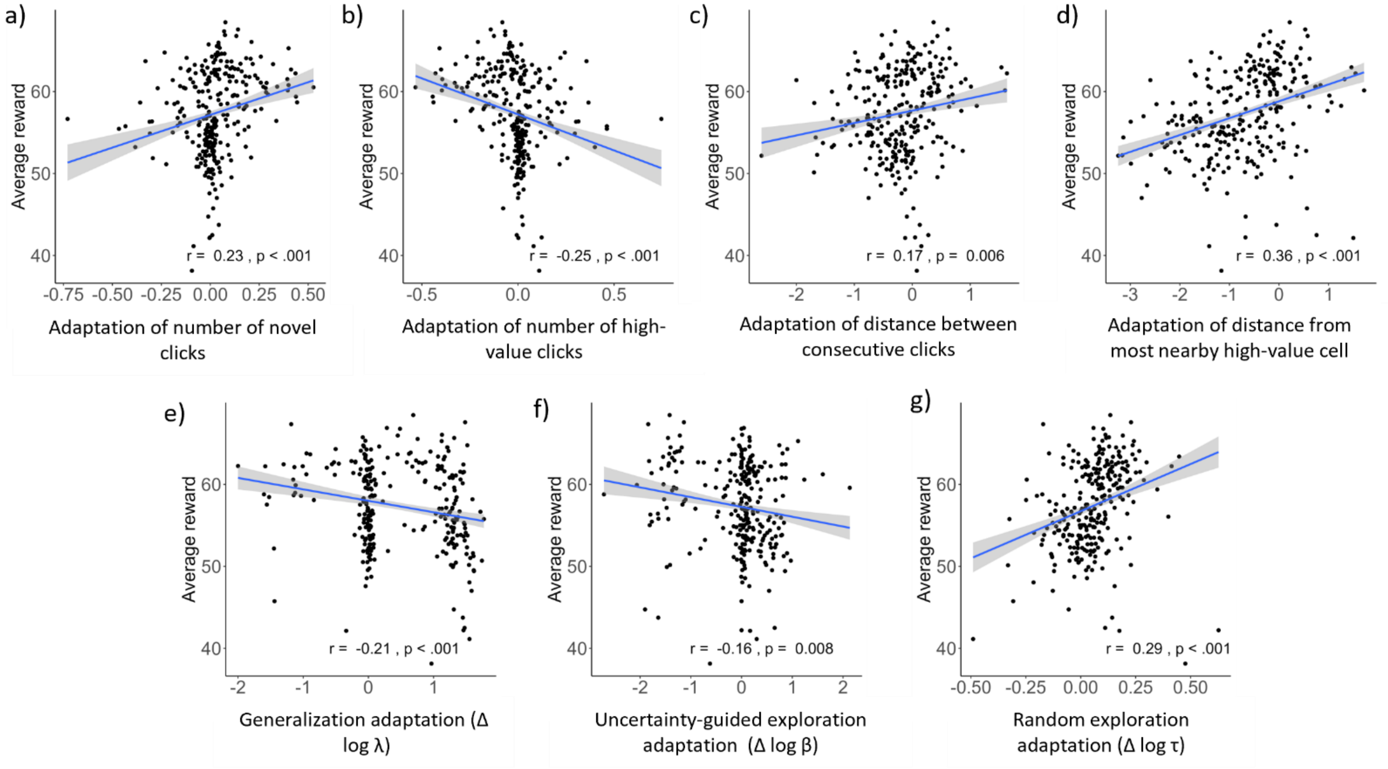


Figure A5: Performance benefits of strategy adaptations. Adaptations are calculated by subtracting the measures of the rough environments from the measures of the smooth environments. Positive adaptations therefore refer to larger measures in smooth environments. Pearson's correlations and *p* values are shown in the respective plots.