D I S S E R T A T I O N

zur Erlangung des akademischen Grades
Doctor rerum naturalium (Dr. rer. nat.) im Fach Psychologie

Guided by generalization and uncertainty:
A theory of human learning and exploration

eingereicht an der Lebenswissenschaftlichen Fakultät der Humboldt–Universität zu Berlin
von **Charley M. Wu, M.Sc.**

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät der Humboldt-Universität zu Berlin
Prof. Dr. Bernhard Grimm

Gutachter/innen:

1. Dr. Björn Meder
2. Prof. Dr. Thomas T. Hills
3. Prof. Dr. Ralph Hertwig

Tag der mündlichen Prüfung: 06.08.2019

Berlin, im April 2019

This thesis is dedicated to Agnieszka for teaching me the meaning of *Verwunderung*.

Acknowledgements

This thesis would not have been possible without the support of my advisers Björn Meder and Jonathan Nelson. I could not have asked for better mentors and better human beings to introduce me to the world of science. I am incredibly grateful for their patience and selfless dedication to answering my countless naïve questions. Somehow they managed to give me the perfect balance of enough freedom to explore, but also enough structure to develop. Any success that might come my way is completely owed to them.

I would also like to thank Eric Schulz for being the Bill to my Ted and the Wayne to my Garth. Since we began collaborating together, Eric has been an oracle of statistical knowledge and a fountain of keen scientific ideas. The projects that eventually developed into this thesis are built upon an absolute mountain of correspondence between us, where I could always count on Eric to give me a thoughtful and honest assessment of my ideas, no matter how misguided or just simply wrong.

I am grateful to have so many people who have helped me along the way. I thank my parents for teaching me the value of hard work and being supportive of me even when I decided to become a Philosophy major. I am thankful to: Gerd Gigerenzer for pushing me to take more risks. Ralph Hertwig for taking me in as an orphan PhD student without a group. Tim Pleskac for always being in my corner and giving me freedom to explore. Nico Schuck for showing me that a real scientist is dedicated to his students. Rob Goldstone for patiently guiding me through unfamiliar territory. Azzurra Ruggeri for reminding me to take life less seriously and to have a good time along the way. Sam Gershman for teaching me good ideas need to also be simple. Maarten Speekenbrink for being a soft-spoken sage of reinforcement learning. Shenghua Luan for teaching me how to own your mistakes. Amit Kothiyal for showing me how the sage does nothing, yet nothing is left undone. Peter Todd for opening my eyes to the profundity of search. Christian Elsner for being a pillar of the MPIB community. Simón Algorta and Marcus Buckmann for showing me the ropes. Laura Martignon for reminding me that life demands music and celebration. Giorgio Coricelli for being such a likable person even though we disagree about so much. Josh Tenenbaum for challenging me to tackle bigger questions. Daniel Barkoczi and Pantelis Analytis for introducing me to the study of collective behavior. Kyanoush Yahosseini for striving toward the Daoist principle of effortless action. Alan Novaes

Tump for his serene enthusiasm for the natural world. Christina Leuker for helping wade me through the administrative quagmire. Christoph Engel, Oliver Kirchkamp, Susanne Büchner, and everyone at the IMPRS Uncertainty school for so kind and hospitable despite my heretical beliefs about optimality. Steve Lewandowski, Simon Farrell, Klaus Oberauer, Bob French, Jörg Rieskamp, and everyone at the computational modeling summer school for giving me one of the most useful tools in my repertoire. Michael Krause for being the best goddamn systems administrator in the history of the Max Planck Society. The Max Planck PhDnet for their tireless efforts in improving the lives of students.

Abstract

How do people navigate the vastness of real-world environments where it is not feasible to explore all possibilities and the exact same situation is rarely encountered twice? The study of human learning has made rapid progress in the past decades, from discovering the neural substrate of reward prediction errors, to using similar principles to build artificial intelligence capable of beating the best human players in skillful games such as Go. Yet this line of research has primarily focused on learning through repeated interactions with the same stimuli. How are humans able to rapidly adapt to novel situations and learn from such sparse examples?

We propose that generalization plays a crucial role in guiding human exploration and learning. Inspired by Roger Shepard's law of generalization, we present a domain general theory of generalization across spatial, conceptual, and structured environments. We use a function learning model to describe how people generalize limited experiences to a wide set of novel possibilities, based on the simple principle that similar actions produce similar outcomes. Our model of generalization generates predictions about the expected reward and underlying uncertainty of unexplored options, where both are vital components in how people actively explore the world.

Across 8 experiments, we show how generalization and uncertainty guides human learning. Chapters 1 and 2 introduce the psychological and mathematical basis of our theory. In Chapter 3 we study search on generated and real-world environments with spatially correlated rewards, where we find robust and recoverable evidence for our model of generalization and exploration. Chapter 4 uses a field study to investigate the source of variability in the sampling behavior of children. Contrary to prevailing theories, we find that children are not simply more random, but they generalize less and seek out uncertainty more eagerly than adults. Chapter 5 extends the scope of our theory to connect search across spatial and conceptual domains. Chapter 6 provides a unifying framework for generalization and search in structured spaces, where we show that the models used in the previous chapters are a special case of a graph generalization model. Finally, Chapter 7 provides a summary of the main contributions and lays the groundwork towards a process-level theory.

Zusammenfassung

Wie navigieren Menschen in unüberschaubaren Umgebungen, wenn es nicht möglich ist, alle Optionen ausführlich zu untersuchen? Die Forschung zum menschlichen Lernen hat in den letzten Jahrzehnten rasche Fortschritte gemacht: Von der Entdeckung des neuronalen Substrats für Vorhersagefehler bis hin zur Verwendung ähnlicher Prinzipien zum Aufbau künstlicher Intelligenz, die in Geschicklichkeitsspielen wie Go die besten menschlichen Spieler schlagen kann. Bisher lag der Forschungsschwerpunkt auf dem Lernen durch wiederholte Interaktionen mit ein und derselben Situation. Dabei ist ungeklärt, wie Menschen sich schnell an neue Situationen anpassen können und aus spärlichen Beispielen lernen können.

Wir schlagen vor, dass die Generalisierung eine wesentliche Rolle dabei spielt, wie Menschen effizient die Welt erforschen und sich in ihr zurechtfinden. Inspiriert von Roger Shepards „law of generalization“ präsentieren wir eine allgemeine Theorie der Generalisierung in räumlichen, konzeptuellen und strukturierten Umgebungen. Wir verwenden ein Funktionslernmodell, um zu beschreiben, wie Menschen begrenzte Erfahrungen auf eine Vielzahl neuer Situationen generalisieren, basierend auf dem einfachen Prinzip, dass ähnliche Aktionen zu ähnlichen Ergebnissen führen. Unser Modell der Generalisierung erzeugt Vorhersagen über die erwartete Belohnung und die damit verbundene Unsicherheit unerforschter Optionen—zwei wichtige Komponenten dafür, wie Menschen die Welt aktiv erkunden.

In acht Experimenten zeigen wir, wie die durch Generalisierung und Ungewissheit bedingte Erforschung das menschliche Lernen lenkt. Die Kapitel 1 und 2 stellen die psychologischen und mathematischen Grundlagen unserer Theorie vor. In Kapitel 3 untersuchen wir die Suche in generierten und natürlichen Umgebungen mit räumlich korrelierten Belohnungen. Hierbei finden wir robuste und wiederherstellbare Beweise für unser Modell der Generalisierung und Erkundung. In Kapitel 4 verwenden wie eine Feldstudie, um die Quelle der Variabilität im Probenahmeverhalten von Kindern zu untersuchen. Im Gegensatz zu den vorherrschenden Theorien stellen wir fest, dass Kinder nicht einfach mehr zufällig Entscheidungen treffen, sondern weniger generalisieren und unsichere Optionen eifriger erkunden als Erwachsene. Kapitel 5 erweitert den Umfang unserer Theorie, um die Suche über räumliche und konzeptuelle Domänen hinweg zu verbinden. Kapitel 6 bietet ein einheitliches Rahmenwerk für Generalisierung und Suche in strukturierten Umgebungen. Hier zeigen wir, dass die in den vorangegan-

genen Kapiteln verwendeten Modelle einen Sonderfall eines Graph-Generalisierungsmodells darstellen. Schließlich liefert Kapitel 7 ein Überblick über die wichtigsten Beiträge und legt die Grundlagen für eine process-level Theorie.

Table of contents

List of figures

List of tables

Nomenclature

Roman Symbols

$\mathcal{D}$	A set of data or observations $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ containing outputs $\mathbf{y}$ observed for inputs $\mathbf{X}$
$\mathbf{X}$	A matrix of data points in $\mathbb{R}^d$, where each row $\mathbf{x}_i \in \mathbf{X}$ is an observation
$\mathbf{x}$	A vector representing a stimulus or a reward-generating option
$\mathbf{y}$	A vector of real-valued scalar outputs (e.g., rewards), where each $y_i \in \mathbf{y}$ corresponds to a single observation
$\mathcal{G}(\mathbf{x}, \mathbf{x}')$	Generaliation gradient. The probability of responding to stimuli $\mathbf{x}$ when $\mathbf{x}'$ is observed
$\mathcal{H}$	Hypothesis space
$\mathcal{H}_K$	Hilbert Space
$\mathbf{K}$	Gram matrix, where the rows and columns are indexed by the elements of $\mathbf{X}$ such that $\mathbf{K}_{x,x'} = k(x,x')$. This is can be interchangeably called a kernel just as $k(x,x')$ is called a kernel
$k(x,x')$	Kernel function (also known as a covariance function)
$\mathcal{A}$	A set of actions, where an agent can selects $a \in \mathcal{A}$ at any point in time.
$\mathcal{S}$	A set of states, where the environment is in some state $s \in \mathcal{S}$ at any point in time.
$R(s,a)$	The reward function for performing action a from state s.
$V(s)$	The value of state s based on the expected sum of discounted future rewards
G	Kalman gain
$\mathcal{GP}$	A Gaussian Process (GP) defines a multivariate normal distribution over functions $f \sim \mathcal{GP}(m(x), k(x,x'))$, where functions $f : \mathcal{X} \to \mathbb{R}^n$ map the input space $\mathcal{X}$ to real-valued scalar outputs
$m(\mathbf{x})$	The posterior mean reward of input $\mathbf{x}$
$v(\mathbf{x})$	The posterior variance of input $\mathbf{x}$

Greek Symbols

α	Parameter governing the rate of diffusion in the diffusion kernel
β	Exploration bonus of the UCB sampling strategy
δ	Prediction error
η	Learning rate, e.g., in the Rescorla-Wagner model or Temporal Difference learning model
λ	Length scale of the RBF kernel
ϕ	A feature map from $\mathbf{X} \rightarrow \mathcal{H}_K$, also known as a basis function
τ	Temperature parameter of the softmax choice algorithm
θ_ε^2	Error variance parameter of the Kalman Filter and Bayesian Mean Tracker models

Acronyms / Abbreviations

AI	Artificial Intelligence
BF	Bayes Factor; the ratio of the likelihood of one particular hypothesis to the likelihood of another
BMT	Bayesian mean tracker.
CDF	Cumulative Distribution Function
DA	Dopamine
GP	Gaussian Process regression.
i.i.d.	Independent and identically distributed
MAB	Multi-armed bandit
MDP	Markov decision process.
MDS	Multidimensional scaling
PDF	Probability Density Function
RBF	Radial Basis Function kernel, sometimes also called the squared exponential kernel or the Gaussian kernel.
RW	Rescorla-Wagner model
SNc	Substantia nigra

UCB Upper Confidence Bound sampling.

VTA Ventral tegmental area

WSLS Win-stay lose-sample.

Chapter 1

Exploring the Landscape of Human Learning

The years of our lives number less than one hundred
Yet we agonize over a thousand years of worry.
The days are short and the bitter nights are long!
Why not take up torches and go wandering out into the dark?
—Han Dynasty Poem (ca. 200 CE)

What principles guide human learning in large problem spaces, where the same situation is rarely encountered twice? Rather than merely groping through the dark, the goal of this thesis is to show how human learning is guided by generalization. Past experiences can be leveraged to make predictions about an endless number of novel situations, thereby guiding search towards promising options. In this chapter, we present an overview of influential theories of learning and generalization in psychology. We trace the development of modern reinforcement learning algorithms to origins in theories of animal and human learning. Despite a deep understanding of the behavioral and neural signatures of learning through repeat interactions with the same stimuli, prevalent theories in this framework lack an account of how people adapt to novel situations. Inspired by Shepard's law of generalization, we propose a theory of generalization using the mechanism of function learning. Functions represent inferred relationships between actions and outcomes, and provide guiding predictions across an entire continuum of possibilities. Across spatial, conceptual, and structured domains, we show that generalization acts as a torch for guiding human exploration in unknown and uncertain environments.

1.1 The gap between human and machine learning

One of the mysteries of human cognition is how people perform rapid inference and generalization from a sparse number of examples (Lake, Salakhutdinov, & Tenenbaum, 2013; Vul, Goodman, Griffiths, & Tenenbaum, 2014). While recent advances in machine learning have produced algorithms that achieve human–level performance in sophisticated games such as Go (Silver et al., 2016), there still exists an intriguing gap between the efficiency of human and machine learning (Lake, Ullman, Tenenbaum, & Gershman, 2017). For example, consider that AlphaGo was trained on approximately 10.5 million virtual games[†] of Go in order to achieve what DeepMind described as "masterful" performance (Silver et al., 2016). If a human were to completely devote themselves to the game of Go, playing 24 games a day without sleep, it would take 1200 years to achieve the same depth of experience as AlphaGo acquired. Recent improvements to this algorithm, as demonstrated in the AlphaStar algorithm designed for the game Starcraft II, still require up to 200 years of real-time play during training (Vinyals et al., 2019). While neither domain can be considered a trivial learning problem, the immense quantity of training data illustrates a qualitative difference in how people and modern computational algorithms learn through experience. Since first posed by Plato in the Meno (380 BCE), the question still remains: how do people learn from such sparse examples?

One potential answer may be found in the strategies people use to actively explore the world. By asking questions (Coenen, Nelson, & Gureckis, 2018; Nelson, 2005; Rothe, Lake, & Gureckis, 2018; Wu, Meder, Filimon, & Nelson, 2017), directing eye-movements towards informative scenes (Cavanagh, Hunt, Afraz, & Rolfs, 2010; Najemnik & Geisler, 2005; Nelson & Cottrell, 2007), or sampling different options (Hertwig, Barron, Weber, & Erev, 2004; Hills & Hertwig, 2010), people are remarkably efficient at acquiring relevant information for making good decisions (Pirolli & Card, 1999; Todd, Hills, & Robbins, 2012). We investigate this capacity for active learning as a potential source of the gap between human and machine learning. Using the metaphor of "learning as search", this thesis aims to understand the mechanisms that guide human exploration through complex and uncertain environments.

This introduction chapter is structured as follows. We first trace the development of psychological theories of learning, specifically related to value or reward-based learning (e.g., which actions lead to rewarding outcomes). Models of associative learning provide a common point of origin for both neural theories of human learning and computational algorithms

[†]160,000 games were used to train the Policy network to classify positions according to expertise, and another 1.6 million games were used to train the policy gradient. Additionally 1.6 billion positions were used to train the value function, while 8 million positions were used to train the rollout policy. Assuming an average of 183.75 positions per game (as was the case in the dataset used for the policy network), this is equivalent to another 8.7 million games. While this does not include the full extent of training data (but rather only what can be easily sorted into games), we find that this sums to a total of about 10.5 million virtual games of Go.

used in large-scale machine learning applications. However, one missing ingredient in the associative learning framework is the ability to generalize previous experiences to novel situations. The most influential solution to this problem comes from Shepard's (1987) law of generalization, which describes generalization as an exponential function of distance in an internal representational space. This raises important questions about how knowledge is organized. A spatial representation is intuitive and largely effective, but lacks the ability to capture structured and asymmetric representations. We propose a kernel representation of similarity that unifies both spatial and structural organizations of knowledge. Kernel similarity forms the basis of a function learning model, which we use to describe how people generalize limited experiences to a wide set of novel possibilities, with functions representing a candidate hypotheses about the relationship between actions and outcomes. Across spatial, conceptual, and structured domains, we show that generalization guides exploration in large problem spaces, by illuminating which unknown regions seem most promising to explore. Combined with a bold heuristic that proposes "optimism in the face of uncertainty", we find evidence for general principles of generalization-guided exploration at the foundation of human learning.

1.1.1 Learning as search

From foraging for resources in a spatial landscape (Hills et al., 2015; Kolling, Behrens, Mars, & Rushworth, 2012; Wu, Schulz, Speekenbrink, Nelson, & Meder, 2018) to searching through hypothesis space in order to learn causal relationships (Bramley, Dayan, Griffiths, & Lagnado, 2017; Mitchell, 1982; Ullman, Goodman, & Tenenbaum, 2012), many aspects of human cognition can be captured by the metaphor of search. At any moment, we are presented with a near infinite number of possibilities, from which we must search for the right action to take. What should I have for lunch? Which path should I take to get to work? Which experiment should I conduct next? Indeed, this is the exact problem for which Kierkegaard (2004/1849) coined the term "despair of the infinite". In such vast decision spaces and with only a finite horizon for exploration, how can one possibly make the right decision?

William James was one of the first to use the metaphor of search in the context of cognition, describing how people retrieve memories similar to a person searching for a misplaced item in a house. "In both cases we visit what seems to us the probable neighborhood of that which we miss" (James, 1890, Ch. 16). But what defines probable? And how are cognitive representations organized into neighborhoods? William James appealed to the intuition that search often begins near *associated* items, and that the "machinery of recall is thus the same as the machinery of association" (James, 1890). A lost cup might be found on the kitchen table or near the sink. A lost key might be by the door or in the hallway.

Edward Thorndike pioneered the empirical study of this "machinery of association" (James, 1890), by putting cats in puzzle boxes and observing that they escaped faster and faster over successive trials (Thorndike, 1898, 1911). Thorndike's "Law of Effect" (1911) stated that actions associated with satisfaction are strengthened, while those associated with discomfort become weakened. Together with Pavlov's (1927) theory of classical conditioning, these early demonstrations of *associative learning* using cats and dogs would form the basis for many theories of human learning in psychology. Both Thorndike and Pavlov were important sources of inspiration for Clark Hull's (1943) theory of "drives", which modulate how stimulus-response pairings are translated into behavior, as well as B.F. Skinner's (1938) approach of shaping behavior through rewards.

1.1.2 Learning by predicting the world

Early approaches to associative learning broadly fall under two categories: classical condition and operant conditioning. Classical condition (e.g., Pavlov, 1927) described learning as a passive coupling of stimulus and response, whereas operant conditioning (e.g., Skinner, 1938; Thorndike, 1911) described learning as a shift of behavior in response to rewards or punishments. Initially developed as a model of classical conditioning, the Rescorla-Wagner[†] (RW) model was greatly influential to both camps of associative learning by reframing learning as an active process of making predictions of the world. Thus, highly surprising outcomes resulted in greater changes to learned responses, since the magnitude of the prediction error would be large (Rescorla & Wagner, 1972). The RW model uses a linear combination of weights $\mathbf{w}_t$ and a vector of conditioned stimulus intensities $\mathbf{x}_t$ (e.g., the sound of a bell) to predict the unconditioned stimulus $V(\mathbf{x}_t)$ (e.g., the arrival of food) on trial t :

$$V(\mathbf{x}_t) = \mathbf{w}_t^\top \mathbf{x}_t \tag{1.1}$$

Learning is modeled by updating the weights on each trial based on the prediction error δ_t:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \eta\,\delta_t \mathbf{x}_t \tag{1.2}$$

$$\delta_t = r_t - V(\mathbf{x}_t), \tag{1.3}$$

where $\eta \in [0,1]$ is the learning rate parameter, and δ_t is the prediction error, calculated as the difference between the acquired reward r_t and the predicted value $V(\mathbf{x}_t)$. Thus, the amount of

[†]The RW model was originally an extension of the Bush and Mosteller (1951a, 1951b) model, but has largely overshadowed the latter in terms of popularity. As pointed out by Glimcher (2011), the RW is often incorrectly attributed as the origin of prediction error updating, although it was first formalized by Bush and Mosteller (1951a, 1951b).

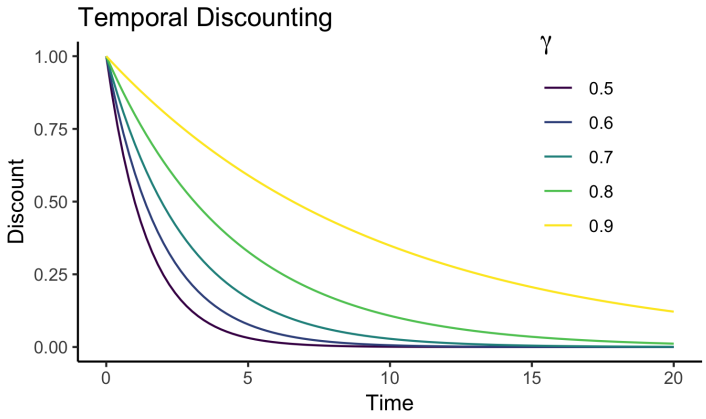

Fig. 1.1 Temporal discounting at different discount rates (γ). Lower values of γ result in more rapid discounting of rewards as a function of time.

learning is modulated by the (un)predictability of rewards, where large prediction errors lead to larger updates and faster learning. The important contribution of Bush and Mosteller (1951a, 1951b) and the RW model (Rescorla & Wagner, 1972) was in recasting the process of learning as an active process of making predictions and recalibrating them based on the magnitude of the error, rather than a solely passive association of rewards.

Temporal difference learning

While the RW model explains learning based on the prediction of immediate rewards, it fails to address an important aspect of what Minsky (1961) called the "credit-assignment problem". Specifically, the RW is unable to distribute credit for delayed or long-term rewards to the previous actions or stimuli that were involved in producing it. Eating a balanced diet and exercising regularly may not result in much immediate reward (relative to feasting on fast food and binge drinking), yet it may be reasonable to expect better long-term rewards. To solve this problem[†], Richard Sutton developed the Temporal Difference (TD) model to learn predictions of value based on long-term future rewards instead of only immediate rewards (Sutton, 1988; Sutton & Barto, 1990). TD learning uses a discount factor $\gamma \in [0,1]$ that controls the rate at which we discount future rewards (Fig 1.1), where the value function $V(\mathbf{x}_t)$ encodes the expected sum of discounted future rewards:

$$V(\mathbf{x}_t) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r_{t+k}\right] \tag{1.4}$$

[†]The TD learning model was also designed to solve the problem of modeling continuous time rather than only discrete time intervals (i.e., trials).

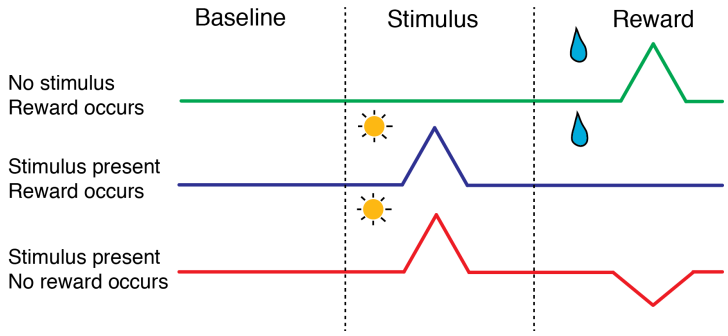

Fig. 1.2 Dopamine responses adapted from Schultz et al. (1997).

Thus, we can modify the RW prediction error equation based on trying to predict discounted future rewards instead of only immediate rewards:

$$\delta_t = r_t + \gamma V(\mathbf{x}_{t+1}) - V(\mathbf{x}_t), \tag{1.5}$$

This value updating equation is identical to the previous RW equation (Eq. 1.3), except for the addition of the discounted future reward term $\gamma V(\mathbf{x}_{t+1})$. Another way to interpret temporal discounting is that rewards are generalized to the sequence of previous states that contributed to the eventual outcome, where the discount factor controls the rate of exponential decay of reward generalizations over time.

Prediction error and dopamine

The principle of prediction error learning has influenced the majority of modern reinforcement learning models (Sutton & Barto, 2018), playing an integral role in large-scale applications from self-driving cars (Michels, Saxena, & Ng, 2005; Sallab, Abdou, Perot, & Yogamani, 2017) to robotics (Kober, Bagnell, & Peters, 2013; OpenAI et al., 2018). Yet aside from being a useful principle in applied domains, a prediction error learning rule has also been widely linked to the firing rate of Dopamine (DA) neurons in the ventral tegmental area (VTA) and substantia nigra (SNc) of the midbrain (Glimcher, 2011; Lak, Stauffer, & Schultz, 2014; Schultz, 2016). Based on a previously developed theoretical framework (Montague, Dayan, Nowlan, Pouget, & Sejnowski, 1993; Montague, Dayan, & Sejnowski, 1996), Schultz et al. (1997) showed that the activity of midbrain Dopamine (DA) neurons specifically correspond to a TD prediction error model of learning.

Thirsty monkeys were studied using single cell recordings of DA neurons under two conditions. In one condition, the monkeys were rewarded with water squirted into their mouths at unpredictable times without any stimulus present, whereas in the second condition, a visual stimulus preceded each water reward. In the no stimulus condition, the unpredictable rewards

were immediately followed by a burst of action potentials in the DA neurons (Fig. 1.2 top). In the visual stimulus condition, DA neurons initially responded the same as the no stimulus condition (i.e., a spike following the reward). However, the DA neurons gradually shifted their response from the actual reward to the visual stimulus that predicted the reward. Eventually, DA neurons no longer responded to the water reward, but only to the stimulus (Fig. 1.2 middle). Additionally, once the monkeys learned to associate the stimulus with reward, they were again shown the stimulus but without any reward occurring. This caused the usual DA response following the stimulus, but then a depressed DA response when the predicted reward failed to occur (Fig. 1.2 bottom), indicating a negative prediction error.

Since then, a host of human brain imaging studies (Berns, McClure, Pagnoni, & Montague, 2001; D'Ardenne, McClure, Nystrom, & Cohen, 2008; O'Doherty, Dayan, Friston, Critchley, & Dolan, 2003; Pagnoni, Zink, Montague, & Berns, 2002; Pessiglione, Seymour, Flandin, Dolan, & Frith, 2006) have replicated these main results, showing that the prediction error response of DA neurons is a robust feature of human learning. However, one criticism of this paradigm is that it fails to account for how people generalize about novel situations (Gardner, Schoenbaum, & Gershman, 2018; Gershman & Niv, 2015). DA neurons are also shown to respond to novel stimuli (Ljungberg, Apicella, & Schultz, 1992), with the response fading as the novelty wears off (Menegas, Babayan, Uchida, & Watabe-Uchida, 2017). This has sometimes been explained as an exploration or shaping bonus (Gershman, 2017; Kakade & Dayan, 2002). However, other theories have argued that DA responses are more sophisticated than the model-free prediction errors of the TD learning model, and that it corresponds to predictions based on inferences about the structure of the environment (Bromberg-Martin, Matsumoto, Hong, & Hikosaka, 2010; Daw, Gershman, Seymour, Dayan, & Dolan, 2011; Gardner et al., 2018).

While the DA prediction error theory provides one of the rare cases of a top to bottom theory of cognition[†], the scope appears limited to settings where learning occurs through repeated interactions with the same stimulus. There is still a need to understand how people generalize to new stimuli and learn to predict rewards in novel situations, since real world learning hardly ever involves the exact same situation twice (Gershman & Daw, 2017).

1.2 Generalization as similarity

A defining feature of human intelligence is that we are "adaptive systems, whose behavior is highly flexible" (Simon, 1990). Yet both the RW and TD learning models are only calibrated

[†]David Marr famously proposed that the study of intelligent systems should be conducted at three different levels of analysis (Marr, 1982; Marr & Poggio, 1976). The three levels, from top to bottom, consist of: the computational level (what is the problem being solved?), the algorithmic level (which representations or processes are used to implement the solution?), and the physical level (how is the system physically realized?).

for previously experienced pairings of stimulus and reward. Thus, a critical deficiency is that they fail to account for how people learn and adapt to novel situations. It only takes a small number of features (e.g., the amount of each ingredient in a meal or a cocktail) such that the total number of possibilities vastly outnumbers what can be explored in a human lifetime. Yet even without experience, one can predict that a tuna–marmalade sandwich or a Jägermeister mayonnaise shot would taste awful (Gershman, Malmaud, & Tenenbaum, 2017). Because we cannot exhaustively explore all possibilities, what principles guide our exploration towards promising yet novel items?

In this section we first explore the massively influential literature on stimulus generalization, where the similarity between stimuli defines the level of generalization. A central question is how to represent similarity, with deep implications for theories about how people represent knowledge. An intuitive approach is to use a spatial representation, where larger distances in representational space correspond to less similarity. However, there are also challenges to the spatial approach, with alternative representations being better suited for capturing structured and asymmetric forms of human knowledge (i.e., tree structures or sequential dynamics). We resolve these issues by introducing a kernel representation of similarity, which can operate on both spatial and structured forms of knowledge. This also allows us to extend the principles of generalization between pairs of items to a function learning framework, where a function represents an inferred relationship between actions and outcomes, producing an adaptive range of predictions about novel situations.

1.2.1 Shepard's universal law of generalization

Inspired by Newton's (1687) laws of physics, Roger Shepard famously proposed that the first law of psychology should be a law of generalization and that it follows an invariant exponential form (Shepard, 1987). Already, Pavlov had noticed that his dogs would salivate as a response to other sounds, distinct from the bell or whistle they had been conditioned on, and that this response was more likely when the pitch was similar (Pavlov, 1927).

Using a stimulus-response paradigm, Shepard (1958) trained participants (both humans and animals) to associate a variety of different stimuli (e.g., a specific color or tone) with a response (e.g., pushing a button or pulling a lever). When the stimuli were modified (e.g., by changing the hue of a color or the pitch of a tone), the probability of responding decayed as an exponential function of distance between the original stimulus $\mathbf{x}$ and modified stimulus $\mathbf{x}'$:

$$\mathcal{G}(\mathbf{x}, \mathbf{x}') \propto \exp\left(-d(\mathbf{x}, \mathbf{x}')\right) \tag{1.6}$$

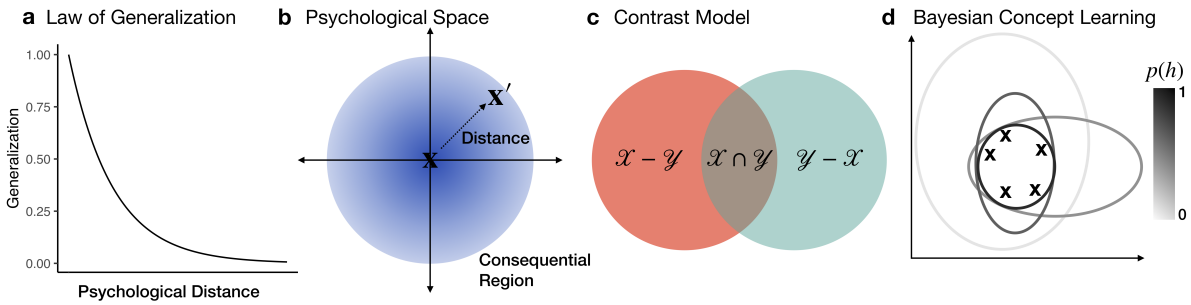

Fig. 1.3 Theories of generalization and similarity. **a**) Shepard's (1987) law of generalization, where the generalization between two stimuli (i.e., probability of confusion) decays as an exponential function of their distance in "psychological space". **b**) An illustration of a psychological space centered on previously encountered stimuli $\mathbf{x}$, where the blue region indicates a circular consequential region of unknown size (darker colors indicate higher likelihood). For some new stimuli $\mathbf{x}'$, a larger distance from $\mathbf{x}$ corresponds to higher uncertainty about both belonging to the same consequential region, producing lower levels of generalization. **c**) Tversky's (1977) contrast model, where the overlap indicates features shared by both x and y, while features exclusive to x are on the left and features exclusive to y are on the right. **d**) Tenenbaum and Griffiths (2001) show that a Bayesian inference model of concept learning can unify both Shepard's law of generalization and Tversky's contrast model. Based on previous experiences of a class of items (black crosses), potential consequential regions (ellipses) are considered as hypotheses (h), where the probability of a hypothesis $p(h)$ is inversely proportion to it's size (Bayesian size principle).

$\mathcal{G}(\mathbf{x}, \mathbf{x}')$ is the generalization gradient indicating the probability of producing the response associated with stimulus $\mathbf{x}$ when $\mathbf{x}'$ is presented (Fig. 1.3a). This exponential generalization function is able to explain behavior across an impressive range of data sets (Attneave, 1950; Cheng, 2000; Ekman, 1954; Guttman & Kalish, 1956; Miller & Nicely, 1955), with diversity in both stimuli (e.g., phonemes, shapes, and colors) and subject (e.g., humans, pigeons, and bees). Recent work has further grounded the universality of the exponential law of generalization in fundamental properties of efficient coding given information rate limits (Sims, 2018) and as a requirement for any perceptual system that is invariant to shifts, stretches, and rotations (S. A. Frank, 2018).

An important feature of the generalization gradient is the notion of distance between stimuli $d(\mathbf{x}, \mathbf{x}')$, where it is assumed that stimuli are represented as coordinates in an internal *psychological space* and that distance is the inverse of similarity (Fig. 1.3b). We discuss this spatial representation of similarity below, but in plain terms the law of generalization implies that similar stimuli produce similar responses, with an exponential decay of response probability as similarity decreases.

1.2.2 Representations of similarity

Similarity is a central theoretical concept in psychology (Goldstone & Son, 2012; Medin, Goldstone, & Gentner, 1993), and is used in a wide variety of contexts, from category learning

(Kruschke, 1993; Nosofsky, 1988b), to semantics (Hills, Jones, & Todd, 2012; Landauer & Dumais, 1997), to the representativeness heuristic (Galesic, Goode, Wallsten, & Norman, 2018; Kahneman & Tversky, 1972). Yet there has been much debate about which metric or representation to use (e.g., Chater & Vitányi, 2003; Hahn, Chater, & Richardson, 2003; Tversky, 1977). Here we discuss influential theories about how to represent similarity, which are fundamentally linked to theories about how knowledge is organized in the brain.

Spatial representations of similarity

Thinking spatially is intuitive. We remember things in terms of places (Dresler et al., 2017; James, 1890; Yates, 2013), describe the world using spatial metaphors (Lakoff & Johnson, 2008; Landau & Jackendoff, 1993), and commonly use concepts like "space" or "distance" in mathematical descriptions of abstract phenomena. Previous theories have proposed that spatial representations have been "exapted" (adapted externally into new domains through evolution) for organizing more abstract forms of conceptual knowledge (Hills, 2006; Hills, Todd, & Goldstone, 2008; Todd et al., 2012). This theory has gained new support from neuroscience, with evidence that grid cells in the entorhinal cortex utilize the same encoding for representing knowledge in both spatial and conceptual domains (Behrens et al., 2018; Constantinescu, O'Reilly, & Behrens, 2016). Grid cells are a type of neuron that encode an individual's location and orientation in space (Hafting, Fyhn, Molden, Moser, & Moser, 2005), the discovery of which won May-Britt and Edvard Moser the Nobel prize in 2014. More specifically, grid cells have also been shown to encode distances in multidimensional feature space (Theves, Fernandez, & Doeller, 2019), providing neurological evidence for a spatial representation of similarity across domains.

Early theories of similarity introduced the notion of a *psychological space*, where stimuli are embedded as geometric coordinates and a measure of distance (commonly Euclidean or Manhattan distance) serves to represent the level of (dis)similarity between stimuli (Ekman, 1954; Torgerson, 1952). These early spatial representations were constructed using subjective ratings of similarity, such that items rated as highly similar were placed in close proximity. Shepard's (1987) law of generalization also relies on a metric representation of similarity. But rather than using similarity ratings, Shepard constructed his psychological space based on the confusability of different stimuli (i.e., responding to stimulus $\mathbf{x}$ when $\mathbf{x}'$ is shown), such that items producing similar responses are embedded in similar locations. More specifically, stimuli embeddings are computed using Multidimensional scaling (MDS; Kruskal, 1964a, 1964b; Shepard, 1962a, 1962b) as a means to a preserve the rank-ordering of the data in a low-dimensional Euclidean space. This monotonic transformation results in a metric space where the same unit of distance in any direction corresponds to the same level of generalization.

Shepard assumed that mental representations of categories or natural kinds implicitly correspond to a *consequential region* in psychological space (Fig. 1.3b). Thus, generalizations result from uncertainty about the distribution of stimuli that define this region (Shepard, 1987). This approach to stimulus generalization has inspired many theories of category learning, where categories (e.g., dog or cat) correspond to different consequential regions and membership in a category is determined on the basis of similarity to previously encountered exemplars (Kruschke, 1992; Nosofsky, 1986). These *exemplar* models also construct a psychological space using either similarity ratings (Nosofsky, 1986) or directly using perceptual features (Kruschke, 1992; Love, Medin, & Gureckis, 2004). A notable deviation from Shepard's original formula is that sometimes generalization is observed to follow a Gaussian function i.e., $\mathcal{G}(\mathbf{x},\mathbf{x}') \propto \exp\left(-d(\mathbf{x},\mathbf{x}')^2\right)$, rather than an exponential function[†] (Nosofsky, 1985, 1988a, 1988b).

Thus, while there is some ambiguity about how to construct a psychological space (similarity judgments, confusability, or perceptual features) or whether the generalization gradient is exponential or Gaussian, the basic formula of a spatial organization of knowledge is supported by neural evidence and has proved to be a useful characterization across multiple domains.

Alternative representations of similarity

One of the main critiques of a spatial representation of similarity comes from Tversky and colleagues (Beals, Krantz, & Tversky, 1968; Tversky, 1977; Tversky & Gati, 1982), who argue that the properties of *symmetry* and the *law of triangle inequality* inherent to any spatial representation are regularly violated by human judgments.

Symmetry is based on the fact that the distance between two points $d(\mathbf{x},\mathbf{x}') = d(\mathbf{x}',\mathbf{x})$ is the same in either direction for any distance metric. An intuitive example of when symmetry is violated is that we can say "an ellipse is like a circle" but not "a circle is like an ellipse". The law of triangle inequality requires that the distance between two points $d(a,b)$ be shorter or equal to a path that also visits a third point c, such that $d(a,b) \leq d(a,c) + d(c,b)$. It is helpful to visualize the points a, b, and c as the vertices of a triangle, where the direct distance between two vertices is always shorter than the sum of the other two sides. A verbal example is that Jamaica is similar to Cuba (because of geography), and Cuba is similar to Russia (because of political history), but Jamaica and Russia are not very similar. Thus, it appears that adding a comparison to Cuba reduces the distance between Jamaica and Russia such

[†] An important caveat is that the exercise of curve fitting is susceptible to identifiablity issues (e.g., the debate between power law vs. exponential learning curves; Myung, Balasubramanian, & Pitt, 2000; Pitt, Myung, & Zhang, 2002). There is also a host of animal learning literature that finds Gaussian instead of exponential generalization gradients (Blough, 1969, 1975; Ghirlanda & Enquist, 2003; Hanson, 1959; Thomas & Bistey, 1964).

that $d(\text{Jamaica}, \text{Russia}) < d(\text{Jamaica}, \text{Cuba}) + d(\text{Cuba}, \text{Russia})$. Thus, contrary to the law of triangle inequality, these critiques suggest mental representations may have inherent structures that cannot solely be captured by a spatial representation of similarity.

Tversky's (1977) solution to these critiques is expressed in his contrast model of similarity (Fig. 1.3c), where the similarity of y to x is a function of their shared and distinct features:

$$S(y,x) = \theta f(\mathcal{Y} \cap \mathcal{X}) - \alpha f(\mathcal{Y} - \mathcal{X}) - \beta f(\mathcal{X} - \mathcal{Y}) \tag{1.7}$$

where $\mathcal{X}$ and $\mathcal{Y}$ are the feature sets of x and y, respectively, and $\theta, \alpha, \beta \geq 0$ are weights for each comparison. Similarity $S(y,x)$ is expressed as the contrast between the shared features $f(\mathcal{Y} \cap \mathcal{X})$ and distinctive features, $f(\mathcal{Y} - \mathcal{X})$ and $f(\mathcal{X} - \mathcal{Y})$.

While the contrast model is commonly considered to be the main alternative to Shepard's metric model of generalization, Tenenbaum and Griffiths (2001) show that a Bayesian inference model of concept learning is able to unite both frameworks. This recasts the problem of generalization as the problem of inferring whether a new stimuli $\mathbf{x}'$ belongs to the same consequential region C as previously observed data $\mathbf{X}$ (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001, Fig; 1.3d). This is performed by averaging over potential hypotheses $h \in \mathcal{H}$ corresponding to different consequential regions that are consistent with the data

$$p(\mathbf{x}' \in C|\mathbf{X}) = \sum_{h:\mathbf{x}' \in h} p(h|\mathbf{x}') \tag{1.8}$$

where $p(h|\mathbf{x}')$ is computed using Bayes' rule, and weighted by the prior $p(h)$ such that more complex hypotheses corresponding to larger consequential regions are less likely (Bayesian size principle; Tenenbaum & Griffiths, 2001). The Bayesian inference framework operates on hypothesis space, which is agnostic to either spatial or feature-based representations of knowledge. Thus, it is able to unify both approaches by producing generalization gradients equivalent to Shepard's (1987) law of generalization, while also encompassing the set-theoretic approach of Tversky's (1977) contrast model.

Extensions of the Bayesian concept learning approach have made notable contributions to understanding how people learn rapidly from sparse data, in tasks ranging from inferring structure (e.g., hierarchically organized tree structures; Kemp & Tenenbaum, 2008, 2009), to learning generative programs (e.g., motor programs for handwritten characters; Lake, Salakhutdinov, & Tenenbaum, 2015; Lake et al., 2017), and causal theories about the world (Griffiths & Tenenbaum, 2009). While powerful, this approach also suffers from issues of computational complexity, since the set of potential hypotheses $\mathcal{H}$ could be prohibitively large or infinite (for continuous features). Thus, the sum in Eq. 1.8 may be difficult or impossible to compute in practice, although a sampling approach can provide approximate solutions within realistic com-

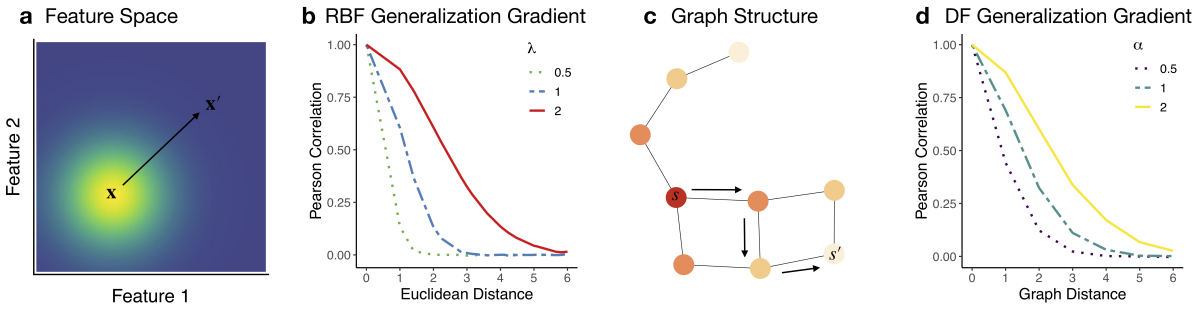

Fig. 1.4 Kernel generalization gradients. **a**) The radial basis function (RBF) kernel defines a Gaussian generalization gradient over feature space, where colors indicate the decay of generalization for increasing distances $d(\mathbf{x}, \mathbf{x}')$. **b**) The RBF generalization gradient computed on a 8x8 feature space, where larger distances correspond to lower assumed correlations of reward. The length-scale λ governs the rate of decay. **c**) A graph structure used in Experiment 7, where the connections between nodes define their relationship. Lighter colors indicate lower levels of generalization relative to node s. **d**) The diffusion kernel (DF) generalization gradient, where larger graph distances (shortest path) correspond to lower assumed correlations of reward. The diffusion parameter α governs the rate of decay.

putational demands (e.g., Dasgupta, Schulz, & Gershman, 2017). Nevertheless, this highlights the need for a theory of similarity that can describe both metric and structured representations of knowledge.

1.2.3 Kernel representations of similarity

Having described the advantages and disadvantages of different types of similarity, we propose a theory of kernel similarity for describing both spatial and structured representations of knowledge. This approach operates on familiar principles of similarity-based generalization, yet manages to evade Tversky's (1977) critiques of symmetry and the law of triangle inequality. Additionally, kernel similarity provides a bridge from generalizing about discrete stimuli or categories, to inferring functions relating a continuous range of potential options to expectations of reward. A small number of experiences can be generalized to produce a value function across the space of possible options. This inferred value function can be used navigate large problem spaces by directing search towards locations that seem promising.

A kernel function $k(\mathbf{x}, \mathbf{x}')$ expresses a similarity metric for all pairs of points in some data set $\mathbf{X}$. Mathematically, the kernel function is equivalent to defining an implicit feature map $\Phi = \phi(\mathbf{X})$ mapping the data onto a Hilbert space $\mathbf{X} \rightarrow \mathcal{H}_K$ (Steinwart, Hush, & Scovel, 2006). A Hilbert space possesses a potentially infinite number of dimensions, yet it never needs to be explicitly computed, serving merely as a useful mathematical abstraction[†]. In practice, the kernel function only requires computing the inner product between pair-wise data points

[†]Another perspective is that Shepard's psychological space maps the data onto a lower dimensional Euclidean space, while a kernel allows us to implicitly represent the data in a higher dimensionality.

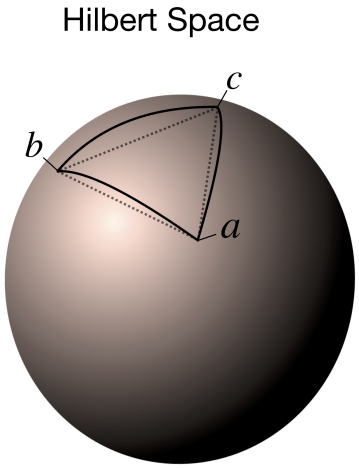

Fig. 1.5 Hilbert space representation of three observed stimuli a, b, and c. The RBF kernel maps all observations to the surface of a unit sphere in Hilbert space. While measuring distance in the original feature space corresponds to tracing a path along the surface of the unit sphere (solid lines), the kernel similarity corresponds to moving through the sphere (dotted lines). Thus, triangle inequality is never violated, because the shortest path between two points is through the sphere, and never involves a detour intersecting a third point on the surface of the sphere. Adapted from Jäkel et al. (2008b).

(Mercer, 1909):

$$k(\mathbf{x},\mathbf{x}') = \left\langle \phi(\mathbf{x}), \phi(\mathbf{x}') \right\rangle \tag{1.9}$$

This is known as the 'kernel trick', and one of the main motivation behind the widespread use of kernel methods in machine learning (Boser, Guyon, & Vapnik, 1992; Schölkopf, Smola, & Müller, 1997) and psychology (Jäkel, Schölkopf, & Wichmann, 2009), since it allows linear methods to describe non-linear relationships (Schölkopf & Smola, 2001). Thus, computing a set of pair-wise comparisons of the stimuli features is sufficient for defining the kernel. Each comparison computes a similarity metric, which is a function of the distance between stimuli in feature space (Fig. 1.4a).

RBF kernel

A common choice of kernel is the radial basis function (RBF), which encodes similarity as a smooth function of the squared euclidean distance between two stimuli $\mathbf{x}$ and $\mathbf{x}'$:

$$k_{RBF}(\mathbf{x},\mathbf{x}') = \exp\left(\frac{||\mathbf{x}-\mathbf{x}'||^2}{2\lambda^2}\right) \tag{1.10}$$

The RBF kernel produces a monotonically decaying generalization gradient as a function of the distance between two stimuli $\mathbf{x}$ and $\mathbf{x}'$, where the length-scale λ encodes the rate of decay (Fig. 1.4b). This representation of similarity is conceptually similar to Shepard's law of generalization (Jäkel, Schölkopf, & Wichmann, 2008a).

Even though the RBF kernel is straightforward to compute using pairwise distance comparisons, Jäkel et al. (2008b) shows that it offers a solution to the problem of triangle inequality. While comparisons of distance in any metric feature space are consistent with the law of triangle inequality, recall that the kernel function involves an implicit mapping onto Hilbert space. Each observation is mapped as a point on the unit sphere in Hilbert space (Fig 1.5). While the shortest path in the original feature space corresponds to a path traced over the surface of the unit sphere (solid line), the shortest distance in Hilbert space is through the sphere (dotted line). Thus, the direct distance between any two points (through the sphere) is always shorter than a detour that traces a path across the surface to visit another point. In other words, the law of triangle inequality is irrelevant for a kernel similarity metric and produces no inconsistencies. Intuitively, we can think about classical forms of spatial similarity representations as constructing a map, which flattens the topology of the world onto a lower dimensional representation. In comparison, kernel similarity creates an implicitly higher dimensional mapping, based only on computing pairwise comparisons.

Diffusion kernel

We can also use a kernel to represent similarity over graph structures (Fig. 1.4c), where relationships are defined based on connectivity rather than their singular features. From social networks to subway maps, graph structures are pervasive in both natural and built environments. Graphs define a more restricted and structured representation of knowledge, where transitions between nodes are only allowed along defined edges. In contrast, any metric space is equivalent to assuming that transitions are unrestricted and symmetrical in all directions.

The diffusion kernel (DF; Kondor & Lafferty, 2002) defines a similarity metric $k(s,s')$ between the nodes of a graph $s_i \in \mathcal{S}$:

$$k_{DF}(s,s') = \exp^{\alpha L} \tag{1.11}$$

where L is the graph Laplacian and α is the diffusion parameter controlling the extent to which generalizations "diffuse" along the graph structure. The graph Laplacian ($L = D - A$) captures the connectivity structure of the graph as a function of the adjacency matrix A, where $A_{i,j}$ describes the connection weight between s_i and s_j, and the degree matrix D, where the diagonals describe the degree of each node. Figure 1.4d shows the generalization gradient of the diffusion kernel, where generalizations decay over longer path distance between nodes.

A more detailed treatment of the diffusion kernel is available in Chapter 6, where we show that it is exactly equivalent to the RBF kernel in the limiting case of an infinitely fine lattice graph. Thus, the diffusion kernel can be seen as an extension of the RBF kernel for

generalization over structured domains, where relationships between stimuli are discrete and non-symmetric. Indeed, the diffusion kernel as a similarity metric operating on a graph structure may offer a potential solution to the problem of symmetry posed by Tversky (1977), since nodes can exert asymmetric influences based on the connectivity structure[†].

Additionally, the diffusion kernel has mathematical equivalencies to the Successor Representation (SR; Dayan, 1993) model of reinforcement learning. The SR originated as a means to improve the generalizability of TD learning (Eq. 1.1.2), by decomposing a learned value function into state-specific rewards and a predictive model of state transitions (e.g., moving my queen in a game of chess changes the set of future states that are available). Recently, the SR has been revived as a neurological theory of how the spatial encoding of grid cells in the entorhinal cortex are selectively skewed towards commonly travelled paths (Gershman, 2018b; Momennejad et al., 2017; Stachenfeld, Botvinick, & Gershman, 2014, 2017). Thus, the evidence for a common spatial organization of knowledge in the brain also points towards a representation that is sensitive to the transition structures that define our world. The shortest path is not as the crow flies, but requires navigating the structures of the environment. As a result, these structures define the cognitive map we use to navigate the world (Tolman, 1948), and can be described using a kernel representation of similarity.

1.3 Generalization through function learning

Having provided a brief overview of generalization and representations of similarity, we now turn to the problem of how to efficiently search for rewards in vast problem spaces. We established that traditional models of associative learning fail to make informative predictions about novel situations, and thus looked to theories of stimulus and category generalization for potential mechanisms to infer the value of novel options. However, rather than making individual generalizations for each new situation (e.g., Shepard, 1987), we can use the same principles of similarity-based generalization to make inferences about the relationship between actions and outcomes in the form of a function (Fig. 1.6 Boyan & Moore, 1995; Carroll, 1963). Functions represent candidate hypotheses about the world, for example, how pressure on the gas pedal is related to the acceleration of a car, or how the amount of water influences the growth rate of a plant. Based on a small number of observations, we can rapidly extrapolate beyond our experience, in order to predict which regions of the search space seem likely to yield the desired outcomes.

[†]For example, a central train station shutting down may have huge impacts on the flow of passengers at other local stations, whereas a local station shutting down would have relatively little influence on the central station.

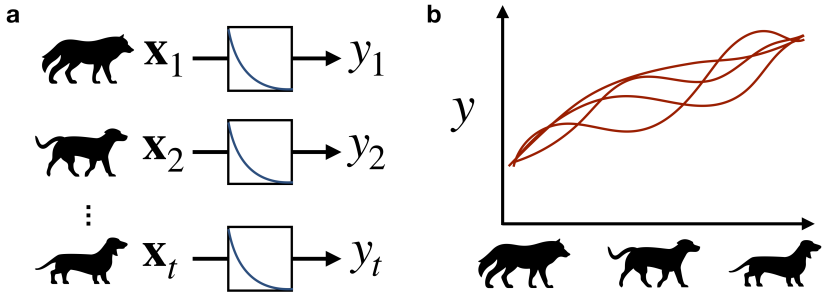

Fig. 1.6 Stimuli generalization and function learning. **a**) Stimuli generalization makes individual predictions for each new stimuli by generalizing from previous observations (e.g., using an exponential generalization gradient). **b**) Function learning provides a mechanism for learning a function mapping inputs to outputs, providing predictions about an entire range of possible inputs (e.g., the size of a dog and excitability). The GP model uses a kernel function to learn a distribution over functions (red lines indicate sampled functions), where each function represents a candidate hypothesis of the relationship between inputs and outputs.

The question of how people explicitly interpolate or extrapolate from limited data has been the central focus of research on human function learning, which has traditionally studied predictions on continuous spaces (e.g., the relationship between two variables; Brehmer, 1974; Busemeyer, Byun, DeLosh, & McDaniel, 1997; Carroll, 1963; Koh & Meyer, 1991). This line of research has revealed inductive biases that guide learning (Kalish, Griffiths, & Lewandowsky, 2007; Kwantes & Neal, 2006; E. Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2017) and which types of functions are easier or harder to learn (E. Schulz, Tenenbaum, Reshef, Speekenbrink, & Gershman, 2015).

Earlier theories of function learning used *rule-based* models that assumed a specific parametric family of functions (e.g., linear or exponential; Brehmer, 1974; Carroll, 1963; Koh & Meyer, 1991). However, the rigidity of rule-based learning struggled to account for order-of-difficulty effects in interpolation tasks (McDaniel & Busemeyer, 2005), and could not capture the biases displayed in extrapolation tasks (DeLosh, Busemeyer, & McDaniel, 1997). An alternative approach relied on *similarity-based* learning, using connectionist networks to associate observed inputs and outputs (DeLosh et al., 1997; Kalish, Lewandowsky, & Kruschke, 2004; McDaniel & Busemeyer, 2005). The similarity-based approach is able to capture how people interpolate, but fails to account for some of the inductive biases displayed in extrapolation and in the partitioning of the input space. In some cases, hybrid architectures were developed to incorporate rule-based functions in a associative framework (e.g., Kalish et al., 2004; McDaniel & Busemeyer, 2005) in an attempt to gain the best of both worlds.

More recently, a theory of function learning based on Gaussian Process (GP) regression was proposed to unite both accounts (Griffiths, Lucas, Williams, & Kalish, 2009; Lucas, Griffiths, Williams, & Kalish, 2015), because of its inherent duality as both a rule-based and a similarity-based model. GP regression is a non-parametric method for performing Bayesian function learning, which has successfully described human behavior across a range of traditional function

learning paradigms (Griffiths et al., 2009; Lucas et al., 2015), and can account for compositional inductive biases in human learning (e.g., combining periodic and long range trends; E. Schulz, Tenenbaum, et al., 2017).

1.3.1 Generalization using Gaussian process regression

Here, we use the GP as a model of how people generalize by learning an implicit value function over the space of possible options. This corresponds to predictions about novel stimuli by extrapolating or interpolating from existing data. The key principle of the GP model is a kernel function, which provides inductive biases encoding expectations about the smoothness of functions (i.e., that similar inputs produce similar outputs). This is exactly the type of kernel function we described in the previous section. While we have shown that a similarity-based approach to generalization has been widely applied in stimulus and category generalization, here we apply it as a model of function learning to guide the efficient search for rewards.

We provide a formal description of the GP function learning model in Chapter 2, but in brief, the GP defines a multivariate normal distribution over functions

$$f \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right), \tag{1.12}$$

where each function f represents a candidate hypothesis for the relationship between stimulus $\mathbf{x}$ and reward y (Fig. 1.6b). A common convention is to set the mean of the distribution to zero $m(\mathbf{x}) = 0$, such that the GP is entirely defined by the kernel (without loss of generality).

Conditioned on observations of reward $\mathcal{D}_t = \{\mathbf{X}_t, \mathbf{y}_t\}$, we can compute a posterior distribution corresponding to predictions about reward. Each prediction is also accompanied by an estimate of the underlying uncertainty, which provides important guidance about which regions of the search space would be most informative to sample. In 8 experiments throughout this thesis, we show that the GP function learning model provides accurate predictions of how people explore large problem spaces, where both predictions of reward and representations of uncertainty play an essential role in guiding efficient search behavior.

In this chapter we focused on providing theory-driven motivations for a kernel representation of similarity, and showed how it resolves challenges faced by other models of generalization and offers a rich set of connections to psychological and neurological theories of how knowledge is organized in the brain. In Chapter 2, we show how the GP model fits into the reinforcement learning framework, where the computational challenges of the exploration-exploitation dilemma motivates the need for efficient models of learning.

1.4 Thesis roadmap

Over the past 100 years, the study of human learning has evolved from observing cats break out of strange boxes (Thorndike, 1911), to understanding the neural substrate of reward prediction errors (Schultz et al., 1997), to using these principles to build artificial intelligence capable of beating the best human players in skillful games such as Go (Silver et al., 2016). While the trajectory of this line of research is certainly impressive, a noticeable gap still exists between the efficiency of human and machine learning. How are people able to rapidly adapt to new situations and learn from so few examples (Lake et al., 2017)? This is the central question of this thesis, where we propose a theory of how generalization guides exploration towards promising options and is responsible for efficient learning in large problem spaces.

Research on stimulus-response learning (Shepard, 1958) and categorization in psychology has defined generalization as a function of similarity (Shepard, 1987), where similar stimuli produce similar responses. This principle of similarity-based generalization can be formalized as a kernel function (Jäkel et al., 2008b), which allows us to not only make pair-wise generalizations, but can make predictions about a potentially infinite set of possible stimuli by approximating a function over the input space (Griffiths et al., 2009). This function learning approach to generalization uses Gaussian Process (GP; Rasmussen & Williams, 2006) regression, which also resolves competing theories of rule-based and similarity-based function learning through virtue of an inherent duality (Lucas et al., 2015). Additionally, the GP approach not only makes predictions about expected values of novel stimuli, but also makes predictions about the underlying certainty. This is a crucial component of the efficient nature of human learning, since building a representation of the uncertainty in the world allows for actively and directedly exploring the parts of the world expected to be most informative. The structure of this thesis is as follows:

Chapter 2 provides a tutorial on the computational models used throughout the thesis. It is designed to provide the reader with an intuitive yet sufficiently technical account of the Reinforcement Learning (RL; Sutton & Barto, 2018) framework, and more specifically the GP function learning model used in subsequent chapters.

Chapter 3 presents three experiments studying how people search for rewards in problem spaces far larger than the horizon of search. Being unable to exhaustively explore all options, participants were instructed to use the spatial correlation of rewards to generalize and infer which options were the most promising to explore. Experiments 1 and 2 used generated environments with varying levels of spatial correlation in rewards, while Experiment 3 used natural environments defined by the yield of various crops, where spatial correlations arise organically, but at heterogeneous levels both within and between environments. We used computational models to predict participant choices, where we show that the GP function

learning model provides the best account of behavior and is able to simulate human-like learning curves.

Chapter 4 presents a field experiment studying how generalization and exploration changes over the lifespan. The goal of this chapter is to add clarity to a popular theory that children have "higher temperature sampling", which cools off over time, leading to a reduction of random exploration. We use a computational model where the specific components are distinctively aligned to different theories about the sources of variability in sampling behavior: the extent of generalization, the eagerness of exploration directed towards uncertain options, and the level of random exploration. The results of Experiment 4 show that children do not simply behave more randomly. Instead, we find that they generalize less than adults, but explore the world more eagerly, in a distinctively directed fashion.

Chapter 5 connects theories of search in spatial environments to that of conceptual domains. We present two longitudinal experiments, where participants performed successive search tasks where either spatial or conceptual features predicted rewards. Experiment 5 used a paradigm where participants were shown both spatial and conceptual features (plant stimuli differing in the number of leaves and berries) simultaneous, but only one set of features predicted rewards. We found a task-order effect, where performance was boosted in the conceptual domain after experience in the spatial domain, but not vice versa. Additionally, while our GP function learning model best predicted behavior in the spatial task, it failed to achieve better performance than a non-generalizing RL model in the conceptual task. However, the GP model parameters were correlated across both tasks, suggesting participants who generalized more or who explored more in one domain, also did so in the other. Experiment 6 used a different task design, where participants used arrow keys to modify a single stimuli displaying only spatial (location on a grid) or conceptual features (rotation and number of stripes of a Gabor patch). In both tasks, the GP function learning model was the best predictor of search behavior, produced human-like learning curves, and was able to prediction of participant judgments about expected reward and confidence. While this supports the hypothesis that similar principles govern generalization across domains, we also found intriguing differences, with participants substituting directed exploration for random exploration in the conceptual task.

Chapter 6 extends the theories and models of the previous chapters from metric spaces to structured environments. From social networks to subway lines, many real-world human environments are defined by graph structures, where generalization is better defined by connectivity structure rather than some notion of spatial or feature similarity. We describe a diffusion kernel as a means to perform inference and learn functions over graph structures. The RBF kernel used in previous chapters is equivalent to the diffusion kernel in the special case of an infinitely fine lattice graph. Thus, the diffusion kernel provides a broader framework of human generalization

in both metric and structured spaces, with additional connections to neural theories of predictive coding (Behrens et al., 2018; Stachenfeld et al., 2017). Experiment 7 presents a function learning task where participants are asked to predict the number of passengers at various stations on generated subway maps, where we control the number of observations at other stations. We show that the diffusion kernel accounts for both predictions and confidence judgments. Experiment 8 is a bandit task similar to Experiments 1-6, but where the distribution of rewards are defined by the connectivity structure of a graph. The diffusion kernel predicts participant choices, produces human-like learning curves, and predicts judgments about expected reward and underlying uncertainty of unobserved nodes. Taken together, this paints a rich picture of how generalization is supported by a common organization of knowledge that is not only sensitive to spatial or feature similarity, but also encodes a predictive map of the transition dynamics that define structured environments.

Lastly, Chapter 7 provides concluding discussions about the work presented and proposes yet unanswered questions to be tackled in future research. These include exploring the social dimension of learning, the difference between description and experience, and building a process model of generalization by accounting for selective attention, working memory, and reaction times.

Chapter 2

A Model of Human Generalization

This chapter provides a tutorial on the methods and computational models used throughout the thesis. We start with an introduction to the reinforcement learning (RL; Sutton & Barto, 2018) framework and show how optimal solutions are only obtainable in limited cases under restrictive assumptions. This motivates the need for efficient learning strategies that balance the exploration-exploitation dilemma, where we use the multi-armed bandit problem as an experimental paradigm for studying human learning.

We first describe common computational solutions that approximate the optimal solutions using tabular methods from Dynamic Programming, yet still require exhaustive exploration of the entire search space. Because this scales poorly to larger problem spaces, an alternative approach is to use value function approximation to learn a global value function.

These two computational approaches map onto two classes of models we use to describe human learners. The option learning model is based on classic theories of associative learning and resembles tabular methods by learning independent value representations for each option. The function learning model is informed by theories of how humans explicitly learn functions and uses Gaussian Process (GP) regression as a method for generalizing previous observations of reward to a potentially infinite set of unexplored possibilities.

The option learning and function learning models are combined with a set of sampling strategies that transform the predictions of expected reward and the underlying uncertainty into a probabilistic choice framework. These sampling strategies represent different computational heuristics for navigating the exploration-exploitation dilemma, where we show that "optimism in the face of uncertainty" provides a remarkably effect strategy for learning.

2.1 Introduction

Consider the problem of choosing where to go for lunch. Should you exploit your usual option by going to the café you always go to? This produces a predictable yet rewarding outcome, but forgoes any opportunity for learning. Or should you explore by visiting a newly opened restaurant? This might lead to a pleasant or unpleasant surprise, although it will certainly be informative for future decisions. The key to efficient learning and adaptive behavior is to balance the two competing goals of exploration and exploitation. This frames the *exploration-exploitation dilemma*, which is a central problem in RL. Because optimal solutions in RL are generally unavailable, an optimal solution to the exploration-exploitation dilemma is likewise non-existent for all but limiting cases (e.g., infinite time). How do people balance the dual demands of exploration and exploitation? And what strategies do human learners use to navigate vast problem spaces where the limited horizon for search is eclipsed by the number of possible options to choose from?

We begin by describing the reinforcement learning (RL; Sutton & Barto, 2018) framework for providing computational solutions to problems where learning occurs through interactions with the environment (Fig. 2.1). Optimal solutions in RL are only available under restrictive conditions, although approximations based on *tabular methods* are commonly used. Nevertheless, tabular methods require multiple visits to all possible states in the problem space, making them poorly suited to real world problems where the number of possible states or actions vastly outnumbers the horizon for search. An alternative method for large problem spaces is to approximate a global value function across all possible states. This is known as *value function approximation* and allows for predictive generalization about the value of unexplored states by interpolating or extrapolating from a limited set of observation.

These two computational approaches provide a useful distinction that map onto two classes of models we use to describe human learners. The *option learning* model is based on classic theories of associative learning and learns independent value representations for each option, similar to tabular methods. The *function learning* model has equivalencies to commonly used neural network implementations of value function approximation, but is also informed by psychological theories about how humans learn explicit functions and generalize about novel stimuli. Combined with a variety of heuristic sampling strategies, these models provide a probabilistic choice framework that allows us to model human search behavior in a wide variety of settings, where generalization plays a crucial role in guiding exploration towards promising options.

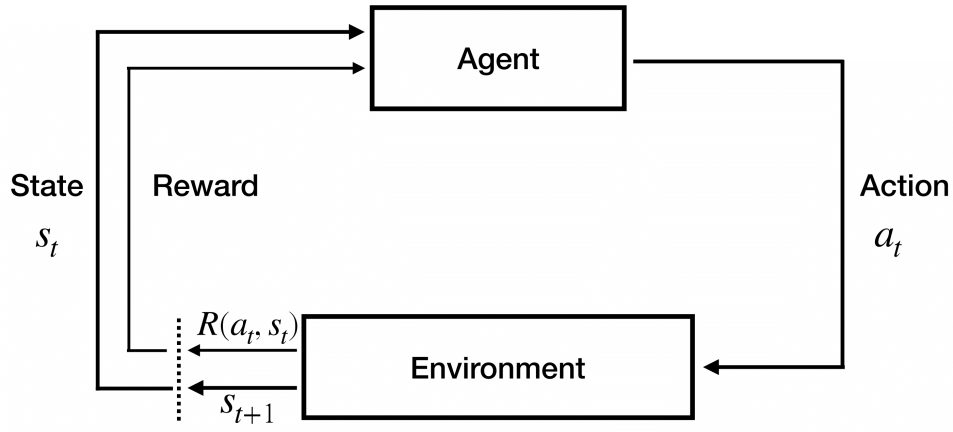

Fig. 2.1 Reinforcement Learning framework adapted from Sutton and Barto (2018).

2.2 Reinforcement learning

The goal in RL is one of acquiring rewards through discovery. Much like an infant learns by experimenting in the world without an explicit teacher (Oudeyer, Kaplan, Hafner, & Whyte, 2005; Riedmiller et al., 2018), an RL agent must be able to effectively explore the space of possibilities and carefully balance the competing requirements of acquiring information and acquiring immediate rewards. Here we present an overview of the RL framework and the computational challenges that are posed. Table 2.1 provides a summary of the nomenclature for convenience.

2.2.1 Markov decision process

Let us consider the problem of an agent (biological or machine) navigating a set of states $s \in \mathcal{S}$, for instance by traveling along a subway network or preparing a meal. The agent selects an action $a \in \mathcal{A}$ at each point in time t (e.g., boarding a train or chopping onions), which results in a new state s' with probability $P(s'|s,a)$. Each state-action pair corresponds to some reward $R(s,a)$, which is typically learned through feedback from the environment. This defines a Markov decision process (MDP) where the agent's goal is to learn a policy π over actions that maximizes the cumulative reward. We can write this in terms of a value function $V_\pi(s)$:

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{s'} P(s'|s,a) \left[R(s',a) + \gamma V_\pi(s') \right] \tag{2.1}$$

Table 2.1 Summary of reinforcement learning terminology

Variable	Description
$\mathcal{S}$	A set of states. The environment is in some state $s \in \mathcal{S}$ at each time point.
$\mathcal{A}$	A set of actions. The agent selects action $a \in \mathcal{A}$ at each time point.
$P(s'\|s,a)$	Transition probability. The probability of ending up in state s' when performing action a from the current state s.
$R(s,a)$	Reward function. The immediate rewards for the state-action pair (s,a).
γ	Temporal discount parameter. In the range $0 \geq \gamma \geq 1$, where smaller values discount long-term future gains more heavily, relative to short-term gains.
$V(s)$	Value function. The expected future discounted returns for state s.
π	A policy. Defines the probability distribution over actions $\pi(a\|s)$ for any state s.

For any state s, the value function $V_\pi(s)$ defines an expectation over all possible actions and all possible outcomes, corresponding to the discounted future rewards[†] that the agent can expect to earn following policy π. The value function is dependent on policy π, which defines a probability distribution over actions $\pi(a|s)$. Intuitively, the policy can be interpreted as an action plan for how the agent should behave in any given state.

Theoretically, we can define an optimal value function $V_*(s)$ using the Bellman (1957) equation:

$$V_*(s) = \max_a \sum_{s'} P(s'|s,a) \left[R(s,a) + \gamma V_*(s') \right] \tag{2.2}$$

The Bellman equation utilizes the Markov assumption of the MDP framework (i.e., that a future state depends only on the current state), to define the value of state s based on the immediate rewards $R(s,a)$ and the discounted value of successor states s'. If the optimal value function can be computed, then the optimal policy π_* becomes rather trivial and merely corresponds to selecting the best action in any state.

While the Bellman equation represents a theoretically optimal solution, computational limitations make it intractable except under restrictive conditions for very simple problems. More commonly, methods from dynamic programming known as *tabular methods* are used to compute approximate solutions to the Bellman equation. We first discuss tabular methods as approximation of the optimal solution that may be suitable for problems with small problem

[†]The temporal discount parameter γ determines how short-term gains are valued relative to long-term gains. See Temporal difference learning (section 1.1.2)

spaces, followed by a discussion of the difficulties in scaling them up to larger problems where we cannot afford to exhaustively explore all possibilities.

2.2.2 Tabular methods for small problem spaces

In small problem spaces there are computational solutions that approximate the optimal solution, but still require exhaustive search. These *tabular* solutions are based on methods from Dynamic Programming (DP; Bellman, 1957), which were first proposed by Minsky (1961) for solving RL problems. While computationally simpler than directly solving the Bellman equation, tabular methods may still be inappropriate when it is not possible to experience all possibilities. For instance, Claude Shannon's (1988) conservative lower bound on the number of possible board positions in the game of chess is 10^{120}, where comparatively, the number of atoms in the observable universe are estimated to only be 10^{80}.

Value iteration

Value iteration resembles the recursive form of the Bellman equation by iteratively visiting all states and updating the value function until a stopping policy terminates search once the algorithm has converged on a "good enough" solution. We begin by initializing the value function as $V_0(s) = 0$ for all states, and then repeatedly visiting each state and iteratively updating the value function until convergence:

$$V_{k+1}(s) = \max_{a \in A} \sum_{s'} P(s'|s,a)[R(s,a) + \gamma V_k(s')] \tag{2.3}$$

The are many types of stopping rules, with one common approach based on the *Bellman residual* (Puterman, 1994; Zhang & Zhang, 2005), which defines a small positive value θ and terminates search when the maximal change in the value function is less than θ:

$$\max_{s \in \mathcal{S}} |V_k(s) - V_{k-1}(s)| < \theta \tag{2.4}$$

However, there are no performance bounds on how well V_n approximates the optimal value function V_* (Geist, Piot, & Pietquin, 2017), where n is the number of iterations before the Bellman residual stopping rule is reached. This is because the local updates in Equation 2.3 may result in a value function V_n that is only a *local* optima, and may be substantially inferior to some *global* optimum that has not been found.

Policy iteration

Policy iteration consists of alternating between evaluating a policy and then improving the policy, which repeats until convergence. Policy evaluation is performed by starting with some policy π_0 (often a random policy), and then visiting each state $s \in \mathcal{S}$ in order to iteratively learn a value function V_{π_0} using Eq. 2.3. The policy π_0 is then improved by again visiting each state $s \in \mathcal{S}$ in order to define a new policy based on selecting the best possible action in each state based on $V_{\pi_0}(s)$:

$$\pi_{k+1} = \arg\max_a \sum_{s'} P(s'|s,a)\left[R(s,a) + \gamma V_{\pi_k}\right] \tag{2.5}$$

This process then repeats over n iterations until some policy π_n emerges that may approximate the theoretically optimal π_*. Policy iteration has many caveats, such as requiring a visit to each individual state $2n$ times, and that convergence is not guaranteed. Indeed, the choice of policy greatly influences the value function that is learned, such that a low-value state for one policy, may in fact be a high-value state for some other policy.

Limitations of tabular methods

Value iteration and policy iteration are known as tabular methods because the value of each state is learned independently from other states (Fig. 2.2 left). We can imagine a large lookup table, where each state corresponds to a single value that is individually updated based on observations of reward. Both value iteration and policy iteration provide computational approximations of the theoretically optimal solution (Eq. 2.2), yet still require multiple visits to every state in the problem space. This is simply not feasible for many real world learning environments, where the number of possible states greatly outnumbers what can be experienced in a human lifetime. In these cases, the value of unobserved states must be defaulted to zero or some prior value.

2.2.3 Value function approximation for large problem spaces

In order to scale to problems with large problem spaces, one solution from the RL framework is to approximate a global value function (Fig. 2.2 right). This transforms the problem of learning a large number of independent values for each state into a function learning problem, where the goal is to learn the appropriate function mapping states onto values $f : s \rightarrow V(s)$. We explore different approaches to function learning in Section 2.3.2, while here we can broadly define an approximate value function $\hat{V}$ parameterized by a weight vector $\mathbf{w}$ and the state s at which it is evaluated:

$$\hat{V}(s,\mathbf{w}) \approx V_\pi(s) \tag{2.6}$$

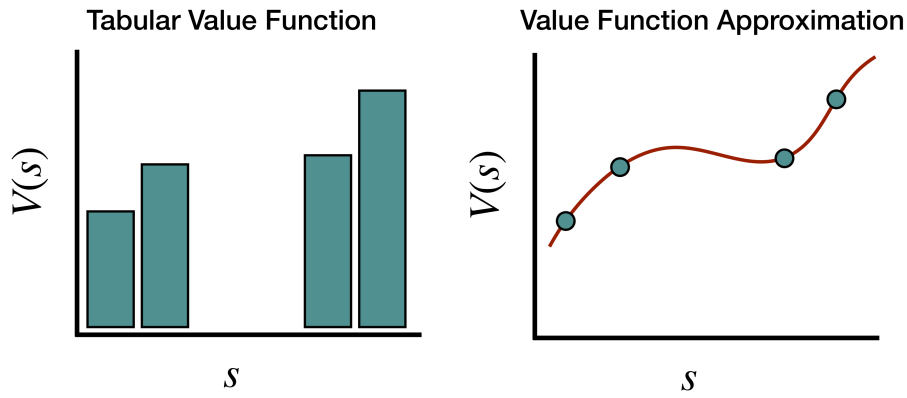

Fig. 2.2 Value functions. Tabular value functions (left) learn the value of each state independently, where unobserved states have a default value of 0 or set to some prior. Value function approximation learns a continuous value function (red line) over the entire state space, where predictions are made about unobserved states through interpolating or extrapolating the value function from the set of observed states (green dots).

The weight vector $\mathbf{w}$ can correspond to the weights of a linear function over features of the state space, or more commonly in machine learning, to the weights of a neural network. Neural networks are increasingly common implementations of RL agents (Mnih et al., 2015; Silver et al., 2016), where whether they be deep or shallow, are specifically a type of universal function approximator (Schölkopf, 2015). The inputs of the neural network are translated into some output by the connection weights, thus corresponding to a function.

A common method for learning the weight vector $\mathbf{w}$ is to use Stochastic Gradient Descent (SGD) to minimize the error between the approximate value function $\hat{V}(s, \mathbf{w})$ and the observed value (Sutton & Barto, 2018). SGD is a type of local optimization with an added component of randomness, which can perform well in very high dimensional environments (i.e., having many weights corresponding to many features), where it is less likely to to become stuck in a local optima (Bottou, 2010). However, the number of function evaluations (i.e., costly observations of reward from the environment) may still be prohibitively large.

One specific approach for approximating a value function is based on using Radial Basis Function (RBF) networks (Poggio & Girosi, 1989, 1990, see Fig. 2.4), which generalize a finite set of observations to nearby regions of the input space. This represents a type of inductive bias that can lead to more efficient function approximation by assuming some level of smoothness of the value function. We discuss this in Section 2.3.2, where we show that GP regression using a RBF kernel learns an equivalent value function, but has the added advantage of making predictions about the underlying uncertainty. Representations of uncertainty are necessary for *directed exploration* strategies that prioritize exploring highly uncertain regions of the state space, leading to more efficient learning.

2.2.4 Multi-armed bandit problem

While RL provides a computational framework for a very general class of problems, we now narrow our focus to the multi-armed bandit (MAB) problem (Robbins, 1952; Wald, 1947) as a simpler subclass of RL problems. In the MAB framework, we can effectively ignore states and state transitions to focus on selecting actions that produce good rewards. Yet, even in this simpler class of problem, optimal solutions are still intractable under finite time horizons[†] (Lusena, Goldsmith, & Mundhenk, 2001; Reverdy, Srivastava, & Leonard, 2014).

Named after a colorful metaphor for a slot machine (i.e., a one-armed bandit), the MAB problem imagines an agent in front of a row of slot machines. Here, actions correspond to the agent selecting one of the available options $\mathbf{x} \in \mathbf{X}$, and acquiring some stochastic reward $R(\mathbf{x})$. Because the rewards are *a priori* unknown, the agent needs to both explore new options to gain information and exploit options known to have high rewards. From choosing which beer to try at a craft brewery or deciding where to go for lunch, the *exploration-exploitation dilemma*, is a ubiquitous feature of human learning.

While many MAB problems are framed such that each option has an independent reward distribution (Gershman, 2018a; Steyvers, Lee, & Wagenmakers, 2009; Wu, Schulz, Gerbaulet, Pleskac, & Speekenbrink, 2019), the experiments in this thesis use environments with correlated rewards, such that similar options yield similar outcomes. This is a common feature of real world environments (e.g., the natural availability of food; see Experiment 3) and is equivalent to the assumption of patchy (Hills et al., 2008; Wilke et al., 2015) or spatially correlated reward distributions (Srivastava, Reverdy, & Leonard, 2015). The correlation of rewards provides traction for generalization, which can be learned and used to efficiently direct search towards promising locations of the search space. Additionally, while many MAB tasks allow for exhaustive sampling of all options, our interest in the problem of generalization motivates the use of short horizons, where the number of possible options is greater than the available number of actions. This makes generalization particularly important and amplifies the difference between efficient and inefficient search strategies.

2.3 Models of human learning

We now turn to the problem of describing and modeling human learners in a MAB task. The distinction between tabular methods and value function approximation map onto two classes of

[†]The Gittins index (Gittins, 1979; Gittins & Jones, 1974) provides an optimal solution for the MAB problem by transforming it into a series of optimal stopping problems for each option, yet requires the prohibitive assumption of infinite time and exponential discounting. Additionally, Lai and Robbins (1985) describe a solution that reaches an asymptotically optimal lower bound on performance, but also in the limit of infinite time.

models we use to describe human learners. Traditional models of human reinforcement learning are similar to tabular methods because they assume humans learn independent associations of rewards for each individual stimulus or option. The Rescorla-Wagner (RW; 1972) and Temporal Difference (TD; Sutton, 1988; Sutton & Barto, 1990) learning models described in Chapter 1 are examples of this approach, while here we introduce a Bayesian variant for modeling human learners (see Bayesian Mean Tracker). In the context of human learning, we call this the *option learning* model.

While value function approximation is commonly implemented using neural networks (Mnih et al., 2015; Silver et al., 2016) in large-scale machine learning problems, here we use a Gaussian Process (GP; Rasmussen & Williams, 2006) implementation as a psychologically intepretable model of human function learning (Griffiths et al., 2009; Lucas et al., 2015; E. Schulz, Tenenbaum, Duvenaud, Speekenbrink, & Gershman, 2016; E. Schulz, Tenenbaum, et al., 2017). GPs also have mathematical equivalencies to neural networks (Neal, 2012), meaning we do not need commit ourselves to either implementation. However, a useful feature of the GP model is that it uses a kernel to encode inductive biases, which we show to be conceptually related to Shepard's law of generalization. We call this the *function learning* model.

Both the option learning and function learning models generate Bayesian predictions of reward for any option $\mathbf{x}$ that can be decomposed into predictions about the expected reward $m(\mathbf{x})$ and the underlying uncertainty $v(\mathbf{x})$. These two components allow us to define a variety of *sampling strategies* that implement heuristic solutions to the exploration-exploitation dilemma. These sampling strategies represent different approaches to balancing the exploitation of high expected rewards $m(\mathbf{x})$ while also exploring optioned deemed to be highly uncertainty $v(\mathbf{x})$. We begin by introducing the option learning and function learning models of human learning, followed by a discussion of potential sampling strategies. Lastly, we also consider several heuristics that do not build a representation of the world (i.e., no predictions about reward), but also capture aspects of human search behavior (Bonawitz, Denison, Gopnik, & Griffiths, 2014; Raichlen et al., 2014).

2.3.1 Option learning

We use a Bayesian Mean Tracker (BMT) as a prototypical model of associative learning that learns independent reward distributions for each option. Formally, the BMT is a type of Kalman filter (Dayan, Kakade, & Montague, 2000; Kruschke, 2008), which originated as a method for controlling dynamic systems (Kalman, 1960), but can be understood as a Bayesian extension of the traditional Rescorla-Wagner (1972) model (Gershman, 2015). The BMT differs from the standard Kalman filter by making the simplifying assumption that the rewards associated with

each option remain constant over time (Speekenbrink & Konstantinidis, 2015). This assumption allows us to focus on how people generalize over the space of possible options, by omitting complex temporal dynamics.

The BMT assumes a normal prior distribution of rewards $\mathcal{N}(m_{j,0}, v_{j,0})$ computed independently for each option j, where $m_{j,0}$ is the prior mean and $v_{j,0}$ is the prior variance. Given some observations $\mathcal{D}_{t-1} = \{\mathbf{X}_{t-1}, \mathbf{y}_{t-1}\}$ of rewards $\mathbf{y}_{t-1}$ for inputs $\mathbf{X}_{t-1}$, the BMT learns the rewards of each option independently by computing independent posterior distributions for the mean μ_j for each option j

$$P(\mu_j|\mathcal{D}_{t-1}) = \mathcal{N}(m_{j,t-1}, v_{j,t-1}) \tag{2.7}$$

In each iteration, the posterior mean $m_{j,t}$ and variance $v_{j,t}$ for the option j chosen at time t are updated based on the prediction error:

$$m_{j,t} = m_{j,t-1} + \delta_{j,t} G_{j,t} \left[y_{j,t} - m_{j,t-1} \right] \tag{2.8}$$

$$v_{j,t} = \left[1 - \delta_{j,t} G_{j,t} \right] v_{j,t-1} \tag{2.9}$$

where $\delta_{j,t} = 1$ if option j was chosen on trial t, and 0 otherwise. Additionally, $y_{j,t}$ is the observed reward for option j at time t, and $G_{j,t}$ is the Kalman gain, which is defined as:

$$G_{j,t} = \frac{v_{j,t-1}}{v_{j,t-1} + \theta_\varepsilon^2} \tag{2.10}$$

where θ_ε^2 is the error variance parameter. Intuitively, the estimated mean of the chosen option $m_{j,t}$ is updated based on the difference between the observed value y_t and the prior expected mean $m_{j,t-1}$ (i.e., prediction error), scaled by the Kalman gain $G_{j,t}$. At the same time, the estimated variance $v_{j,t}$ is reduced by a factor of $1 - G_{j,t}$, which is in the range $[0, 1]$. The error variance (θ_ε^2) can be interpreted as an inverse sensitivity, where smaller values result in more substantial updates to the mean $m_{j,t}$, and larger reductions of uncertainty $v_{j,t}$.

Similar to the tabular methods described above, the BMT uses repeated interactions with the environment to learn about the rewards of each option independently. Thus, predictions about unobserved options are defaulted to some prior (Fig. 2.5a). However, one distinction is that the BMT computes a Bayesian posterior distribution, making predictions about both the expected mean and the estimated uncertainty. Repeated observations about an option reduce the uncertainty, reflecting higher confidence over the course of learning.

2.3.2 Function learning

The function learning model uses Gaussian Process (GP; Rasmussen & Williams, 2006) regression as a model of how people learn an implicit value function over the space of possible options. The GP model is informed by research on how people perform explicit function learning (Griffiths et al., 2009; Lucas et al., 2015), and has equivalencies to neural network methods used in large-scale machine learning applications (Neal, 2012; Poggio & Girosi, 1989, 1990).

Like the BMT model described above, the GP also makes Bayesian predictions about reward that can be decomposed into a prediction about the expected reward $m(\mathbf{x})$ and the underlying uncertainty $v(\mathbf{x})$, but where $\mathbf{x}$ can be any point in a continuous input space. Unlike the BMT model, the observations about rewards at input $\mathbf{x}$ also influence predictions about rewards at other inputs $\mathbf{x}'$ based on the kernel similarity $k(\mathbf{x},\mathbf{x}')$. Thus, the GP uses function learning as a mechanism for continuous generalization across the space of possible inputs, making predictions about a potentially infinite set of possibilities from a small set of observations.

In order to introduce the mathematics of the GP regression framework, we first begin with the traditional weight-space view of regression. This approach requires making parametric assumptions about the type of function to be learned, which are often inflexible and unforgiving in the case of a mismatch with the environment. We transition to the function-space view of regression used in the GP framework, which is capable of learning any stationary function by matching the complexity of the function to the complexity of the data.

Weight-space: linear regression

Let us consider the simple example of learning a function f from a set of observations $\mathcal{D} = \{\mathbf{X},\mathbf{y}\}$, where $\mathbf{x}_i \in \mathbf{X}$ are the inputs and $y_i \in \mathbf{y}$ are the outputs. For example, imagine we are making a soup, and want to learn how the quantities of each ingredient contribute towards the overall taste. The input vector $\mathbf{x}_i$ represents a specific set of features (e.g., the amount of each ingredient) that we will use to predict the output y_i (e.g., the taste). We can write this function as:

$$y_i = f(\mathbf{x}_i) + \varepsilon \tag{2.11}$$

where we assume Gaussian noise $\varepsilon \sim \mathcal{N}(0,\sigma_\varepsilon^2)$ on each observation. The typical weight-space view of function learning uses a linear combination of weights such that

$$y_i = \mathbf{w}^\top \mathbf{x}_i + \varepsilon \tag{2.12}$$

where $\mathbf{w} = [w_0, w_1, \cdots, w_d]$ are the weights of the model and we redefine $\mathbf{x}_i = [1, \mathbf{x}_i]$ by appending a value of 1 for the intercept term (w_0). The maximum likelihood estimate (MLE) of the weights can be found by minimizing the Residual Sum of Squares (RSS):

$$RSS(\mathbf{w}) = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = ||\mathbf{y} - \mathbf{X}\mathbf{w}||^2 \tag{2.13}$$

where $\hat{y}_i$ is the model prediction for input $\mathbf{x}_i$ based on $\mathbf{w}^\top \mathbf{x}_i$, and where predictions for each of the n observations in $\mathcal{D}$ are given by $\mathbf{X}\mathbf{w}$. A simple analytic solution is available through the Moore-Penrose pseudoinverse (Penrose, 1955):

$$\hat{\mathbf{w}} = \left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top \mathbf{y} \tag{2.14}$$

where $\hat{\mathbf{w}}$ is the maximum likelihood estimate (MLE)[†] that minimizes the prediction error defined in Eq. 2.13. This defines a simple format for learning any linear function (Fig. 2.3a). However, not all functions are linear. Imagine a function mapping the happiness of child to the amount of force used to push them on a swing, or your own happiness as a function of the amount of pizza you eat. A linear function would assume that the highest levels of happiness would be on either extreme, of no force or maximum force, or no pizza or infinite pizzas. The reader is not recommended to try this at home.

The weight-space solution to learning non-linear functions proposes using a feature map to transform the input space $\mathcal{X}$ into a feature space $\Phi = \phi(\mathbf{X})$. One common example is the polynomial kernel where $\phi(\mathbf{x}_i) = 1 + \mathbf{x}_i + \mathbf{x}_i^2 + \cdots + \mathbf{x}_i^k$, with k representing the order of the largest polynomial (Fig. 2.3b). Thus, we can substitute $\mathbf{x}_i$ with the feature mapped $\phi(\mathbf{x}_i)$ in equation 2.12:

$$y_i = \mathbf{w}^\top \phi(\mathbf{x}_i) + \varepsilon \tag{2.15}$$

The use of a feature map in the weight-space linear regression framework makes parametric assumptions about the class of function being learned. With the case of the polynomial kernel, the choice of k makes rigid assumptions about the complexity of the function to be learned. A mismatch in choosing either an overly complex or overly simple functional form can lead to inaccurate representations of the data and produce poor predictions about unobserved inputs

[†]Linear regression can also be extended to the Bayesian framework by assuming a prior over weights, for instance a Gaussian prior $p(\mathbf{w}) = \mathcal{N}(0, \Sigma)$. In this case, the likelihood can be described as $p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{X}^\top \mathbf{w}, \sigma_\varepsilon^2 \mathbf{I})$, and the posterior distribution over weights is given by $p(\mathbf{w}|\mathbf{y}, \mathbf{X}) = \mathcal{N}\left(\frac{1}{\sigma_\varepsilon^2} \mathbf{A}^{-1} \mathbf{X}\mathbf{X}^\top, \mathbf{A}^{-1}\right)$, where $\mathbf{A} = \Sigma^{-1} + \sigma_\varepsilon^{-2} \mathbf{X}\mathbf{X}^\top$. In the Bayesian framework, finding the right weights for the model is equivalent to performing Bayesian inference. See Williams (1998) and Chapter 3 of Bishop (2006) for more details.

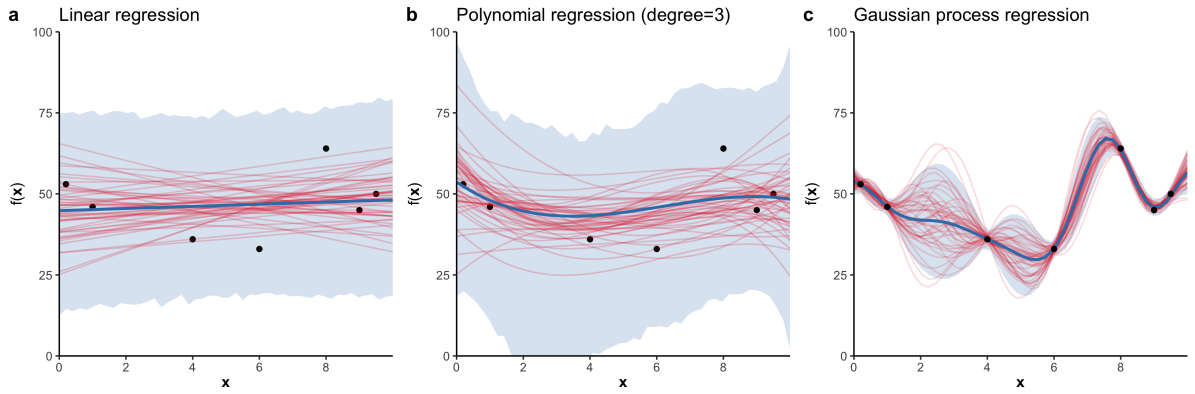

Fig. 2.3 Types of regression. **a**) Linear regression using the Bayesian framework. **b**) Regression with a polynomial kernel of degree 3. **c**) Gaussian process regression with a RBF kernel. Each black point is an observation, while the blue lines indicate the posterior mean and the blue ribbon indicates the 95% confidence interval. The red lines are examples of candidate functions computed by sampling from the posterior distribution of weights (a,b) or by sampling directly from the posterior distribution over functions (c).

(Pitt & Myung, 2002). We now turn to the GP regression framework where the complexity of the function is directly determined by the complexity of the data.

Function-space: Gaussian Process regression

A Gaussian Process (GP) defines a distribution over functions $f : \mathcal{X} \rightarrow \mathbb{R}^n$ that map the input space $\mathcal{X}$ to real-valued scalar outputs (Rasmussen & Williams, 2006). Each function f can be modelled as a random draw from a multivariate normal distribution:

$$f \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right), \tag{2.16}$$

where $m(\mathbf{x})$ is a mean function specifying the expected output of the function given input $\mathbf{x}$, and $k(\mathbf{x}, \mathbf{x}')$ is a kernel (or covariance) function specifying the covariance between outputs. A common convention is to set the mean function to zero, such that GP prior is completely defined by the kernel (without loss of generality).

We use the Radial Basis Function (RBF) kernel to model the covariance between points as an exponentially decreasing function of their squared Euclidean distance:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\lambda^2}\right) \tag{2.17}$$

The length-scale parameter λ modifies the rate of generalization decay, with larger λ-values corresponding to slower decays, generalization over larger distances, and smoother functions. As $\lambda \rightarrow +\infty$, the RBF kernel assumes functions approaching linearity, whereas as $\lambda \rightarrow 0$, there

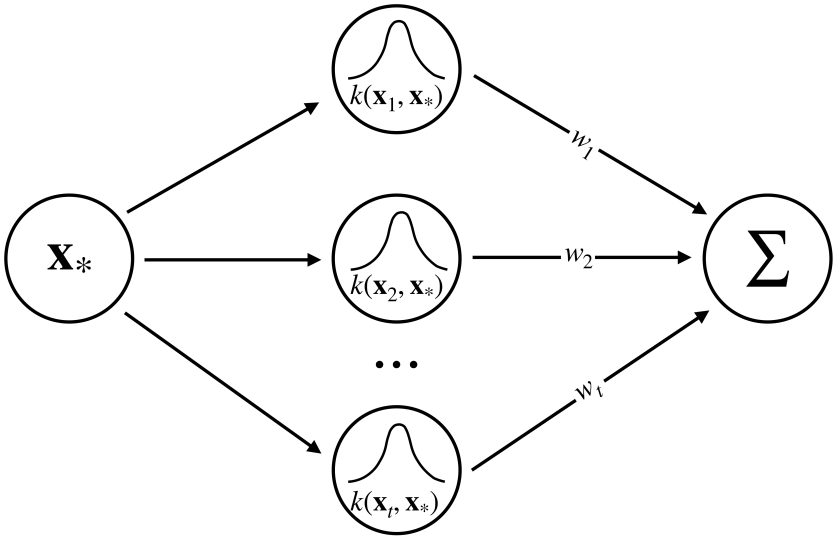

Fig. 2.4 RBF network (adapted from Jäkel et al., 2008a). The posterior mean of the GP can also be interpreted as a weighted sum of the similarities between a target input $\mathbf{x}_*$ and the previous observations $\mathbf{x}_i \in \mathbf{X}_t$. This is equivalent to a single layer RBF network that has featured prominently in machine learning approaches to value function approximation (Sutton & Barto, 2018) and as a theory of the neural architecture of human generalization (Poggio & Bizzi, 2004).

ceases to be any spatial correlation, with the implication that learning happens independently for each input without generalization (similar to traditional models of associative learning).

Given some observations $\mathcal{D}_t = \{\mathbf{X}_t, \mathbf{y}_t\}$ of observed outputs $\mathbf{y}_t$ at inputs $\mathbf{X}_t$, we can compute the posterior distribution $p(f(\mathbf{x}_*)|\mathcal{D}_t)$ for any target input $\mathbf{x}_*$. The posterior is a normal distribution with mean and variance defined as:

$$m(\mathbf{x}_*|\mathcal{D}_t) = \mathbf{k}_{*,t}^\top (\mathbf{K} + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{y}_t \tag{2.18}$$

$$v(\mathbf{x}_*|\mathcal{D}_t) = k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_{*,t}^\top (\mathbf{K} + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{k}_{*,t} \tag{2.19}$$

where $\mathbf{K}_t$ is the $t \times t$ covariance matrix evaluated at each pair of observed inputs, $\mathbf{k}_{*,t} = [k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_t, \mathbf{x}_*)]$ is the covariance between each observed input and the target input $\mathbf{x}_*$, $\mathbf{I}$ is the identity matrix, and σ_ε^2 is the noise variance. Thus, for any input (including unobserved inputs not in $\mathbf{X}_t$), we can make Bayesian predictions about the expected reward $m(\mathbf{x}_*|\mathcal{D}_t)$ and also the level of uncertainty $v(\mathbf{x}_*|\mathcal{D}_t)$ (see Figs. 2.3c and 2.5a).

Generalization as a similarity-weighted sum

We can also return to the weight-space view of regression and rewrite the posterior mean function of the GP as:

$$m(\mathbf{x}_*|\mathcal{D}_t) = \sum_{i=1}^{t} w_i k(\mathbf{x}_i, \mathbf{x}_*) \tag{2.20}$$

where each $\mathbf{x}_i$ is a previously observed input and the weights are collected in the vector $\mathbf{w} = \left[k(\mathbf{X}_t, \mathbf{X}_t) + \sigma_\varepsilon^2 \mathbf{I}\right]^{-1} \mathbf{y}_t$. Intuitively, this means that GP regression is equivalent to a linearly-weighted sum using basis functions $k(\mathbf{x}_i, \mathbf{x}_*)$ to project observed inputs onto a feature space (Rasmussen & Williams, 2006; E. Schulz, Speekenbrink, & Krause, 2018). To generate new predictions for an unobserved input $\mathbf{x}_*$, each output y_i is weighted by the similarity between observed states $\mathbf{x}_i \in \mathbf{X}_t$ and the target input $\mathbf{x}_*$.

Formally, Eq. 2.20 is equivalent to a single layer neural network known as a RBF network (Jäkel et al., 2008a, see Fig. 2.4), which has been proposed as a theory of the neural architecture of generalization in vision and motor control (Poggio & Bizzi, 2004). However, one important difference is that the GP implementation described above also computes uncertainty estimates while the RBF network only predicts expected reward.

2.4 Sampling strategies

Having defined BMT option learning and GP function learning as candidate models for how people learn from experience and make predictions about reward, we now describe various sampling strategies that convert these predictions into valuations of where to sample next. These strategies can be understood as heuristics for trading off between exploring highly uncertain options in order to improve future behavior and exploiting known options for immediate reward. Beyond just predicting which options will produce high-value rewards, efficient learning also requires selectively sampling informative options that reduce our uncertainty about the world.

Both the option learning and function learning models generate normally distributed predictions about each option $\mathbf{x}$, which have mean $m_t(\mathbf{x})$ and variance $v_t(\mathbf{x})$ (Fig. 2.5a). For the Option Learning model, we let $m_t(\mathbf{x}) = m_{j,t}$ and $v_t(\mathbf{x}) = v_{j,t}$, where j is the index of the option characterized by $\mathbf{x}$. Each sampling strategy transforms these Bayesian predictions into a value $q(\mathbf{x})$, which is then put into a softmax choice rule (Eq. 2.27) to produce probabilistic predictions about where participants search next at time $t+1$.

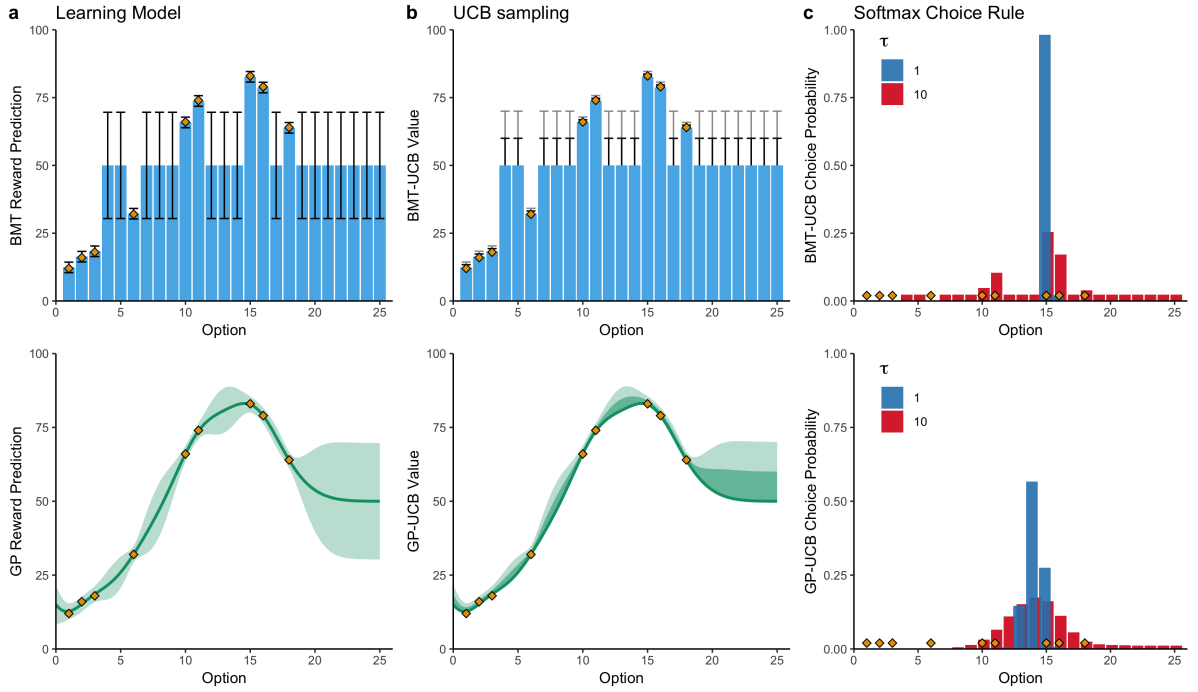

Fig. 2.5 Illustration of learning models, contrasting how the BMT-UCB option learning model (top row) and the GP-UCB function learning model (bottom row) make predictions about search behavior. **a**) Based on a set of observed rewards (orange diamonds) the BMT option learning model makes predictions about the expected reward (blue bars) and the estimated uncertainty (error bars showing 95% CI). Unobserved options are defaulted to a prior mean and prior uncertainty (set to $m_0 = 50$ and $v_0 = 10^2$). In contrast, the GP function learning model uses observed rewards (orange diamonds) to generalize across the search space, with the green line indicating expected rewards and the shaded ribbon indicating the uncertainty (95% CI). **b**) The UCB sampling strategy computes a valuation of each option based on a weighted sum of the expected reward and uncertainty estimate, where β defines how much the reduction of uncertainty is valued relative to exploiting high value rewards. The top panel shows the BMT-UCB strategy where the dark error bars indicate $\beta = 1$ and the light error bars indicate $\beta = 2$. UCB values are uniformly large for unexplored options. The bottom panel shows the GP-UCB strategy, where the dark green ribbon shows the UCB value for $\beta = 1$ and the light green shows $\beta = 2$. The highest UCB values are in unexplored regions near to high observations of reward. **c**) The softmax choice rule transforms the valuations of any sampling strategy into a probabilistic choice framework, where the temperature parameter τ controls the level of undirected sampling noise. The top panel shows the probabilistic predictions of the BMT-UCB model ($\beta = 1$), where the blue bars indicate a low temperature $\tau = 1$ and red indicates a high temperature $\tau = 10$. Note that the highest probabilities tend to be for previously chosen options (indicated by orange diamonds). In contrast, the predictions of the GP-UCB model ($\beta = 1$) are directed towards regions of the search space similar to previous observations of high rewards, which prominently includes unobserved options.

2.4.1 Upper Confidence Bound Sampling

Upper confidence bound (UCB) sampling (Auer, 2002) represents an optimistic approach to the explore-exploit dilemma because it places a positive value on uncertain options. The value of an option is computed using a simple weighted sum:

$$\text{UCB}(\mathbf{x}) = m_t(\mathbf{x}) + \beta\sqrt{v_t(\mathbf{x})}, \tag{2.21}$$

where the exploration factor β determines how much the reduction of uncertainty is valued relative to the exploitation of high-value options. Sometimes referred to as "optimism in the face of uncertainty" (Bubeck & Cesa-Bianchi, 2012), the intuition behind UCB sampling is that search is directed towards highly uncertain options (Fig 2.5b). As uncertainty diminishes through successive observations of reward, the overall uncertainty in the environment diminishes and the expected reward term begins to play a larger role in determining the value of an option.

The optimism of UCB sampling contrasts with economic approaches to determining value under uncertainty. A common utility function in the Mean-Variance Theory (MVT) framework (Markowitz, 1952) is to assign a negative value to uncertainty (Meyer, 2014). However, in the RL domain where optimal solutions are generally unavailable, UCB sampling combined with GP regression is one of the only known algorithms with theoretical guarantees (Srinivas, Krause, Kakade, & Seeger, 2010). The exploration term in Eq. 2.21 guarantees that UCB sampling achieves a constant fraction of the maximum information gain in each sample, thus providing a lower bound on performance. This makes UCB a highly competitive algorithm in Bayesian optimization, which has been successfully applied to robotics (Cully, Clune, Tarapore, & Mouret, 2015), ecology (Gotovos, Casati, Hitz, & Krause, 2013), and protein biology (Romero, Krause, & Arnold, 2013), where search is costly and efficient exploration is valuable.

2.4.2 Pure Exploitation and Pure Exploration

Upper Confidence Bound sampling can be decomposed into a Pure Exploitation component and a Pure Exploration component:

$$\text{PureExploit}(\mathbf{x}) = m_t(\mathbf{x}) \tag{2.22}$$

$$\text{PureExplore}(\mathbf{x}) = \sqrt{v_t(\mathbf{x})} \tag{2.23}$$

These strategies are meant to illustrate the important contribution of both components in the UCB framework. However, equivalent approaches are widely used in a variety of decision making and learning problems. The PureExploit strategy is equivalent to greedy methods used

in decision making (Meder, Nelson, Jones, & Ruggeri, 2018), while the PureExplore strategy implements the information gain maximizing solution (Lindley, 1956; Nelson, 2005) when combined with the GP framework Krause and Ong (2011).

2.4.3 Probability of Improvement.

We can also define a sampling strategy that values an option based on the probability of improving upon the best observed outcome (Močkus, 1975). At any point in time t, the best observed outcome can be described as $\mathbf{x}^+ = \arg\max_{\mathbf{x}_i \in \mathbf{X}_t} m_t(\mathbf{x}_i)$. The probability of improvement (POI) strategy evaluates each option based on high likely it improves on $\mathbf{x}^+$:

$$\begin{aligned} \text{POI}(\mathbf{x}) &= P\left(f(\mathbf{x}) \geq f(\mathbf{x}^+)\right) \\ &= \Phi\left(\frac{m_t(\mathbf{x}) - m_t(\mathbf{x}^+)}{\sqrt{v_t(\mathbf{x})}}\right) \end{aligned} \tag{2.24}$$

where $\Phi(\cdot)$ is the cumulative density function (CDF) of a normal distribution (since both the BMT and GP produce normally distributed posterior predictions). A similar strategy is also used in Gershman et al. (2017) in predicting how people sample novel food combinations.

2.4.4 Expected Improvement

Rather than evaluating an option by the extent of the probability of improvement , the Expected Improvement (EXI) strategy evaluates an option based on *how much* (in the expectation) it promises to be better than the best observed outcome $\mathbf{x}^+$:

$$\text{EXI}(\mathbf{x}) = \begin{cases} \Phi(Z)(m_t(\mathbf{x}) - m_t(\mathbf{x}^+)) + \sqrt{v_t(\mathbf{x})}\phi(Z), & \text{if } \sqrt{v_t(\mathbf{x})} > 0 \\ 0, & \text{if } \sqrt{v_t(\mathbf{x})} = 0 \end{cases} \tag{2.25}$$

where $\Phi(\cdot)$ is the normal CDF, $\phi(\cdot)$ is the probability density function (PDF) of a normal distribution, and $Z = (m_t(\mathbf{x}) - m_t(\mathbf{x}^+))/\sqrt{v_t(\mathbf{x})}$.

2.4.5 Probability of Maximum Utility.

The Probability of Maximum Utility (PMU) uses a form of probability matching (Gaissmaier & Schooler, 2008; Neimark & Shuford, 1959) to sample each option according to the probability that it results in the highest reward of all options in a particular context (Speekenbrink & Konstantinidis, 2015). It can also be understood as an implementation of Thompson sampling

(Gershman, 2018a; Thompson, 1933), where we use Monte Carlo sampling from each option's predictive distributions, and then assign a probability to each option proportional to the number of times it has the highest sampled payoff:

$$\text{PMU}(\mathbf{x}) = P\left(f(\mathbf{x}_j) > f(\mathbf{x}_{i\neq j})\right) \tag{2.26}$$

We compute 1,000 Monte Carlo samples from the posterior distribution of each option, and then evaluate the proportion of times a given option turns out to be the maximum.

2.4.6 Softmax choice rule

All sampling strategies produce a valuation for each option $q(\mathbf{x})$. We then use a softmax choice rule to compute a probability distribution over options

$$p(\mathbf{x}) = \frac{\exp(q(\mathbf{x})/\tau)}{\sum_{j=1}^{N}\exp(q(\mathbf{x}_j)/\tau)}, \tag{2.27}$$

where $q(\mathbf{x})$ is the predicted value of each option $\mathbf{x}$ for a given sampling strategy (e.g., $q(\mathbf{x}) = \text{UCB}(\mathbf{x})$ for UCB sampling), and τ is the temperature parameter (Fig. 2.5c). Lower values of τ produce more concentrated probability distributions centered on the option with the maximum value, where the limit of $\tau \to 0$ is equivalent to an argmax choice rule. Larger values of τ produce noisier choice distributions, where in the limit of $\tau \to \infty$ produces a uniform distribution over choices. We can also understand τ as controlling the level of noisey, *undirected* exploration, which can be contrasted with the UCB β parameter that controls the level of *directed* exploration, targeted towards highly uncertain options.

2.5 Heuristic strategies

We also compare various heuristic strategies that make predictions about search behaviour without learning about the distribution of rewards. These are not combined with either the option learning or function learning models, and make predictions about behavior without maintaining either a tabular or function-like representation of the world (Gigerenzer, Todd, & the ABC Research Group, 1999; Hertwig, Hoffrage, & Martignon, 1999; Todd & Gigerenzer, 2007).

2.5.1 Win-stay lose-sample

We consider a form of a win-stay lose-sample (WSLS) heuristic (Bonawitz et al., 2014), where a *win* is defined as finding a payoff with a higher or equal value than the previously best observed outcome. When the decision-maker "wins", we assume that any option within a Manhattan distance ≤ 1 in the search space is chosen (i.e., a repeat selection or selecting any of the four cardinal neighbours) with equal probability. *Losing* is defined as the failure to improve, and results in sampling any unexplored option with equal probability.

2.5.2 Local search

Local search predicts that search decisions have a tendency to stay local to the previous choice. We use inverse Manhattan distance (IMD) to quantify locality:

$$\mathrm{IMD}(\mathbf{x}, \mathbf{x}') = \frac{1}{\sum_{i=1}^{n} |x_i - x_i'|} \tag{2.28}$$

where $\mathbf{x}$ and $\mathbf{x}'$ are vectors corresponding to options embedded in $\mathbb{R}^n$. For the special case where $\mathbf{x} = \mathbf{x}'$, we set $\mathrm{IMD}(\mathbf{x}, \mathbf{x}') = 1$ to avoid a divide by zero. This local search strategy, when coupled with a softmax choice rule behaves similarly to a Lévy flight (Mandelbrot, 1982), where small steps are more likely relative to longer jumps. Lévy flights have been used to model the search patterns of human hunter–gatherers (Raichlen et al., 2014), predatory birds (Humphries, Weimerskirch, Queiroz, Southall, & Sims, 2012), and honey bees (Reynolds et al., 2007). While sometimes hailed as an optimal foraging strategy under random reward distributions (Viswanathan et al., 1999), the lack of responsiveness to encounters of reward in the environment make Lévy flights poorly adapted to patchy or correlated resource distributions (Hills, Kalff, & Wiener, 2013).

2.6 Conclusion

In this chapter, we have provided a summary on the RL framework and highlighted the computational challenges of learning in large problem spaces. Optimal solutions are generally unavailable, while approximate solutions from Dynamic Programming still require exhaustive exploration of all possibilities. We contrast the tabular approach of these approximate solutions with value function approximation, which learns a global value function over the space of possible options. This provides a useful distinction that maps onto two classes of human learning models. The option learning model is based on classic theories of associative learning, while the function learning model is inspired by theories of explicit function learning in humans. Both

models produce predictions about expected rewards and the underlying uncertainty, which we combine with a variety of heuristic sampling strategies to navigate the exploration-exploitation dilemma.

This set of candidate models for human learning is explored in Chapter 3 where we present a large-scale model comparison across three different MAB problems with spatially correlated rewards. Chapter 4 uses a more restricted set of models to test clear hypotheses about developmental changes in generalization and exploration, where we also include an ε-greedy model as an alternative form of random exploration. Chapter 5 compares option learning and function learning models across spatial and conceptual domains, while Chapter 6 extends the function learning framework to describe generalization over graph structures where relationships between options are discrete and asymmetric.

Chapter 3

Generalization and Exploration in Vast Spaces

The sage steers by the torch of uncertainty and doubt
—Zhuangzi

How do people explore the vast (and sometimes infinite) decision spaces that define our lives, where the number of possibilities can never be exhaustively explored? Here the traditional explore-exploit dilemma breaks down, because aside from knowing when to explore, you also need to know *where* to explore.

This chapter presents three experiments studying how people explore large decision spaces under limited search horizons, using both generated and real-world search environments. Using a model of generalization based on Gaussian process (GP) regression combined with an optimistic sampling strategy, we are able to describe participant behavior and simulate human-like learning curves. Our results also suggest that a tendency towards undergeneralization is a beneficial bias in a wide variety of settings.

This chapter is based on the following publications:

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2017). Mapping the unknown: The spatially correlated multi-armed bandit. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. (pp. 1357–1362). Austin, TX: Cognitive Science Society.

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, 915–924.

3.1 Introduction

Many aspects of human behavior can be understood as a type of search problem (Todd et al., 2012), from foraging for food or resources (Kolling et al., 2012), to searching through a hypothesis space to learn causal relationships (Bramley et al., 2017), or more generally, learning which actions lead to rewarding outcomes (Sutton & Barto, 2018). In a natural setting, these tasks come with a vast space of possible actions, each corresponding to some reward that can only be observed through experience. In such problems, one must learn to balance the dual goals of exploring unknown options, while also exploiting familiar options for immediate returns. This frames the *exploration-exploitation dilemma*, typically studied using the multi-armed bandit framework (Palminteri, Lefebvre, Kilford, & Blakemore, 2017; Reverdy et al., 2014; Speekenbrink & Konstantinidis, 2015; Steyvers et al., 2009), 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 (Gershman & Daw, 2017; Lee, Shimojo, & O'Doherty, 2014). This highlights an intriguing gap between human and machine learning, where traditional approaches to reinforcement learning typically learn about the distribution of rewards for each state independently (Sutton & Barto, 2018). Such an approach falls short in more realistic scenarios where the size of the problem space is far larger than the search horizon, and it becomes infeasible to observe all possible options (Lake et al., 2017; Wilson, Geana, White, Ludvig, & Cohen, 2014). 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 prior experience to unobserved states (Tesauro, 1992). The function learning approach approximates a global value function over all options, including ones not experienced yet (Gershman & Daw, 2017). This allows for generalization to vast and potentially infinite state spaces, based on a small number of observations. Additionally, function learning scales to problems with complex sequential dynamics and has been used in tandem with restricted search methods, such as Monte Carlo sampling, for navigating intractably large search trees (Mnih et al., 2015; Silver et al., 2016). While restricted search methods have been proposed as models of human reinforcement learning in planning tasks (Huys et al., 2015; Solway & Botvinick, 2015), here we focus on situations in which a rich model of environmental structure supports learning and generalization (Guez, Silver, & Dayan, 2013).

Function learning has been successfully utilized for adaptive generalization in various machine learning applications (Rasmussen & Kuss, 2004; Sutton, 1996), although relatively little is known about how humans generalize *in vivo* (e.g., in a search task; but see Reverdy et al., 2014). Building on previous work exploring inductive biases in pure function learning contexts (Lucas et al., 2015; E. Schulz, Tenenbaum, et al., 2017) and human behavior in univariate function optimization (Borji & Itti, 2013), 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 uni- and bivariate multi-armed bandits with up to 121 arms, we compare a diverse set of computational models in their ability to predict individual human behavior. In all experiments, the majority of subjects are best captured by a model combining function learning using Gaussian Process (GP) regression, with an optimistic Upper Confidence Bound (UCB) sampling strategy that directly balances expectations of reward with the reduction of uncertainty. Importantly, we recover meaningful and robust estimates about the nature of human generalization, showing the limits of traditional models of associative learning (Dayan & Niv, 2008) in tasks where the environmental structure supports learning and inference.

The main contributions of this paper are threefold:

1. We introduce the *spatially correlated multi-armed bandit* as a paradigm for studying how people use generalization to guide search in larger problems space than traditionally used for studying human behavior.

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.

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.

3.2 The spatially correlated multi-armed bandit

A useful inductive bias in many real world search tasks is to assume a spatial correlation between rewards (Srivastava et al., 2015) or clumpiness of resource distributions (Wilke et al., 2015). 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. 3.1)

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. Additionally, since recent work has connected both spatial and conceptual representations to a common neural substrate (Constantinescu et al., 2016), our results in a spatial domain provide potential pathways to other search domains, such as contextual (E. Schulz, Konstantinidis, & Speekenbrink, 2017; Stojic, Analytis, & Speekenbrink, 2015; Wu, Schulz, Garvert, Meder, & Schuck, 2018) or semantic search (Abbott, Austerweil, & Griffiths, 2015; Hills et al., 2012).

3.2.1 Methods

Participants and design

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

Experiments 1 and 2 used a 2×2 between-subjects design, where participants were randomly assigned to one of two different payoff structures (*Accumulation condition* vs. *Maximization condition*) and one of two different classes of environments (*Smooth* vs. *Rough*), whereas Experiment 3 used environments from real-world agricultural datasets (Table B.1), and manipulated only the payoff structure (random assignment between subjects). Each grid world represented a (either uni- or bivariate) function, with each observation including normally distributed noise, $\varepsilon \sim \mathcal{N}(0,1)$. The task was presented over either 16 rounds (Exp. 1) or 8 rounds (Exp. 2 and 3) on different grid worlds, which were randomly drawn (without replacement) from the same class of environments. Participants had either a short or long search horizon (Exp. 1: $[5,10]$; Exp. 2 and 3: $[20,40]$) to sample tiles on the grid, including repeat clicks. The search

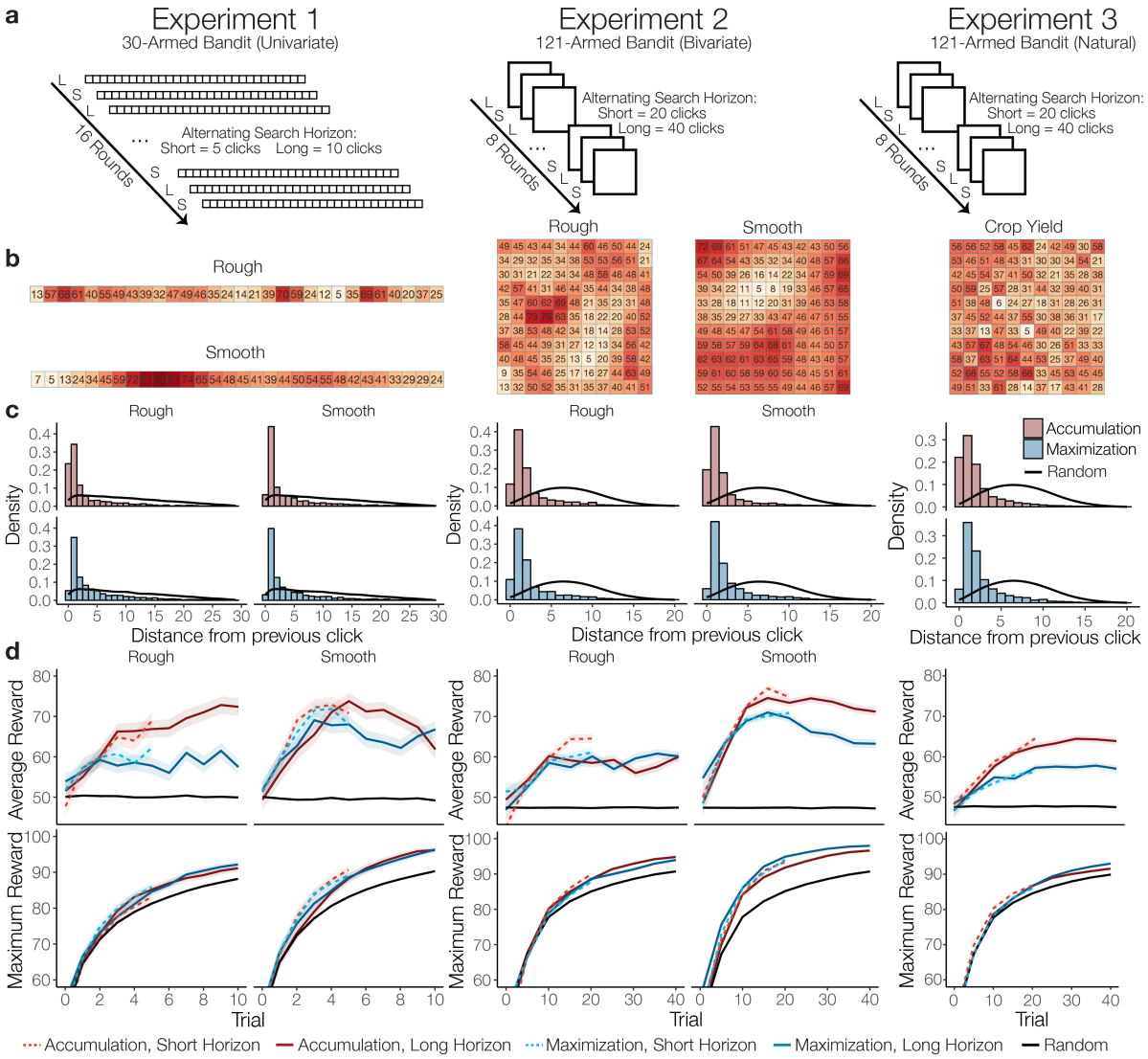

Fig. 3.1 Procedure and behavioral results. Experiments 1 and 2 used a 2×2 between-subject design, manipulating the type of environment (Rough or Smooth) and the payoff condition (Accumulation or Maximization), while Experiment 3 manipulated only payoff conditions (between subjects) and used a set of natural environments where rewards reflect normalized crop yields from various agricultural datasets. **a)** Experiment 1 used a 1D array of 30 possible options, while Experiments 2 and 3 used a 2D array (11×11) with 121 options. Experiments took place over 16 (Exp. 1) or 8 rounds (Exp. 2 and 3), with a new environment sampled without replacement for each round. Search horizons alternated between rounds (within subject), with the horizon order counter-balanced between subjects. **b)** Examples of fully revealed search environments, where tiles were initially blank at the beginning of each round, except for a single randomly revealed tile. Rough and Smooth environments differed in the extent of spatial correlations, while Crop Yield environments have no fixed level of correlation (see Table B.1). **c)** Locality of sampling behavior compared to a random baseline simulated over 10,000 rounds (black line), where distance is measured using Manhattan distance and the y-axis indicates the probability density of different distances (with a different maximum range for Exp. 1 compared to Exp. 2 and 3). **d)** Average reward earned (Accumulation goal) and maximum reward revealed (Maximization goal), where coloured lines indicate the assigned payoff condition and shaded regions show the standard error of the mean. Short horizon trials are indicated by lighter colours and dashed lines, while black lines are a random baseline simulated over 10,000 rounds.

horizon alternated between rounds (within subject), with initial horizon length counterbalanced between subjects by random assignment.

Materials and procedure

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

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

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

where the location on the grid corresponds to the rows and columns of a field and the payoffs reflect the normalized yield of various crops (see Table B.1 for full details).

Search horizons. We chose two horizon lengths (Short=5 or 20 and Long=10 or 40, for Experiment 1 and Experiments 2 and 3, respectively) that were fewer than the total number of tiles on the grid (30 or 121), and varied them within subject (alternating between rounds and counterbalanced). Horizon length was approximately equivalent between Experiments 1 and Experiments 2 and 3, as a fraction of the total number of options (short $\approx \frac{1}{6}$; long $\approx \frac{1}{3}$).

3.3 Results

3.3.1 Experiment 1

Participants ($n = 81$) searched for rewards on a 1×30 grid world, where each tile represented a reward-generating arm of the bandit (Fig. 3.1a). The mean rewards of each tile were spatially correlated, with stronger correlations in *Smooth* than in *Rough* environments (between subjects; Fig. 3.1b). Participants were either assigned the goal of accumulating the largest average reward (*Accumulation condition*), thereby balancing exploration-exploitation, or of finding the best overall tile (*Maximization condition*), an exploration goal directed towards finding the global maximum. Additionally, the search horizons alternated between rounds (within subject; *Short* $= 5$ vs. *Long* $= 10$), with the order counter-balanced between subjects. We hypothesized that if function learning guides search behavior, participants would perform better and learn faster in smooth environments (E. Schulz et al., 2015), in which stronger spatial correlations reveal more information about nearby tiles .

Looking first at sampling behavior, the distance between sequential choices was more localized than chance ($t(80) = 39.8$, $p < .001$, $d = 4.4$, $BF > 100$; Fig. 3.1c), as has also been observed in semantic search Hills et al. (2012) and causal learning (Bramley et al., 2017) domains. Participants in the Accumulation condition sampled more locally than those in the Maximization condition ($t(79) = 3.33$, $p = .001$, $d = 0.75$, $BF = 24$), corresponding to the increased demand to exploit known or near-known rewards. Comparing performance in different environments, the learning curves in Fig. 3.1d show that participants in Smooth environments obtained higher average rewards than participants in Rough environments ($t(79) = 3.58$, $p < .001$, $d = 0.8$, $BF = 47$), consistent with the hypothesis that spatial patterns in the environment can be learned and used to guide search. Surprisingly, longer search horizons (solid vs. dashed lines) did not lead to higher average reward ($t(80) = 0.60$, $p = .549$, $d = 0.07$, $BF = 0.2$). We analyzed both average reward and the maximum reward obtained for each

subject, irrespective of their payoff condition (Maximization or Accumulation). Remarkably, participants in the Accumulation condition performed best according to both performance measures, achieving higher average rewards than those in the Maximization condition ($t(79) = 2.89$, $p = .005$, $d = 0.7$, $BF = 8$), and performing equally well in terms of finding the largest overall reward ($t(79) = -0.73$, $p = .467$, $d = -0.2$, $BF = 0.3$). Thus, a strategy balancing exploration and exploitation—at least for human learners—may achieve the global optimization goal *en passant*.

3.3.2 Experiment 2

Experiment 2 had the same design as Experiment 1, but used a 11×11 grid representing an underlying bivariate reward function (Fig. 3.1 center) and longer search horizons to match the larger search space (*Short* $= 20$ vs. *Long* $= 40$). We replicated the main results of Experiment 1, showing participants ($n = 80$) sampled more locally than a random baseline ($t(79) = 50.1$, $p < .001$, $d = 5.6$, $BF > 100$; Fig. 3.1c), Accumulation participants sampled more locally than Maximization participants ($t(78) = 2.75$, $p = .007$, $d = 0.6$, $BF = 5.7$), and participants obtained higher rewards in Smooth than in Rough environments ($t(78) = 6.55$, $p < .001$, $d = 1.5$, $BF > 100$; Fig. 3.1d). 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 ($t(78) = -0.62$, $p = .534$, $d = -0.1$, $BF = 0.3$), yet in Experiment 2 the Accumulation condition did not lead to substantially better performance than the Maximization condition in terms of average reward ($t(78) = -1.31$, $p = .192$, $d = -0.3$, $BF = 0.5$). Again, short search horizons led to the same level of performance as longer horizons, ($t(79) = -0.96$, $p = .341$, $d = -0.1$, $BF = 0.2$), suggesting that learning occurs rapidly and peaks rather early.

3.3.3 Experiment 3

Experiment 3 used the same 121-armed bivariate bandit as Experiment 2, but rather than generating environments with fixed levels of spatial correlations, we sampled environments from 20 different agricultural datasets (Wright, 2017), where payoffs correspond to the normalized yield of various crops (e.g., wheat, corn, and barley). These datasets 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 payoffs (see Appendix B.4 for selection criteria). The crucial difference compared to

Experiment 2 is that these natural datasets 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 ($t(79) = 50.1$, $p < .001$, $d = 5.6$, $BF > 100$), with participants in the Accumulation condition sampling more locally than those in the Maximization condition ($t(78) = 3.1$, $p = .003$, $d = 0.7$, $BF = 12$). In the natural environments, we found that Accumulation participants achieved a higher average reward than Maximization participants ($t(78) = 2.7$, $p = .008$, $d = 0.6$, $BF = 6$), with an effect size similar to Experiment 1. There was no difference in maximum reward across payoff conditions ($t(78) = 0.3$, $p = .8$, $d = 0.06$, $BF = 0.2$), 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 ($t(78) = 2.1$, $p = .04$, $d = 0.2$, $BF = 0.4$). The results of Experiment 3 corroborate the results of Experiments 1 and 2, showing that our findings in simulated environments are very similar to human behavior in natural environments.

3.4 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 Fig. B.6 and Table B.3 for extended results). These 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 (Lindley, 1956), psychology (Nelson, 2005), and philosophy of science (Crupi & Tentori, 2014). Models of learning form inductive beliefs about the value of possible options (including unobserved options) conditioned on previous observations, while 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 behavior in bandit tasks (Steyvers et al., 2009), yet do not maintain a model of the world (see section 2.5). By far the best predictive models used *Gaussian Process* (GP) regression (Rasmussen & Williams, 2006; E. Schulz, Speekenbrink, & Krause, 2018) as a mechanism for generalization, and *Upper Confidence Bound* (UCB) sampling (Auer, 2002) as an optimistic solution to the exploration-exploitation dilemma.

Function learning provides a possible explanation of how individuals generalize from previous experience to unobserved options, by adaptively learning an underlying function mapping options onto rewards. We use GP regression as an expressive model of human function learning, which has known equivalencies to neural network function approximators (Neal,

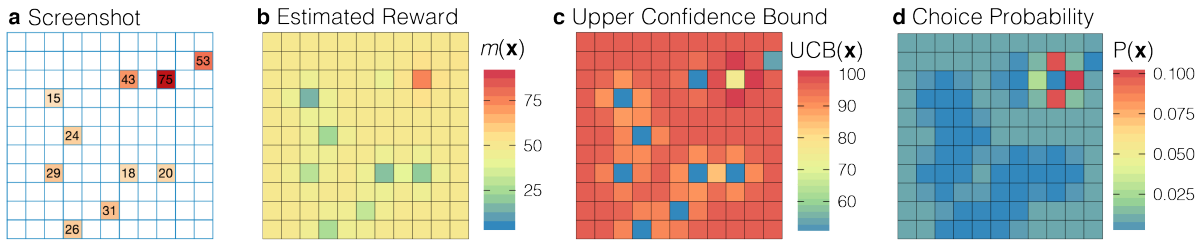

Fig. 3.2 Overview of the GP-UCB Model specified using median participant parameter estimates from Experiment 2 (see Table B.3). **a**) Screenshot of Experiment 2. Participants were allowed to select any tile until the search horizon was exhausted. **b**) Estimated reward (not shown, the estimated uncertainty) as predicted by the GP Function Learning model, based on the points sampled in Panel a. **c**) Upper confidence bound of predicted rewards. **d**) Choice probabilities after a softmax choice rule. $P(\mathbf{x}) = \exp(\text{UCB}(\mathbf{x})/\tau)/\sum_{j=1}^{N}\exp(\text{UCB}(\mathbf{x}_j)/\tau)$, where τ is the temperature parameter (i.e., higher temperature values lead to more random sampling).

2012), yet provides psychologically interpretable parameter estimates about the extent to which generalization occurs. GP function learning can guide search by making predictions about the expected mean $m(\mathbf{x})$ and variance $v(\mathbf{x})$ for each option $\mathbf{x}$ in the global state space (see Fig. 3.2a-b), conditioned on a finite number of previous observations of rewards $\mathbf{y}_t = [y_1, y_2, \ldots, y_t]^\top$ at inputs $\mathbf{X}_t = [\mathbf{x}_1, \ldots, \mathbf{x}_t]$. Similarities between options are modelled by a *Radial Basis Function* (RBF) kernel:

$$k_{\text{RBF}}(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\lambda^2}\right), \tag{3.1}$$

where λ governs how quickly correlations between points $\mathbf{x}$ and $\mathbf{x}'$ (e.g., two tiles on the grid) decay towards zero as their distance increases. We use λ as a free parameter, which can be interpreted psychologically as the extent to which people generalize spatially. Since the GP prior is completely defined by the RBF kernel, the underlying mechanisms are similar to Shepard's universal gradient of generalization (Shepard, 1987), 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 $v(\mathbf{x})$ from the function learning model, UCB sampling produces valuations of each option $\mathbf{x}$ using a simple weighted sum:

$$\text{UCB}(\mathbf{x}) = m(\mathbf{x}) + \beta\sqrt{v(\mathbf{x})}, \tag{3.2}$$

where β is a free parameter governing how much the reduction of uncertainty is valued relative to expectations of reward (Fig. 3.2c). To illustrate, an exploration bonus of $\beta = 0.5$ suggests participants would prefer a hypothetical option $\mathbf{x}_1$ predicted to have mean reward $m(\mathbf{x}_1) = 60$

and uncertainty $\sqrt{v(\mathbf{x}_1)} = 10$, over an option $\mathbf{x}_2$ predicted to have mean reward $m(\mathbf{x}_2) = 64$ and standard deviation $\sqrt{v(\mathbf{x}_2)} = 1$. This is because sampling $\mathbf{x}_1$ is expected to reduce a large amount of uncertainty, even though 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 rewards and exploring to reduce uncertainty (Kaufmann, Cappé, & Garivier, 2012) 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):

$$\mathrm{PureExploit}(\mathbf{x}) = m(\mathbf{x}) \tag{3.3}$$

$$\mathrm{PureExplore}(\mathbf{x}) = \sqrt{v(\mathbf{x})} \tag{3.4}$$

Figure 3.2 shows how the GP-UCB model makes inferences about the search space and uses UCB sampling (combined with a softmax choice rule) to make probabilistic predictions about where the participant will sample next. We refer to this model as the *Function Learning Model* and contrast it with an *Option Learning Model*. The Option Learning Model uses a Bayesian mean tracker (BMT) to learn about the distribution of rewards for each option independently (see subsection 2.3.1). The Option learning Model is a traditional associative learning model, and can be understood as a variant of a Kalman filter where rewards are assumed to be time-invariant (Speekenbrink & Konstantinidis, 2015). Like the Function Learning Model, the Option Learning Model also generates normally distributed predictions with mean $m(\mathbf{x})$ and variance $v(\mathbf{x})$, which we combine with the same set of sampling strategies and the same softmax choice rule to make probabilistic predictions about search. For both models, we use the softmax temperature parameter (τ) to estimate the amount of undirected exploration (i.e., higher temperatures correspond to more noisy sampling; Fig. 3.2d), in contrast to the β parameter of UCB, which estimates the level of exploration directed towards reducing uncertainty.

3.5 Modelling results

Cross-validation. We performed model comparison using cross-validated maximum likelihood estimation (MLE), where each participant's data was separated by horizon length (short or long) and we iteratively form a training set by leaving out a single round, compute a MLE on the training set, and then generate out-of-sample predictions on the remaining round. 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

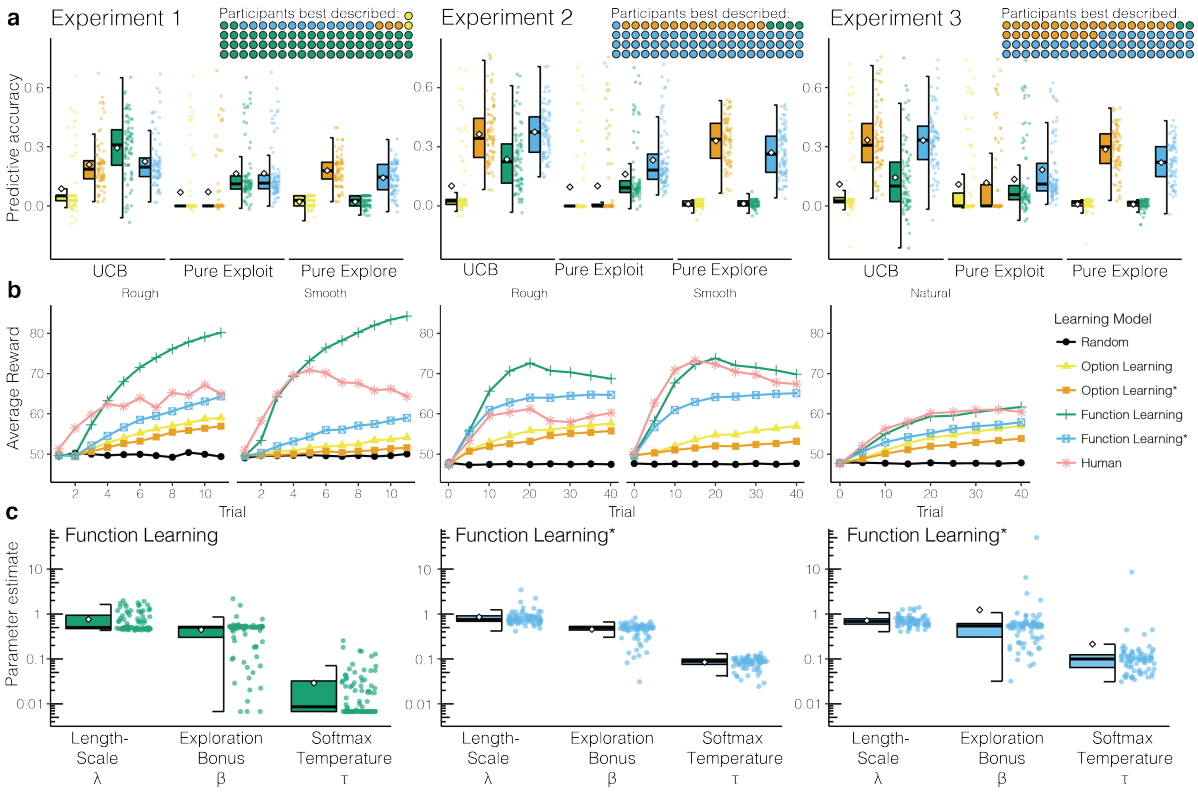

Fig. 3.3 Modelling results. **a**) Cross-validated predictive accuracy of each model (higher is better), with box plots indicating the IQR, the median (horizontal line), mean (diamond), and 1.5x IQR (whiskers). Each individual participant is shown as a single dot, with the number of participants best described shown as an icon array (inset; aggregated by sampling strategies). Asterisks (*) indicate a localized variant of the Option Learning or Function Learning models, where predictions are weighted by the inverse distance from the previous choice (see Methods). **b**) Learning curves of participants and model simulations. Each simulated learning model uses UCB sampling and is specified using participants 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 1.5x IQR (whiskers), median (horizontal line), and mean (diamond).

participant, and out-of-sample predictions for 120 choices in Experiment 1 and 240 choices in Experiments 2 and 3 (per participant). Prediction error (computed as log loss) was summed up over all rounds, and is reported as *predictive accuracy*, using a pseudo-R^2 measure that compares the total log loss prediction error for each model to that of a random model:

$$R^2 = 1 - \frac{\log \mathcal{L}(\mathcal{M}_k)}{\log \mathcal{L}(\mathcal{M}_{\text{rand}})}, \tag{3.5}$$

where $\log \mathcal{L}(\mathcal{M}_{\text{rand}})$ is the log loss of a random model and $\log \mathcal{L}(\mathcal{M}_k)$ is model k's out-of-sample prediction error. Moreover, we calculated each model's protected probability of exceedance using its predictive log-evidence (Rigoux, Stephan, Friston, & Daunizeau, 2014; Stephan, Penny, Daunizeau, Moran, & Friston, 2009). This probability is defined as the

probability that a particular model is more frequent in the population than all the other models, averaged over the probability of the null hypothesis that all models are equally frequent (thereby correcting for chance performance).

Localized models. All models (except for local search) include a localized variant, which introduced a locality bias by weighting the predicted value of each option $q(\mathbf{x})$ by the inverse Manhattan distance (IMD) to the previously revealed tile. This is equivalent to a multiplicative combination with the Local Search model (subsection 2.5.2), and is similar to a "stickiness parameter" (Christakou, Murphy, et al., 2013; Gershman, Pesaran, & Daw, 2009), although we implement it here without the introduction of any additional free parameters. Localized models are indicated with an asterisk (e.g., Function Learning*).

Simulated learning curves. We use participants' cross-validated parameter estimates to specify a given model and then simulate performance. At each trial, model predictions correspond to a probabilistic distribution over options, which was then sampled and used to generate the observation for the next trial. In order to correspond with the manipulations of horizon length, payoff condition, and environment type, each simulation was performed at the participant level, producing data resembling a virtual participant for each replication. Iterating over each round, we selected the same environment as seen by the participant and then simulated data using the cross-validated parameters that were estimated using that round as the left-out round. Thus, just as model comparison was performed out-of-sample, the generated data was also out-of-sample, based on parameters that were estimated on a different set of rounds than the one being simulated. We performed 100 replications for each participant in each experiment, which were then aggregated to produce the learning curves in Figure 3.3b.

3.5.1 Experiment 1

Participants were better described by the Function Learning Model than the Option Learning Model ($t(80) = 14.10$, $p < .001$ $d = 1.6$, $BF > 100$, 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 behavior, with the Pure Exploitation ($t(80) = -8.85$, $p < .001$, $d = -1.0$, $BF > 100$) and Pure Exploration ($t(80) = -16.63$, $p < .001$, $d = -1.8$, $BF > 100$) variants each making 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. 3.3a), penalizing options farther away from the previous selected option (without introducing additional free parameters). While the Option Learning* Model was better than the standard Option Learning Model ($t(80) = 16.13$, $p < .001$, $d = 1.8$, $BF > 100$), the standard Function Learning Model still outperformed its localized variant ($t(80) = 5.05$, $p < .001$, $d = 0.6$, $BF > 100$). 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. Lastly, we also calculated each model's protected exceedance probability [†] (Rigoux et al., 2014; Stephan et al., 2009) using its out-of-sample log loss. 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 Table B.3 for comparison with additional models) while also correcting for chance. Doing so, we found that the Function Learning-UCB Model reached a protected probability of $\text{pxp} = 1$, indicating that it vastly outperformed all of the other models.

Figure 3.3b shows simulated learning curves of each model in comparison to human performance, where models were specified using parameters from participants' estimates (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 while others produce continuously increasing learning curves (see Figure B.5 for individual learning curves). While aggregated learning curves should be analyzed with caution (Myung, Kim, & Pitt, 2000), we find an overlap between elements of human intelligence responsible for successful performance in our task, and elements of participant behavior 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. 3.3c) indicated that people tend to underestimate the extent of spatial correlations, with median per-participant λ estimates significantly lower than the ground truth ($\lambda_{Smooth} = 2$ and $\lambda_{Rough} = 1$) for both Smooth (Wilcoxon signed rank test; $\hat{\lambda}_{Smooth} = 0.5$, $Z = -7.1$, $p < .001$, $r = 1.1$, $BF > 100$) and Rough environments ($\hat{\lambda}_{Rough} = 0.5$, $Z = -3.4$, $p < .001$, $r = 0.55$, $BF > 100$). This can be interpreted as a tendency towards undergeneralization. Additionally, we found that the estimated exploration bonus of UCB sampling (β)

[†] See section A.5 for a description of the Bayesian model selection framework for group studies.

was reliably greater than zero ($\hat{\beta} = 0.51$, $Z = -7.7$, $p < .001$, $r = 0.86$, $BF > 100$, compared to lower estimation bound), reflecting the valuation of sampling uncertain options, together with exploiting high expectations of reward. Lastly, we found relatively low estimates of the softmax temperature parameter ($\hat{\tau} = 0.01$), suggesting that the search behavior of participants corresponded closely to selecting the very best option, once they had taken into account both the exploitation and exploration components of the available actions.

3.5.2 Experiment 2

In a more complex bivariate environment (Fig. 3.3a), the Function Learning Model again made better predictions than the Option Learning Model ($t(79) = 9.99$, $p < .001$, $d = 1.1$, $BF > 100$), although this was only marginally the case when comparing localized Function Learning* to localized Option Learning* ($t(79) = 2.05$, $p = .044$, $d = 0.2$, $BF = 0.9$). In the two-dimensional search environment of Experiment 2, adding localization improved predictions for both Option Learning ($t(79) = 19.92$, $p < .001$, $d = 2.2$, $BF > 100$) and Function Learning ($t(79) = 10.47$, $p < .001$, $d = 1.2$, $BF > 100$), in line with the stronger tendency towards localized sampling compared to Experiment 1 (see Fig. 3.1c). 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 ($t(79) = -6.44$, $p < .001$, $d = -0.7$, $BF > 100$) and Pure Explore ($t(79) = -12.8$, $p < .001$, $d = -1.4$, $BF > 100$) 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 learned slowly and only marginally outperformed a random sampling strategy (Fig. 3.3b), whereas both variants of the Function Learning Model achieved performance comparable to that of human participants[†]. Median per-participant parameter estimates (Fig. 3.3c) from the Function Learning*-UCB Model showed that while participants generalized somewhat more than in Experiment 1 ($\hat{\lambda} = 0.75$, $Z = -3.7$, $p < .001$, $r = 0.29$, $BF > 100$), they again underestimated the strength of the underlying spatial correlation in both Smooth ($\hat{\lambda}_{Smooth} = 0.78$, $Z = -5.8$, $p < .001$, $r = 0.88$, $BF > 100$; comparison to $\lambda_{Smooth} = 2$) and Rough environments ($\hat{\lambda}_{Rough} = 0.75$, $Z = -4.7$, $p < .001$, $r = 0.78$, $BF > 100$; comparison to $\lambda_{Rough} = 1$). This suggests a robust tendency to undergeneralize. There were no differences in the estimated exploration bonus β between Experiment 1 and 2 ($\hat{\beta} = 0.5$, $Z = 0.86$, $p = .80$, $r = 0.07$, $BF = 0.2$),

[†]Note that the learning curves for Experiment 2 differ from the published version. A correction has been submitted to the editor.

although the estimated softmax temperature parameter τ was larger than in Experiment 1 ($\hat{\tau} = 0.09$; $Z = -8.89$, $p < .001$, $r = 0.70$, $BF = 34$). Experiment 2 therefore replicated the main findings of Experiment 1. When taken together, results from the two experiments provide strong evidence that human search behavior is best explained by function learning paired with an optimistic trade-off between exploration and exploitation.

3.5.3 Experiment 3

Using natural environments without a fixed level of spatial correlations, we replicated key results from the prior experiments: Function Learning made better predictions than Option Learning ($t(79) = 3.03$, $p = .003$, $d = 0.3$, $BF = 8$); adding localization improved predictions for both Option Learning ($t(79) = 18.83$, $p < .001$, $d = 2.1$, $BF > 100$) and Function Learning ($t(79) = 14.61$, $p < .001$, $d = 1.6$, $BF > 100$); and the combined UCB algorithm performed better than using only a Pure Exploit ($t(79) = 12.97$, $p < .001$, $d = 1.4$, $BF > 100$) or a Pure Explore strategy ($t(79) = 5.87$, $p < .001$, $d = 0.7$, $BF > 100$). However, the difference between the localized Function Learning* and the localized Option Learning* was negligible ($t(79) = 0.32$, $p = .75$, $d = 0.04$, $BF = 0.1$). This is perhaps due to the high variability across environments, which makes it harder to predict out-of-sample choices using generalization behavior (i.e., 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 $\text{pxp} = 0.98$ that the Function Learning* model was more frequent in the population than all the other models, followed by $\text{pxp} = 0.01$ 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 Figure 3.3b 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 behavior (Palminteri, Wyart, & Koechlin, 2017). The parameter estimates from the Function Learning* Model are largely consistent with the results from Experiment 2 (Fig. 3.3c), but with participants generalizing slightly less ($\hat{\lambda}_{natural} = 0.68$, $Z = -3.4$, $p < .001$, $r = 0.27$, $BF = 10$), and exploring slightly more, with a small increase in both directed exploration ($\hat{\beta}_{natural} = 0.54$, $Z = -2.3$, $p = .01$, $r = 0.18$, $BF = 5$) and undirected exploration ($\hat{\tau}_{natural} = 0.1$, $Z = -2.2$, $p = .02$, $r = 0.17$, $BF = 4$) parameters. Altogether, the parameter estimates are highly similar to the previous experiments.

3.5.4 Robustness and recovery

We conducted both model and parameter recovery simulations to assess the validity of our modelling results (see Fig. B.1 and Fig. B.2). 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 (Fig. B.1), suggesting robustness to Type I errors and ruling out model overfitting (i.e., the Function Learning Model did not best predict data generated by the Option Learning Model). Parameter recovery was performed to ensure that each parameter in the Function Learning-UCB Model robustly captured separate and distinct phenomena. In all cases, the generating and recovered parameter estimates were highly correlated (Fig. B.2). 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 (Wilson et al., 2014) as a separate phenomena from noisy, undirected exploration (Daw, O'doherty, Dayan, Seymour, & Dolan, 2006).

3.5.5 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. We therefore 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 structure of the environment (Fig. 3.4). These simulations were performed by first generating search environments by sampling from a GP prior specified using a *teacher* length-scale (λ_0), and then simulating search in this environment by specifying the GP-UCB Model with a *student* length-scale (λ_1). Instead of a discrete grid, we chose a set-up common in Bayesian optimization (Metzen, 2016) with continuous bivariate inputs in the range $x, y = [0, 1]$, allowing for a broader set of potential mismatched alignments (see Fig. B.3 for simulations using the exact design of each experiment).

We find that undergeneralization largely leads to better performance than overgeneralization, and that this effect is more pronounced over time t (i.e., longer search horizons). Estimating the best possible alignment between λ_0 and λ_1 revealed that underestimating λ_0 by an average of about 0.21 produces the best scores over all scenarios. These simulation results show that the systematically lower estimates of λ captured by our models are not necessarily a flaw in human cognition, but can sometimes lead to better performance. Indeed, simulations based on

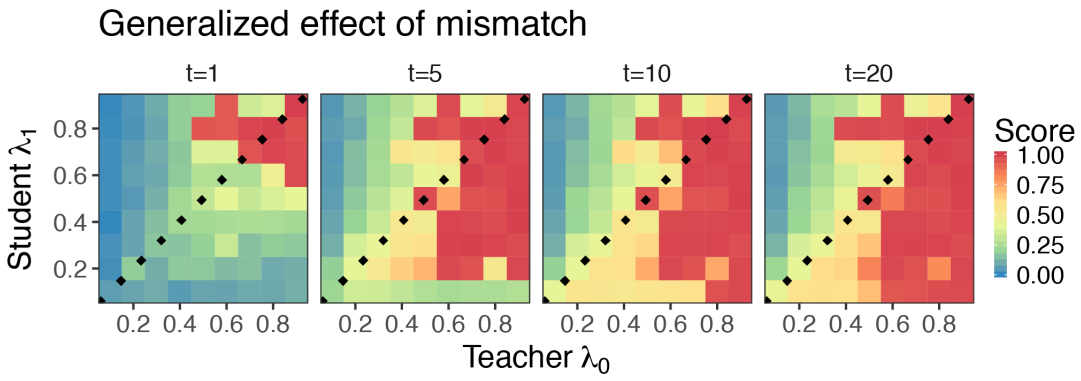

Fig. 3.4 Mismatched length-scale (λ) simulation results. The teacher length-scale λ_0 is on the x-axis, the student length-scale λ_1 is on the y-axis, and each panel is performance at a different trial t. The teacher λ_0 values were used to generate environments, while the student λ_1 values were used to parameterize the Function Learning Model to simulate search performance. The dotted lines show where $\lambda_0 = \lambda_1$ and mark the difference between undergeneralization and overgeneralization, with points below the line indicating undergeneralization. We report the median score (over 100 replications) as a standardized measure of performance, such that 0 shows the lowest possible and 1 the highest possible log unit-performance.

the natural environments used in Experiment 3 (which had no fixed level of spatial correlations) revealed that the range of participant λ estimates were highly adaptive to the environments they encountered (Fig. B.3c). Undergeneralization might not be a bug, but rather an important feature of human behavior.

3.6 Discussion

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

While GP function learning combined with a UCB sampling algorithm has been successfully applied to search problems in ecology (Gotovos et al., 2013), robotics (Cully et al., 2015; Deisenroth, Fox, & Rasmussen, 2015), and biology (Sui, Gotovos, Burdick, & Krause, 2015), 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 GP-UCB and human learning rates (Borji & Itti, 2013; Reverdy et al., 2014), 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 behavior using participants' parameter estimates, we competitively assessed which models best predicted human behavior. The vast majority of participants were best described by the GP-UCB model or its localized variant. Parameter estimates from the best-fitting GP-UCB models suggest there was a systematic tendency to undergeneralize the extent of spatial correlations, which we found can sometimes lead to better search performance than even an exact match with the underlying structure of the environment (Fig. 3.4).

Altogether, our modelling framework yielded highly robust and recoverable results (Fig. B.1) and parameter estimates (Fig. B.2). Whereas previous research on exploration bonuses has had mixed results (Daw et al., 2006; Speekenbrink & Konstantinidis, 2015; Wilson et al., 2014), 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 (Srinivas et al., 2010). This shows a remarkable concurrence between intuitive human strategies and state-of-the-art machine learning research.

3.6.1 Limitations and extensions

One potential limitation is that our payoff manipulation failed to induce superior performance according to the relevant performance metric. While 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. While this may have been insufficient for observing clear performance differences (see Table B.2 for parameter differences), the practical difference between these two goals is murky even in the Bayesian optimization literature, where the strict goal of finding the global optimum is often abandoned based purely on computational concerns (Močkus, 2012). Instead, the global optimization goal is frequently replaced by an approximate measure of performance, such as cumulative regret (Srinivas et al., 2010), which closely aligns to our Accumulation payoff condition. In our experiments, remarkably, participants assigned to the Accumulation goal payoff condition also performed best relative to the maximization criterion.

In addition to providing the best model of human behavior, the Function Learning Model also offers many opportunities for theory integration. The Option Learning Model can itself be reformulated as special case of GP regression (Reece & Roberts, 2010). When the length-scale of the RBF kernel approaches zero ($\lambda \rightarrow 0$), the Function Learning Model assumes state independence, as in the Option Learning Model. Thus, there may be a continuum of reinforcement learning models, ranging from the traditional assumption of state *in*dependence to the opposite extreme, of complete state *inter*-dependence. Moreover, GPs also have equivalencies to Bayesian neural networks (Neal, 2012), suggesting a further link to distributed function learning models (LeCun, Bengio, & Hinton, 2015). Indeed, one explanation for the impressive performance of deep reinforcement learning (Mnih et al., 2015) is that neural networks are specifically a powerful type of function approximator (Schölkopf, 2015).

Lastly, both spatial and conceptual representations have been connected to a common neural substrate in the hippocampus (Constantinescu et al., 2016), suggesting a potential avenue for applying the same GP-UCB model for modelling human learning using contextual (E. Schulz, Konstantinidis, & Speekenbrink, 2017; Stojic et al., 2015; Wu, Schulz, Garvert, et al., 2018), semantic (Abbott et al., 2015; Hills et al., 2012), or potentially even graph-based features. One hypothesis for this common role of the hippocamus is that it performs predictive coding of future state transitions (Stachenfeld et al., 2017), also known as "successor representation" (Dayan & Niv, 2008). In our task, where there are no restrictions on state transitions (i.e., each state is reachable from any prior state), it may be the case that the RBF kernel driving our GP Function Learning model performs the same role as the transition matrix of a successor representation model, where state transitions are learned via a random walk policy. This link is explored in more detail in Chapters 5 and 6.

3.6.2 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. Additionally, 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 modelling framework than traditional multi-armed bandit problems used for studying human behavior. Our empirical and modelling results show how function learning, combined with optimistic search strategies, may provide the foundation of adaptive behavior in complex environments.

Chapter 4

The Developmental Trajectory of Generalization and Search

For the child enamoured with stamps and engravings,
The universe is equal to your vast appetite!
—Baudelaire

How do children and adults differ in their search for rewards? Children are prone to higher variability in their sampling behavior, with an influential hypothesis describing development as a "cooling off" process (Gopnik et al., 2017), where initially random behavior is reduced over the lifespan. We use computation modeling to add clarity to this theory. Our model parameters correspond specifically to three different hypotheses, attributing developmental differences to changes in (i) random sampling, (ii) in exploration directed towards uncertain options, or (iii) the extent of generalization.

Using a field study conducted in museums, we compare children and adults in their ability to generalize about unobserved outcomes and balance the exploration-exploitation dilemma. Our results show that children generalize less than adults, yet are hungrier for information, by using more directed exploration. We do not, however, find differences in terms of random sampling. This account is corroborated by judgments about unobserved options and a forced choice task, providing broad support for a richer picture of developmental differences.

This chapter is based on the following manuscript, where all analyses and figures were either originally produced by me or have been independently reproduced with minor variations:

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (in press). Searching for rewards like a child means less generalization and more directed exploration. *Psychological Science*.

4.1 Introduction

Alan Turing (1950) famously believed that in order to build a General Artificial Intelligence, one must create a machine that can learn like a child. Indeed, recent advances in machine learning often contain references to child-like learning and exploration (Lake et al., 2017; Riedmiller et al., 2018). Yet little is known about how children actually explore and search for rewards in their environments, and in what ways their behavior differs from adults.

In the course of learning through interactions with the environment, all organisms (biological or machine) are confronted with the *exploration-exploitation dilemma* (Mehlhorn et al., 2015). This dilemma highlights two opposing goals. The first is to explore unfamiliar options that provide useful information for future decisions, yet may result in poor immediate rewards. The second is to exploit options known to have high expectations of reward, but potentially forgo learning about unexplored options.

In addition to balancing exploration and exploitation, another crucial ingredient for adaptive search behavior is a mechanism that can *generalize* beyond observed outcomes, thereby guiding search and decision making by forming inductive beliefs about novel options. For example, from a purely combinatorial perspective, it only takes a few features and a small range of values to generate a pool of options vastly exceeding what could ever be explored in a lifetime. Nonetheless, humans of all ages manage to generalize from limited experiences in order to choose from amongst a set of potentially unlimited possibilities. Thus, a model of human search also needs to provide a mechanism for generalization.

Previous research has found extensive variability and developmental differences in children's and adults' search behavior, which not only result from a progressive refinement of basic cognitive functions (e.g., memory or attention), but also derive from systematic changes in the computational principles driving behavior (Palminteri, Kilford, Coricelli, & Blakemore, 2016). In particular, developmental differences in learning and decision making have been explained by appealing to three hypothesized mechanisms: children sample more randomly, explore more eagerly, or generalize more narrowly than adults.

In this chapter, we investigate how these three mechanisms are able to explain developmental differences in exploration-exploitation behavior. We provide a precise characterization of these competing ideas in a formal model, which is used to predict behavior in a search task, where noisy and continuous rewards are spatially correlated. Using behavioral markers, interpreting parameter estimates from computational models, and analyzing judgments about unexplored options, our results converge on the finding that children generalize less, but engage in more directed exploration than adults. We do not, however, find reliable developmental differences in random exploration. These results enrich our understanding of maturation in learning and

decision making, demonstrating that children explore using uncertainty-guided mechanisms rather than simply behaving more randomly.

4.1.1 A tale of three mechanisms

Development as cooling off

Because optimal solutions to the exploration-exploitation dilemma are generally intractable (Bellman, 1952; Gittins & Jones, 1979), heuristic alternatives are frequently employed. In particular, learning under the demands of the exploration-exploitation trade-off has been described using at least two distinct strategies (Wilson et al., 2014). One such strategy is increased *random exploration*, which uses noisy, random sampling to learn about new options.

A key finding in the psychological literature is that children tend to try out more options than adults (Cauffman et al., 2010; Mata, Wilke, & Czienskowski, 2013). This has been interpreted as evidence for higher levels of random exploration in children, and has been loosely compared to algorithms of simulated annealing from computer science (Gopnik et al., 2017), where the amount of random exploration gradually reduces over time. Children can be described as having higher temperature parameters, where the learner initially samples very randomly across a large set of possibilities, before eventually focusing on a smaller subset (Gopnik, Griffiths, & Lucas, 2015). This temperature parameter is expected to "cool off" with age, leading to lower levels of random exploration in late childhood and adulthood.

Development as reduction of directed exploration

A second strategy to tackle the exploration-exploitation dilemma is to use *directed exploration* by preferentially sampling highly uncertain options in order to gain more information and reduce uncertainty about the environment. Directed exploration has been formalized by introducing an "uncertainty bonus" that values the exploration of lesser known options (Auer, 2002), with behavioral markers found in a number of studies (M. J. Frank, Doll, Oas-Terpstra, & Moreno, 2009; Wu, Schulz, Speekenbrink, et al., 2018).

Directed exploration treats information as intrinsically valuable by inflating rewards by their estimated uncertainty (Auer, 2002). This leads to a more sophisticated *uncertainty-guided sampling* strategy that could also explain developmental differences. Indeed, the literature on self-directed learning shows that children are clearly capable of exploring their environment in a systematic, directed fashion. Already infants tend to value the exploration of uncertain options (L. E. Schulz, 2015), and children can balance theory and evidence in simple exploration tasks (Bonawitz, van Schijndel, Friel, & Schulz, 2012) and are able to efficiently adapt their search behavior to different environmental structures (Nelson, Divjak, Gudmundsdottir, Martignon, &

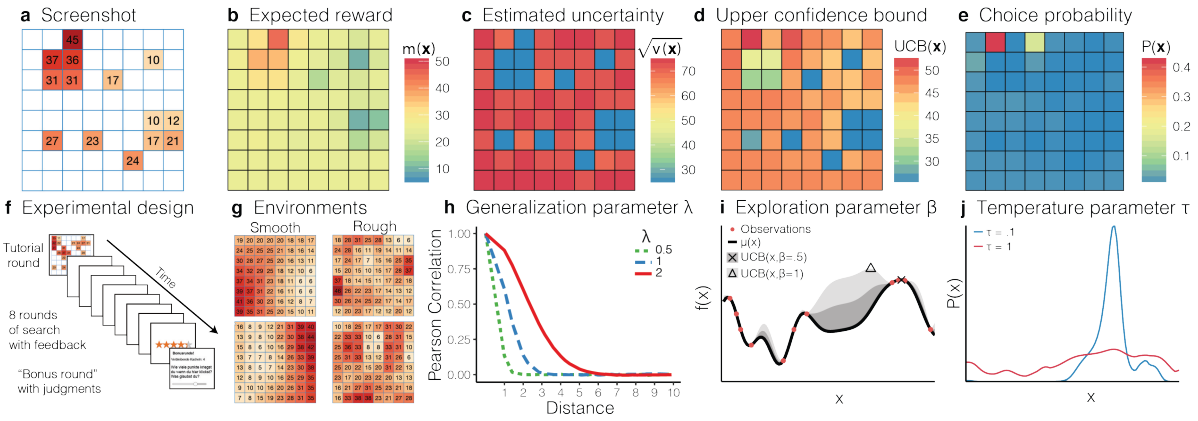

Fig. 4.1 Overview of task and model. **a**) Screenshot of experiment in the middle of a round with partially revealed grid. **b**) Expected reward and **c**) estimated uncertainty, based on observations in **a**) using Gaussian Process regression as a model of generalization. **d**) Upper confidence bounds of each option based on a weighted sum of panels b and c. **e** Choice probabilities of softmax function. Panels (b-e) use median participant parameter estimates. **f**) Overview of the experimental design and **g**) types of environments. **h**) Correlations of rewards between different options decay exponentially as a function of their distance, where higher values of λ lead to slower decays and broader generalizations. **i**) An illustration of UCB sampling using a univariate example, where the expected reward (black line) and estimated uncertainty (gray ribbon; for different values of β) are summed up. Higher values of β value the exploration of uncertain options more strongly (compare the argmax of the two beta values, indicated by the cross and the triangle). **j**) Overview of softmax function, where higher values of the temperature parameter τ lead to increased random exploration.

Meder, 2014; Ruggeri & Lombrozo, 2015). Moreover, children can sometimes even outperform adults in the self-directed learning of unusual relationships (Lucas, Bridgers, Griffiths, & Gopnik, 2014). Both directed and random exploration do not have to be mutually exclusive mechanisms, with recent research finding signatures of both types of exploration in adolescent and adult participants (Gershman, 2018a; Somerville et al., 2017; Wilson et al., 2014).

Development as refined generalization

Rather than explaining development as a change in how we explore given some beliefs about the world, *generalization-based accounts* attribute developmental differences to the way we form our beliefs in the first place. Many studies have shown that human learners use structured knowledge about the environment to guide exploration (Acuna & Schrater, 2009; E. Schulz, Konstantinidis, & Speekenbrink, 2017), where the quality of these representations and the way that people utilize them to generalize across experiences can have a crucial impact on search behavior. Thus, development of more complex cognitive processes (Blanco et al., 2016), leading to broader generalizations, could also account for the observed developmental differences in sampling behavior.

The notion of generalization as a mechanism for explaining developmental differences has a long standing history in psychology. For instance, Piaget (1964) assumed that children

learn and adapt to different situational demands by the processes of assimilation (applying a previous concept to a new task) and accommodation (changing a previous concept in the face of new information). Expanding on Piaget's idea, Klahr (1982) proposed generalization as a crucial developmental process, in particular the mechanism of regularity detection, which supports generalization and improves over the course of development. More generally, the implementation of various forms of decision making (Hartley & Somerville, 2015) could be constrained by the capacity for complex cognitive processes, which become more refined over the life span. For example, although younger children attend more frequently to irrelevant information than older children (Hagen & Hale, 1973), they can be prompted to attend to the relevant information by marking the most relevant cues, whereupon they eventually select the best alternative (Davidson, 1996). Thus, children may indeed be able to apply uncertainty-driven exploratory strategies, but lack the appropriate task representation to successfully implement them.

4.2 A task to study generalization and exploration

In Experiment 4, we study the behavior of both children and adults in a *spatially correlated multi-armed bandit* task (Fig. 4.1a; Wu, Schulz, Speekenbrink, et al., 2018, see also White, 2013, for a similar task), where rewards are distributed on a grid characterized by spatial correlation (i.e., high rewards cluster together; Fig. 4.1g) and the search horizon is vastly smaller than the number of options. Efficient search and accumulation of rewards in such an environment requires two critical components. First, participants need to learn about the underlying spatial correlation in order to generalize from observed rewards to unseen options. This is crucial because there are considerably more options than can be explored within the limited search horizon. Second, participants need a sampling strategy that achieves a balance between exploring new options and exploiting known options with high rewards.

4.3 Methods

Participants

We recruited 55 younger children (range: 7 to 8, 26 female, M_{age}=7.53; SD=0.50), 55 older children (range: 9 to 11, 24 female, M_{age}=9.95; SD=0.80), and 50 adults (range: 18 to 55, 25 female, M_{age}=33.76; SD= 8.53) from museums in Berlin, Germany. We determined the different age groups and the number of participants per group based on prior active learning research (Ruggeri & Lombrozo, 2015) and before data collection commenced. Participants were

paid up to €3.50 for taking part in the experiment, contingent on performance (range: €2.00 to €3.50, M_{reward}=€2.67; SD=0.50). Informed consent was obtained from all participants.

Design

The experiment used a between-subjects design, where participants were randomly assigned to one of two different classes of environments (Fig. 4.1g), with *smooth environments* having stronger spatial correlations than *rough environments*. We generated 40 of each class of environments from a radial basis function kernel (see below), with $\lambda_{smooth} = 4$ and $\lambda_{rough} = 1$. On each round, a new environment was sampled (without replacement) from the set of 40 environments, which was then used to define a bivariate function on the grid, with each observation including additional normally distributed noise $\varepsilon \sim \mathcal{N}(0,1)$. The task was presented over ten rounds on different grid worlds drawn from the same class of environments. The first round was a tutorial round and the last round was a bonus round in which participants sampled for 15 trials and then had to generate predictions for five randomly chosen tiles on the grid. Participants had a search horizon of 25 trials per grid, including repeat clicks.

Materials and procedure

Participants were introduced to the task through a tutorial round and were required to correctly complete three comprehension questions prior to continuing the task. At the beginning of each round, one random tile was revealed and participants could click on any of the tiles (including re-clicks) on the grid until the search horizon was exhausted. Clicking an unrevealed tile displayed the numerical value of the reward along with a corresponding color aid, where darker colors indicated higher rewards. Per round, observations were scaled to a randomly drawn maximum value in the range of 35 to 45, so that the value of the global optima could not be easily guessed. Re-clicked tiles could show some variations in the observed value due to noise. For repeat clicks, the most recent observation was displayed numerically, while the color of the tile corresponded to the mean of all previous observations. In the bonus round, participants sampled for 15 trials and were then asked to generate predictions for five randomly selected and previously unobserved tiles. Additionally, participants had to indicate how certain they were about their prediction on a scale from 0 to 10. Afterwards, they had to select one of the five tiles before continuing with the round.

Participants were awarded up to five stars at the end of each round (e.g., 4.6 out of 5), based on the ratio of their average reward to the global maximum. The performance bonus was calculated based on the average number of stars earned in each round, excluding the tutorial

round. 5 out of 5 stars corresponded to €3.50, while each half star interval reduced the bonus by €0.50 until a minimum bonus of €0.50.

4.3.1 A combined model of generalization and exploration

We use a formal model that combines generalization with a sampling strategy accounting for both directed and random exploration (Wu, Schulz, Speekenbrink, et al., 2018), and use it to predict each participant's out-of-sample search behavior. The generalization component is based on *Gaussian Process* (GP) regression, and to briefly summarize, is a Bayesian function learning approach theoretically capable of learning any stationary function (Rasmussen & Williams, 2006) and has been found to effectively describe human behavior in explicit function learning tasks (Lucas et al., 2015). The GP component is used to adaptively learn a value function, which generalizes the limited set of observed rewards over the entire search space using Bayesian inference.

The GP prior is completely determined by the choice of a kernel function $k(\mathbf{x}, \mathbf{x}')$, which encodes assumptions about how points in the input space are related to each other. A common choice of this function is the *radial basis function* (RBF) kernel:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\lambda^2}\right), \tag{4.1}$$

where the length-scale parameter λ encodes the extent of spatial generalization between options (tiles) in the grid. The assumptions of this kernel function are similar to the gradient of generalization historically described by Shepard (1987), which also models generalization as an exponentially decaying function of the stimulus similarity distance (see Fig. 4.1h), which has been observed across a wide range of stimuli and organisms. As an example, generalization with $\lambda = 1$ corresponds to the assumption that the rewards of two neighboring tiles are correlated by $r = 0.6$, and that this correlation decays to zero for options further than three tiles apart. We treat λ as a free parameter in our model comparison in order to assess age-related differences in the capacity for generalization.

Given different possible options $\mathbf{x}$ to sample from (i.e., tiles on the grid), GP regression generates normally distributed beliefs about rewards with expectation $m(\mathbf{x})$ and estimated uncertainty $s(\mathbf{x})$ (Fig. 4.1b,c). A sampling strategy is then used to map the beliefs of the GP onto a valuation for sampling each option at a given time. Crucially, such a sampling strategy must address the exploration-exploitation dilemma. One frequently applied heuristic for solving this dilemma is *Upper Confidence Bound* (UCB) sampling (Srinivas et al., 2010), which evaluates each option based on a weighted sum of expected reward and estimated

uncertainty:

$$\mathrm{UCB}(\mathbf{x}) = m(\mathbf{x}) + \beta\sqrt{v(\mathbf{x})} \tag{4.2}$$

where β models the extent to which uncertainty is valued positively and therefore directly sought out. This strategy corresponds to directed exploration because it encourages the sampling of options with higher uncertainty according to the underlying generalization model (see Fig. 4.1i). We treat the exploration parameter β as a free parameter to assess how much participants value the reduction of uncertainty (i.e., engage in directed exploration). As described before, an exploration bonus of $\beta = 0.5$ means participants would prefer an option x_1 expected to have reward $m(x_1) = 30$ and uncertainty $\sqrt{v(x_1)} = 10$, over option x_2 expected to have reward $m(x_2) = 34$ and uncertainty $\sqrt{v(x_2)} = 1$. This is because sampling x_1 is expected to reduce a larger amount of uncertainty, even though x_2 has a higher expected reward ($\mathrm{UCB}(x_1|\beta = 0.5) = 35$ vs. $\mathrm{UCB}(x_2|\beta = 0.5) = 34.5$).

Finally, we use a softmax function to map the upper confidence bound values, $\mathrm{UCB}(\mathbf{x})$, of our proposed Gaussian Process-Upper Confidence Bound sampling model onto choice probabilities:

$$p(\mathbf{x}) = \frac{\exp(\mathrm{UCB}(\mathbf{x})/\tau)}{\sum_{j=1}^{N}\exp(\mathrm{UCB}(\mathbf{x}_j)/\tau)}, \tag{4.3}$$

where τ is the temperature parameter governing the amount of randomness in sampling behavior. If τ is high (higher temperatures), then participants are assumed to sample more randomly, whereas if τ is low (cooler temperatures), the choice probabilities are concentrated on the highest valued options (Fig. 4.1j). Thus, τ encodes the tendency towards random exploration. We treat τ as a free parameter to assess the extent of random exploration in children and adults (see Appendix section C.1 for alternative implementations such as ε-greedy sampling and estimation of optimal parameters).

In summary, GP-UCB contains three different parameters: the length-scale λ capturing the extent of generalization, the exploration bonus β describing the extent of directed exploration, and the temperature parameter τ modulating random exploration. These three parameters directly correspond to the three postulated mechanisms of developmental differences in complex decision making tasks and can also be robustly recovered (see Fig. C.2).

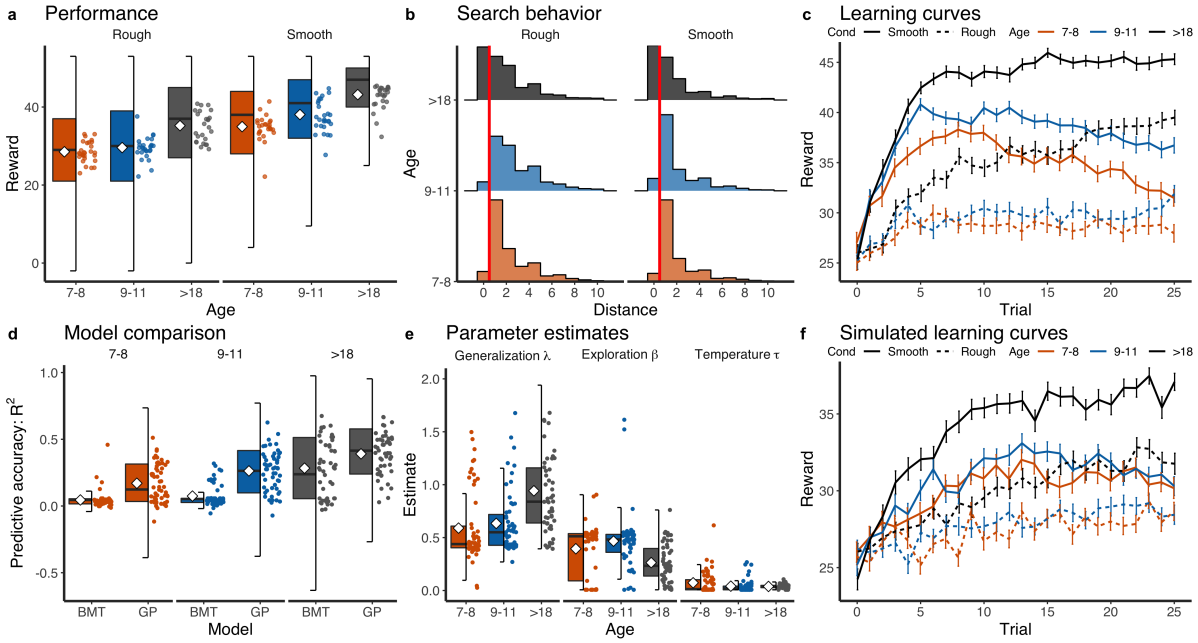

Fig. 4.2 Experiment 4 main results. **a**) Tukey box plots of rewards, showing the distribution of all choices for all participants, with the horizontal line representing the median and box showing the interquartile range of the distribution. Each dot is the participant-wise mean and diamonds indicate group means. **b**) Histograms of distances between consecutive choices by age group and condition, with a distance of zero corresponding to a repeat click. The vertical red line marks the difference between a repeat click and sampling a different option. **c**) Mean reward over trials by condition (solid lines for smooth and dashed lines for rough environments) and age group (color). Error bars indicate the standard error of the mean. **d**) Tukey box plots showing the results of the model comparison between Gaussian Process (GP) and Bayesian Mean Tracker (BMT) models by age group. Each point is a single subject and group means are shown as a diamond. **e**) Tukey box plot of cross-validated parameters retrieved from the GP-UCB model by age group, where each point is the mean estimate per subject and diamonds indicate the group means. Outliers are removed for readability, but are included in all statistical tests (see Appendix C.2). **f**) Learning curves simulated by GP-UCB model using mean participant parameter estimates. Error bars indicate the standard error of the mean.

4.4 Results

4.4.1 Behavioral results

Participants gained higher rewards in smooth than in rough environments (Fig. 4.2a; $t(158) = 10.51$, $p < .001$, $d = 1.66$, $BF > 100$), suggesting they made use of the spatial correlations and performed better when correlations were stronger. Adults performed better than older children (Fig. 4.2a; $t(103) = 4.91$, $p < .001$, $d = 0.96$, $BF > 100$), who in turn performed somewhat better than younger children ($t(108) = 2.42$, $p = .02$, $d = 0.46$, $BF = 2.68$). Analyzing the distance between consecutive choices (Fig. 4.2b) revealed that participants sampled more locally (smaller distances) in smooth compared to rough environments ($t(158) = -3.83$, $p < .001$, $d = 0.61$, $BF > 100$). Adults sampled more locally than older children ($t(103) = -3.9$,

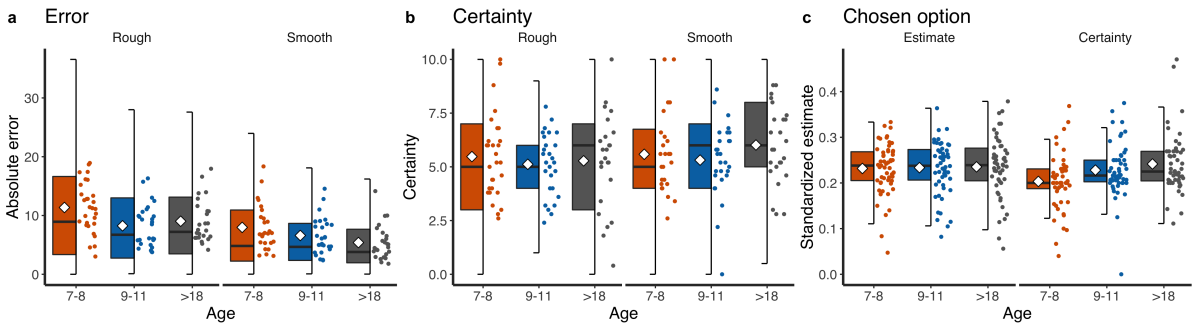

Fig. 4.3 Bonus round results. **a**) Absolute error of participant predictions about the rewards of unobserved tiles. **a**)Certainty judgments, where 0 is least certain and 10 is most certain. **a**) Standardized predictions and certainty estimates, which shows how much the estimated reward and certainty influenced choice (relative to judgments about non-chosen options). All figures show Tukey box plots (over all data points), with participant means as dots and group means as diamonds.

$p < .001$, $d = 0.76$, $BF > 100$), but there was no difference between younger and older children ($t(108) = 1.76$, $p = .08$, $d = 0.34$, $BF = 0.80$). Importantly, adults sampled fewer unique options than older children (14.5 vs. 21.7; $t(103) = 6.77$, $d = 1.32$, $p < .001$, $BF > 100$), whereas the two children groups did not differ in how many options they sampled (21.7 vs. 22.7; $t(108) = 1.27$, $d = 0.24$, $p = .21$, $BF = 0.4$).

Looking at the learning curves (i.e., average rewards over trials; Fig. 4.2c), we found a positive rank-correlation between mean rewards and trial number (Spearman's $\rho = .12$, $t(159) = 6.12$, $p < .001$, $BF > 100$). Although this correlation did not differ between the rough and smooth condition ($t(158) = -0.43$, $p = .67$, $d = 0.07$, $BF = 0.19$), it was significantly higher for adults than for older children (0.29 vs 0.08, $t(103) = 5.90$, $p < .001$, $d = 1.15$, $BF = 0.19$), $BF > 100$). The correlation between trials and rewards did not differ between younger and older children (0.04 vs 0.08; $t(108) = -1.87$, $p = .06$, $d = 0.36$, $BF = 0.96$). Therefore, adults learned faster, while children explored more extensively (see Appendix C.7 for further behavioral analyses).

4.4.2 Model comparison

We compared the GP-UCB model with an alternative model that does not generalize across options but is a powerful Bayesian model for reinforcement learning across independent reward distributions (*Bayesian Mean Tracker*; BMT, see subsection 2.3.1). Model comparisons are based on leave-one-round-out cross-validation error, where we fit each model combined with the UCB sampling strategy to each participant using a training set omitting one round, and then assess predictive performance on the hold-out round. Repeating this procedure for every participant and all rounds (apart from the tutorial and the bonus rounds), we calculated the standardized *predictive accuracy* for each model (pseudo-R^2 comparing out-of-sample log

loss to random chance), where 0 indicates chance-level predictions and 1 indicates theoretically perfect predictions (see Table C.1 for full model comparison with additional sampling strategies).

The results of this comparison are shown in Fig. 4.2d. The GP-UCB model predicted participants' behavior better overall ($t(159) = 13.28$, $p < .001$, $d = 1.05$, $BF > 100$), and also for adults ($t(49) = 5.98$, $p < .001$, $d = 0.85$, $BF > 100$), older ($t(54) = 10.92$, $p < .001$, $d = 1.48$, $BF > 100$) and younger children ($t(54) = 6.77$, $p < .001$, $d = 0.91$, $BF > 100$). The GP-UCB model predicted adults' behavior better than that of older children ($t(103) = 4.33$, $p < .001$, $d = 0.85$, $BF > 100$), which in turn was better predicted than behavior of younger children ($t(108) = 3.32$, $p = .001$, $d = 0.63$, $BF = 25$).

4.4.3 Developmental differences in parameter estimates

We analyzed the mean participant parameter estimates of the GP-UCB model (Fig. 4.2e) to assess the contributions of the three mechanisms (generalization, directed exploration, and random exploration) towards developmental differences.

We found that adults generalized more than older children, as indicated by larger λ-estimates (Mann-Whitney-$U = 2001$, $p < .001$, $r_\tau = 0.32$, $BF > 100$), whereas the two groups of children did not differ significantly in their extent of generalization ($U = 1829$, $p = .06$, $r_\tau = 0.15$, $BF = 1.7$). Furthermore, older children valued the reduction of uncertainty more than adults (i.e., higher β-values; $U = 629$, $p < .001$, $r_\tau = 0.39$, $BF > 100$), whereas there was no difference between younger and older children ($U = 1403$, $p = .51$, $r_\tau = 0.05$, $BF = 0.2$). Critically, whereas there were strong differences between age groups for the parameters capturing generalization and directed exploration, there was no reliable difference in the softmax temperature parameter τ, with no difference between older children and adults ($W = 1718$, $p = .03$, $r_\tau = 0.17$, $BF = 0.7$) and only anecdotal differences between the two groups of children ($W = 1211$, $p = .07$, $r_\tau = 0.14$, $BF = 1.4$). This suggests that the amount of random exploration did not reliably differ by age group (see Appendix C.1 for other implementations of random exploration). Thus, our modeling results converge on the same conclusion as the behavioral results. Children explore more than adults, yet instead of exploring randomly, children's exploration behavior seems to be directed toward options with high uncertainty. Additionally, our parameter estimates are robustly recoverable (see Fig. C.2) and can be used to simulate learning curves that reproduce the differences between the age groups as well as between smooth and rough conditions (Fig. 4.2f).

4.4.4 Bonus round

In the bonus round, each participant predicted the expected rewards and the underlying uncertainty for five randomly sampled unrevealed tiles after having made 15 choices on the grid. We first calculated the mean absolute error between predictions and the true expected value of rewards (Fig. 4.3a). Prediction error was higher for rough compared to smooth environments ($t(158) = 4.93$, $p < .001$, $d = 0.78$, $BF > 100$), reflecting the lower degree of spatial correlation that could be utilized to evaluate unseen options. Surprisingly, older children were as accurate as adults ($t(103) = 0.28$, $p = .78$, $d = 0.05$, $BF = 0.2$), but younger children performed worse than older children ($t(108) = 3.14$, $p = .002$, $d = 0.60$, $BF = 15$). Certainty judgments did not differ between the smooth and rough environments ($t(158) = 1.13$, $p = .26$, $d = 0.18$, $BF = 0.2$) nor between the different age groups (max-$BF = 0.1$).

Of particular interest is how judgments about the expectation of rewards and perceived uncertainty relate to the eventual choice from amongst the five options (implemented as a 5-alternative forced choice). We standardized the estimated reward and confidence judgment of each participant's chosen tile by dividing by the sum of the estimates for all five options (Fig. 4.3c). Thus, larger standardized estimates reflect a larger contribution of either high reward or high certainty on the choice. Whereas there was no difference between age groups in terms of the estimated reward of the chosen option (max-$BF = 0.1$), we found that younger children preferred options with higher uncertainty slightly more than older children ($t(108) = 2.22$, $p = .03$, $d = 0.42$, $BF = 1.8$), and substantially more than adults ($t(103) = 2.82$, $p = .006$, $d = 0.55$, $BF = 7$). This further corroborates our previous analyses, showing that the sampling behavior of children is more directed toward uncertain options than that of adults.

4.5 Discussion

We examined three potential sources of developmental differences in a complex learning and decision-making task: random exploration, directed exploration, and generalization. Using a paradigm that combines both generalization and search, we found that adults gained higher rewards and exploited more strongly, whereas children sampled more unique options, thereby gaining lower rewards but exploring the environment more extensively. Using a computational model with parameters directly corresponding to the three hypothesized mechanisms of developmental differences, we found that children generalized less and were guided by directed exploration more strongly than adults. They did not, however, explore more randomly than adults.

Our results shed new light on the developmental trajectories in generalization and exploration, casting children not as merely more random sampling behavior, but as directed

explorers who are hungry for information in their environment. Our conclusions are drawn from converging evidence combining analysis of behavioral data and computational modeling. Moreover, our findings are highly recoverable and also hold for other formalizations of random exploration instead of using the softmax temperature parameter (see Appendix section C.1).

Related work by Somerville et al. (2017) also found no developmental difference in random exploration, but in contrast, found increased directed exploration across early adolescence, which stabilized in adulthood. We believe that our results are not necessarily incompatible with that finding. Somerville and colleagues defined directed exploration using horizon-sensitive exploration (i.e,. strategic planning of exploration), whereas we define directed exploration as uncertainty-guided exploration via a greedy upper confidence bound algorithm. Thus, children may have higher tendencies towards directed exploration in a stepwise-greedy fashion, but fail to exhibit such tendencies when planning ahead for multiple steps, perhaps due to cognitive limitations. This opens up further possibilities for studying different mechanisms of directed exploration and how they relate to one another.

Our results provide strong evidence for developmental differences in directed exploration driven by both expected rewards and the associated uncertainty. These findings complement existing research on age-related differences in risk- and uncertainty-related behavior (Josef et al., 2016). For instance, adolescents and adults systematically differ in their tolerance of options with outcomes that have unknown probabilities, providing converging evidence that uncertainty is valued differently depending on age (Tymula et al., 2012). Importantly, in our task a sampling strategy that only seeks to reduce uncertainty (PureExplore; see subsection 2.4.2) is inferior to the "optimistic" UCB strategy in predicting children's and adults' behavior (Table C.1). This result demonstrates how reward expectations and uncertainty interact to produce decision-making behavior that balances the exploration-exploitation trade-off adaptively as a function of age. Nonetheless, future work could attempt to further disentangle different interpretations of uncertainty seeking formally, for example, by not familiarizing participants with the underlying environments or by manipulating the variance of outcomes directly.

Furthermore, it is surprising that there were no meaningful differences between the parameter estimates of younger and older children. Since this indicates that directed exploration might emerge even earlier than expected, future studies could apply our paradigm to investigate exploration behavior in even younger children.

Our results showing a developmental increase in generalization can also be related to previous findings showing a developmental increase in the use of task structure knowledge in model-based reward learning (Decker, Otto, Daw, & Hartley, 2016). Because the generalization parameter λ can be mathematically equated to the speed of learning about the underlying function (Sollich, 1999), generalization and learning are inextricably linked in our task. There

are however other uses of the term "generalization" in the psychological literature. For example, children are known to generalize words or categories more broadly, a tendency that decreases over time, trading-off with the capacity to form more precise episodic memories (Keresztes et al., 2017). While we focus on generalization in the sense used by Shepard (i.e., generalization across stimuli), it is an outstanding question how this type of generalization relates to word and category generalization. It would be a fruitful avenue for future research to connect these two domains in a unifying theory of generalization.

In our current study, we have assessed only environments with stationary reward distributions. However, given that children displayed increased exploration behavior, we believe that they could perform especially well in environments that change over rounds. Whether or not children would outperform adults in changing environments remains an important question for future research. Ultimately, our results suggest that to fulfill Alan Turing's dream of creating a child-like AI, we need to incorporate generalization and curiosity-driven exploration mechanisms (Riedmiller et al., 2018).

Chapter 5

From Spatial to Conceptual Search

Everything that distinguishes humanity from animals depends on the ability to volatilize perceptual metaphors into a schema, to dissolve images into a concept.

—Nietzsche

The idea of a "cognitive map" was originally developed to explain planning and generalization in spatial domains as a representation of inferred relationships between experiences. More recently, research has suggested that a common neurological substrate governs the organization of both spatial and conceptual knowledge. Is there a common computational framework for understanding generalization and search across domains?

We present two studies using a within-subject design to compare individual search behavior in successive tasks, where either spatial or conceptual features predicted rewards. Converging evidence from search behavior, judgments about unobserved options, and model parameters indicate similar principles govern how people generalize and search for rewards. However, we also find some intriguing differences. Participants showed a reduced capacity for generalization in the conceptual domain, that led to less directed exploration in favor of increased random exploration. These results enrich our understanding of how people represent and navigate the spatial and conceptual worlds we live in.

This chapter is based on the following publication in addition to new unpublished research:

Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2018). Connecting conceptual and spatial search via a model of generalization. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1183–1188). Austin, TX: Cognitive Science Society.

5.1 Introduction

The ability to search for rewards comes in many shapes. We can wander through a foreign city in search of new and delicious foods, or search through an online store to find a laptop with the features that we like. We can even skim over parts of a thesis chapter to find sections more interesting than the introduction. While these tasks differ in a number of ways, all of them require the exploration of possibilities and the use of generalization to predict outcomes of unexplored options. Here, we ask if generalization and search in different domains employ common computational mechanisms.

Breaking from the classical stimulus-response school of reinforcement learning, Tolman (1948) argued that both rats and humans extract a cognitive representation from experience, described as a "cognitive map" of the environment. Rather than merely representing stimulus-response associations, cognitive maps also encode inferred relationships between experiences or options, such as the distances between locations in space, thereby facilitating planning ahead and generalization. While cognitive maps were first identified as representations of physical spaces, Tolman also hypothesized that similar principles may underlie the organization of knowledge more broadly (Tolman, 1948).

The idea of a cognitive map has been widely adopted in research on brain signals underlying spatial navigation (O'Keefe & Nadel, 1978). In a similar vein, studies on reinforcement learning have emphasized that relations between states may also be encoded as a cognitive map (Behrens et al., 2018; Schuck, Cai, Wilson, & Niv, 2016; Stachenfeld et al., 2017). Most recently, a number of studies suggest that the same neural representations may underlie the organization of spatial and non-spatial relational information in the brain (Constantinescu et al., 2016; Garvert, Dolan, & Behrens, 2017; Kaplan, Schuck, & Doeller, 2017). This is consistent with behavioral evidence for generalized cognitive search processes (Hills et al., 2008) in both semantic and spatial domains. Thus, a common representation (in the form of a cognitive map) may be the basis for generalized search processes operating on multiple domains. Whereas the capacity to *generalize* from past experiences to unobserved states and actions has been studied for decades (e.g., Shepard, 1987), the link between our ability to generalize in a wide range of tasks and the encoding of experiences in a cognitive map-like format has not yet been explored.

To assess this link, we investigate how people search for rewards in both spatially and conceptually correlated multi-armed bandit tasks (E. Schulz, Konstantinidis, & Speekenbrink, 2017; Stojic et al., 2015; Wu, Schulz, et al., 2017; Wu, Schulz, Speekenbrink, et al., 2018). In two experiments employing a longitudinal design, we study how participants perform search guided by either spatial or conceptual features. In spatial domains, the spatial correlation of rewards can be used to guide generalization based on the spatial similarity between options. In conceptual domains, participants can use feature-based similarity to navigate conceptual space

in the search for rewards. In both domains and in both experiments, search takes place in state spaces larger than the available search horizon and with noisy rewards, thereby inducing an exploration-exploitation dilemma. Generalization is constrained by the level of environmental correlation, which we vary between participants.

Our results show that participants are able to learn and generalize in both spatial and conceptual domains, with correlated performance across the two tasks. In Experiment 5, where both spatial and conceptual features were shown simultaneously (with one relevant and the other irrelevant), we find a task order effect, where performing the spatial task first boosted performance in the conceptual task, but not vice versa. We also find that both performance and model parameters were correlated across the two tasks, indicating that participants who generalized more or explored more eagerly in one domain, also did so in the other. However, there were also systematic differences, indicating that the irrelevant spatial features were more difficult to ignore and that participants may have had a linear prior on the conceptual features (i.e., assuming more leaves or berries predicted higher reward). Experiment 6 addressed potential confounds by showing only task-specific features and changing the conceptual stimuli for Gabor patches. Here we find that a Gaussian Process (GP) model of generalization best predicts behavior in both domains, predicts judgments of both reward and uncertainty, and can simulate learning curves replicating differences in tasks and environment manipulations. We also find a intriguing difference in model parameters, indicating participants reduced directed exploration and increased random exploration in the conceptual domain. Overall, our results provide important clarification about the similarities and differences in how people represent and search in spatial and conceptual spaces.

5.2 Experiment 5: Branch search

Participants searched for rewards in two successive multi-armed bandit tasks with correlated rewards (Fig. 5.1). In one task, rewards were spatially correlated (*Spatial task*), meaning options with similar spatial locations yielded similar rewards. In the other task, rewards were conceptually correlated (*Conceptual task*), such that options with similar features (i.e., the number of leaves $\in [1,5]$ and the number of berries $\in [1,5]$) yielded similar rewards. Figure 5.1c,d shows examples of fully revealed environments representing the same underlying reward function, but mapped to either spatial or conceptual features. In both tasks, the search space was represented by a 5×5 two-dimensional grid, where each of the 25 options represented a different arm of the bandit, which could be clicked to obtain (noisy) rewards. Each tile of the grid contained one of 25 unique conceptual stimuli, which were randomly shuffled between rounds and always visible. Thus, we presented information about both spatial and conceptual

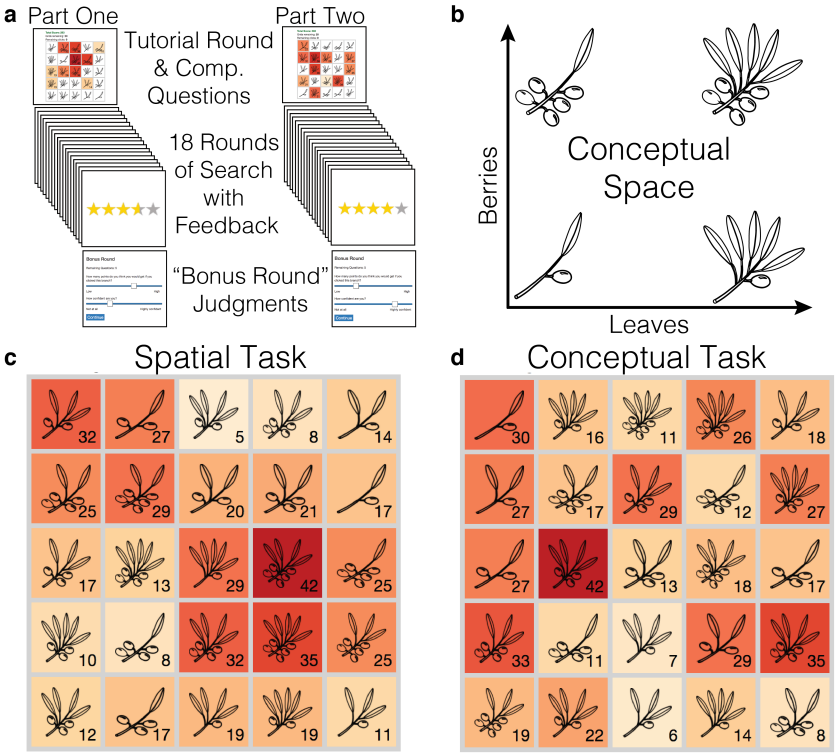

Fig. 5.1 Experiment 5 design. **a**) Searching for spatially or conceptually correlated rewards in a two part multi-armed bandit experiment, where task order was counter-balanced across subjects. **b**) Illustration of conceptual space, with four edge cases shown. **c, d**) Two fully revealed environments (representing the same underlying reward function), represented as a 5×5 grid, where participants could click on the tiles to obtain rewards, revealing a numeric payoff value and a corresponding color aid (larger rewards are darker; unexplored tiles are initially white). **c**) Spatial search task, where rewards are spatially correlated, with nearby tiles yielding similar rewards. **d**) Conceptual search task, where options with high feature similarity (i.e., number of leaves and number of berries) had similar rewards, independent of spatial location.

features in both search tasks, but only one of them was relevant for generalization and predicting rewards. At the beginning of each round only a single randomly chosen option was revealed (i.e., displayed the numerical reward and corresponding color aid), whereby subjects had a limited horizon of 10 actions in each round (40% of the total search space; similar to Wu, Schulz, et al., 2017; Wu, Schulz, Speekenbrink, et al., 2018), thereby inducing an exploration-exploitation trade-off.

5.2.1 Methods

Participants and Design.

72 participants were recruited through Amazon Mechanical Turk for a two part experiment (requiring 95% approval rate and 100 previously approved HITs, while excluding those who participated in Experiments 1-3), where only those who completed part one were invited for

part two. In total, 64 participants completed both parts of the experiment and were included in the analyses (26 Female; mean age=34, SD=11). Participants were paid $1.25 for each part of the experiment, with those completing both parts being paid an additional performance-contingent bonus of up to $3.00. Participants earned $4.94 $\pm$ 0.29 and spent 26 $\pm$ 13 minutes completing both parts. There was an average gap of 5.6 $\pm$ 4.7 hours between the two parts of the experiment.

Task order varied between subjects, with participants completing the Spatial and Conceptual task in counterbalanced order in separate sessions. We also varied between subjects the extent of reward correlations in the search space by randomly assigning participants to one of two different classes of environments (*Smooth* vs. *Rough*), with smooth environments corresponding to stronger correlations, and the same environment class used for both tasks.

Materials and Procedure

Each search task comprised 20 rounds (i.e., grids), with a different reward function sampled without replacement from the set of assigned environments. The reward function specified how rewards mapped onto either the spatial or conceptual features. Between rounds, the locations of each stick stimuli were randomly shuffled. In each round, participants had a limited search horizon of 10 available actions (i.e., clicks), which could be used to either explore unrevealed options or to exploit known options. Participants were instructed to accumulate as many points as possible, which were later converted into monetary payoffs.

For both tasks, the first round was an interactive tutorial and the last round was a "bonus round" (Fig. 5.1a). In the *tutorial round*, participants were shown instructions for the given task alongside an interactive grid, which functioned identically to subsequent rounds. Participants were told that options with either similar spatial features (Spatial task) or similar conceptual features (Conceptual task) would yield similar rewards. Three comprehension questions (different for spatial and conceptual tasks) were used to ensure full understanding of the task (specifically whether spatial or conceptual features predicted reward) before participants were allowed to continue. In the *bonus round*, participants made explicit judgments about the expected rewards and their estimated uncertainty of five unrevealed tiles in the middle of the round (i.e., after five clicks), in order to tap into beliefs supported by generalization. All behavioral and computational modeling analyses exclude the tutorial and bonus rounds, except for the analysis of the bonus round judgments.

Spatial and Conceptual Search Tasks

At the beginning of each round, one random tile was revealed (i.e., showing numerical payoff and color aid) and participants could click any of the 25 tiles in the grid until the search horizon of ten clicks was exhausted, including re-clicking previously revealed tiles. Clicking an unrevealed tile displayed the numerical value of the reward along with a corresponding color aid, where darker colors indicated higher point values. Previously revealed tiles could also be re-clicked, although there were variations in the observed value due to noise. Each observation included normally distributed noise, $\varepsilon \sim \mathcal{N}(0,1)$, where the rewards for each round were scaled to a uniformly sampled maximum value in the range of 35 to 45, so that the value of the global optima could not be easily guessed. For repeat clicks, the most recent observation was displayed numerically, while hovering over the tile would display the entire history of observations. The color of the tile corresponded to the mean of all previous observations.

Participants were awarded up to five stars based on their performance at the end of each round (e.g., 4.4 out of 5), based on the ratio of their average reward to the global maximum. The performance bonus (up to $3.00) was calculated based on the average number of stars earned in each round, excluding the tutorial round.

Judgments about expected reward and confidence

In both tasks the last round was a "bonus round", which solicited judgments about the expected reward and estimated uncertainty of five unrevealed options. Participants were informed that the goal of the task remained the same (maximize cumulative rewards), but that after five clicks, they would be asked to provide judgments about five randomly selected options, which had not yet been explored (sampled uniformly from unexplored options). Judgments about expected rewards were elicited using a slider, which changed the displayed value and color of the selected tile from 0 to 50 (in increments of 1). Judgments about uncertainty were elicited using a slider from 0 to 10 (in increments of 1), with the endpoints labeled 'Not at all' and 'Highly confident'. After providing the five judgments, participants were asked to choose one of the five selected options to reveal, and subsequently completed the round like all others.

Environments

All environments were sampled from a GP prior parameterized with a *radial basis function* (RBF) kernel (see below for details), where the length-scale parameter (λ) determines the rate at which the correlations of rewards decay over (spatial or conceptual) distance. Higher λ-values correspond to stronger correlations. We generated 40 samples of each type of environments, using $\lambda_{Smooth} = 2$ and $\lambda_{Rough} = 1$, which was identical to Experiments 1-2. On each round, one

environment was sampled without replacement from the assigned class of environments and used as the underlying reward function in each task.

5.2.2 Modeling Generalization and Exploration

We use a combination of a *learning model* with a *decision strategy* to make predictions about each individual participant's search decision. The learning model forms beliefs about the expectations of rewards $m(\mathbf{x})$ and the associated uncertainty $v(\mathbf{x})$ for each option $\mathbf{x}$, which are then used by the decision strategy to make probabilistic predictions about search decisions. We apply leave-one-round-out cross validation to estimate the free parameters of our models and use out-of-sample model predictions to compare models for predicting human search behavior. Additionally, we compare predictions of the learning models to judgments made by participants about the expected reward and estimated uncertainty of five unrevealed options.

Learning models

Function Learning. We use Gaussian Process (GP) regression (Rasmussen & Williams, 2006; E. Schulz, Speekenbrink, & Krause, 2018) as a *Function Learning* model for inducing an underlying value function mapping the features of the search task onto rewards, as a method for generalization. A $\mathcal{GP}$ defines a distribution $P(f)$ over possible functions $f(\mathbf{x})$ that map inputs $\mathbf{x}$ to output y. In our case, either the spatial features (i.e., x- and y-coordinates on the grid) or conceptual features (i.e., number of leaves and berries) of each option serve as inputs $\mathbf{x}$ to predict reward y. Crucially, learning a value function by using either spatial or conceptual similarity allows for predictive generalization of unobserved options (see Wu, Schulz, Speekenbrink, et al., 2018).

A $\mathcal{GP}$ is completely defined by a mean function $m(\mathbf{x})$ and a kernel function, $k(\mathbf{x},\mathbf{x}')$:

$$m(\mathbf{x}) = \mathbb{E}\left[f(\mathbf{x})\right] \tag{5.1}$$

$$k(\mathbf{x},\mathbf{x}') = \mathbb{E}\left[(f(\mathbf{x}) - \mu(\mathbf{x}))(f(\mathbf{x}') - \mu(\mathbf{x}'))\right] \tag{5.2}$$

We fix the prior mean to the median value of payoffs, $m(\mathbf{x}) = 25$, while the kernel function $k(\mathbf{x},\mathbf{x}')$ encodes prior assumptions (or inductive biases) about the underlying function. Here, we use the *radial basis function* (RBF) kernel:

$$k_{\text{RBF}}(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{||\mathbf{x}-\mathbf{x}'||^2}{2\lambda^2}\right) \tag{5.3}$$

The RBF kernel models similarity by assuming correlations between two options $\mathbf{x}$ and $\mathbf{x}'$ decay as an exponential function of their (spatial or conceptual) distance. The length-scale parameter λ determines how far correlations extend, with larger values of λ assuming stronger correlations over longer distances, whereas $\lambda \to 0^+$ assumes complete independence of options. We use recovered parameter estimates of λ to learn about the extent to which each participant generalize about unobserved rewards.

Option Learning. The *Option Learning* model uses a Bayesian Mean Tracker (BMT) and is an associative learning model (Speekenbrink & Konstantinidis, 2015). In contrast to the GP Function Learning model, the Option Learning model learns the rewards of each option independently by computing independent posterior distributions for the mean μ_j for each option j:

$$p(\mu_{j,t}|\mathcal{D}_{t-1}) = \mathcal{N}(m_{j,t}, v_{j,t}) \tag{5.4}$$

The rewards of each option j are learned independently, with the posterior mean $m_{j,t}$ and variance $v_{j,t}$ only updated when selected at trial t, based on the observed reward y_t:

$$m_{j,t} = m_{j,t-1} + \delta_{j,t} G_{j,t} \left[y_t - m_{j,t-1} \right] \tag{5.5}$$

$$v_{j,t} = \left[1 - \delta_{j,t} G_{j,t} \right] v_{j,t-1} \tag{5.6}$$

where $\delta_{j,t} = 1$ if option j was chosen on trial t, and 0 otherwise. Additionally, the learning factor $G_{j,t}$ is defined as:

$$G_{j,t} = \frac{v_{j,t-1}}{v_{j,t-1} + \theta_\varepsilon^2} \tag{5.7}$$

where θ_ε^2 is the error variance, which is estimated as a free parameter. Intuitively, the estimated mean of the chosen option $m_{j,t}$ is updated based on the prediction error $y_t - m_{j,t-1}$, multiplied by the learning factor $G_{j,t}$. At the same time, the estimated variance $v_{j,t}$ is reduced by a factor of $1 - G_{j,t}$, which is in the range $[0,1]$. The error variance θ_ε^2 can be interpreted as an inverse sensitivity, where smaller values result in more substantial updates to the mean $m_{j,t}$, and larger reductions of uncertainty $v_{j,t}$. We set the prior mean to the median value of payoffs $m_{j,0} = 25$ and the prior uncertainty $v_{j,0} = 250$.

Decision Strategy

Both Function Learning and Option learning models generate normally distributed predictions about the expected reward $m_t(\mathbf{x})$ and estimated uncertainty $v_t(\mathbf{x})$ for each option. These estimates are used by the decision strategy for evaluating the quality $q(\mathbf{x})$ of each option and making a prediction about where to sample next. We use *Upper Confidence Bound sampling* (UCB) to compute a weighted sum of the expected reward $m(\mathbf{x})$ and the estimated uncertainty $s(\mathbf{x})$:

$$q_{UCB}(\mathbf{x}) = m(\mathbf{x}) + \beta\sqrt{v(\mathbf{x})} \tag{5.8}$$

where the exploration factor β determines how the reduction of uncertainty trades off against exploiting high expected rewards; a strategy that has been found to predict search behavior in a variety of contexts (E. Schulz, Konstantinidis, & Speekenbrink, 2017; Wu, Schulz, Speekenbrink, et al., 2018).

We then use a softmax function to convert the value of an option $q(\mathbf{x})$ into a choice probability:

$$P(\mathbf{x}) = \frac{\exp(q(\mathbf{x})/\tau)}{\sum_{j=1}^{N}\exp(q(\mathbf{x}_j)/\tau)} \tag{5.9}$$

where τ is the temperature parameter. Whereas β encodes exploration directed towards uncertain options, τ encodes undirected (noisy) exploration as a distinct (Wilson et al., 2014) and separately recoverable (Wu, Schulz, Speekenbrink, et al., 2018) phenomenon. As $\tau \to 0$ the highest-value arm is chosen with a probability of 1 (i.e., argmax); when $\tau \to \infty$, predictions converge to random choice.

5.2.3 Behavioral results

Overall, rewards were lower in the conceptual task than in the spatial task (paired t-test: $t(63) = -3.7$, $p < .001$, $d = 0.34$, $BF = 50$), although Figure 5.2a shows how this is largely due to an effect of task order. When the spatial task was performed first (Fig. 5.2a left), we find no performance differences between the two tasks ($t(34) = 0.6$, $p = .55$, $d = 0.07$, $BF = 0.2$). However, in the reverse order (conceptual first), mean rewards were lower in the conceptual task than in the spatial task ($t(28) = -5.6$, $p < .001$, $d = 0.68$, $BF > 100$). Thus, searching first for spatially correlated rewards improved performance in the conceptual domain (two-sample t-test: $t(62) = 2.6$, $p = .01$, $d = 0.66$, $BF = 4.6$), but not vice versa. Across the two tasks,

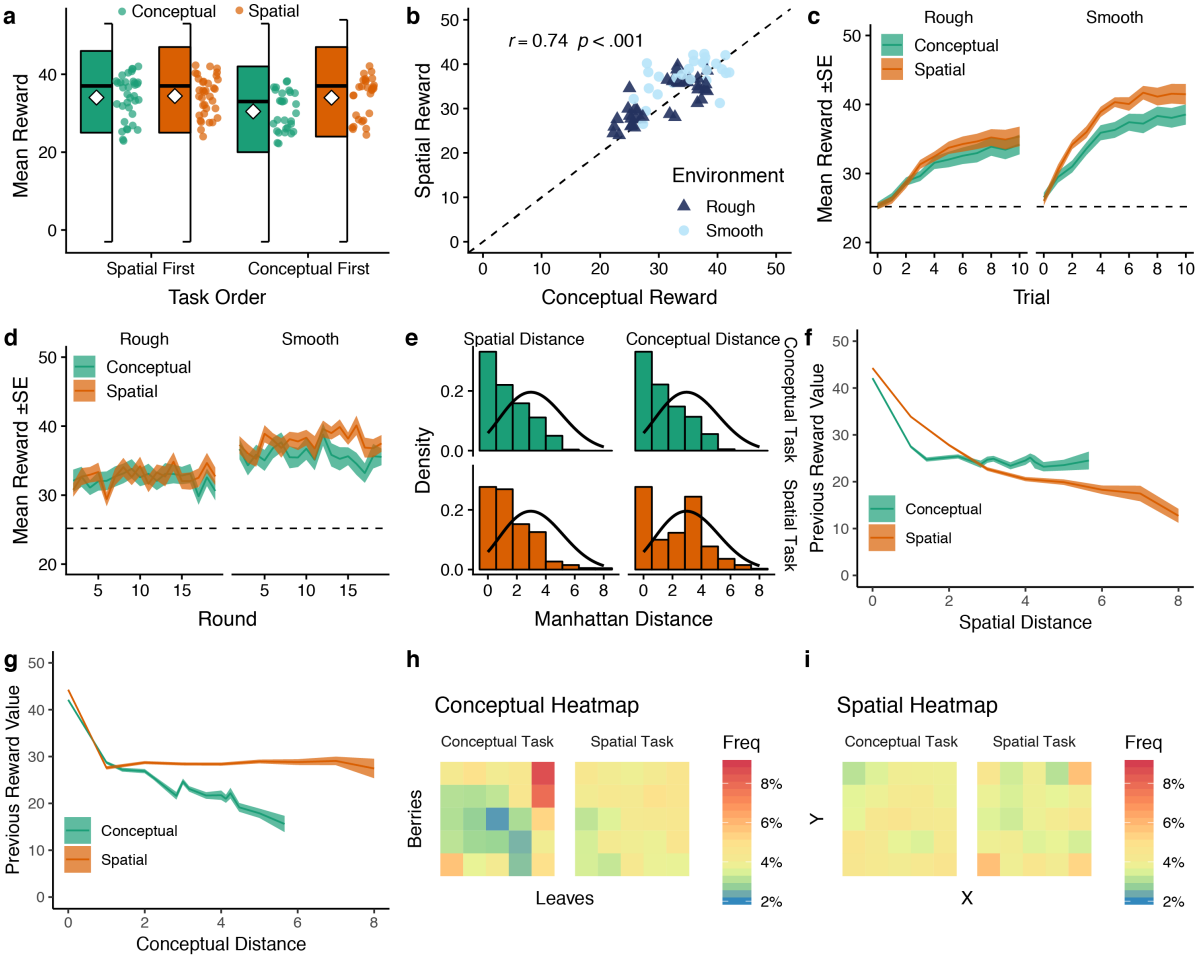

Fig. 5.2 Experiment 5 behavioral results. **a**) Mean performance, where each dot is a single participant, and the boxplot indicates the median and 1.5 IQR, with the diamond showing group means. Comparing performance across the counter-balanced task order, we found performing the spatial task first boosted performance on the conceptual task. **b**) Performance in the spatial and the conceptual tasks were correlated, where each point is a single participant, with the dashed line indicating $y = x$. **c**) Learning over trials and **d**) learning over rounds, where lines indicate aggregated means and the ribbon shows the standard error. The dashed line provides comparison to a random baseline. **e**) Histogram of distances between choices, measured either in terms of spatial distance (left column) or conceptual distance (right column). The black line provides a comparison to a random baseline (10k replications). **f-g**) Relationship between the value of the previous reward and the Manhattan distance to the next selected option (f: spatial distance; g: conceptual distance), showing mean (line) and standard error (ribbon). **h-i**) Heatmaps of chosen options (h: leaf and berry features; i: grid location) aggregated over all participants, where the color shows the frequency of each option relative to the task (left column: conceptual; right column: spatial). Colors are centered on yellow representing random chance (1/25), with orange and red indicating higher than chance, while green and blue were lower than chance.

performance was highly correlated (Pearson's $r = .74$, $p < .001$, $BF > 100$; Fig. 5.2b), with participants who performed better in one task, also performing better in the other.

The learning curves in Figure 5.2c show that participants systematically found higher rewards over subsequent trials ($r = .51$, $p < .001$, $BF > 100$), performed better in smooth than in rough environments ($t(62) = 4.4$, $p < .001$, $d = 1.1$, $BF > 100$), and that the performance

gap between spatial and conceptual performance was larger in smooth environments. Looking only at participants assigned to smooth environments, performance was better in the spatial task than in the conceptual task ($t(27) = 3.2$, $p = .003$, $d = 0.59$, $BF = 11$), consistent with the larger gap between learning curves in Figure 5.2c. We did not find any systematic improvements over rounds ($r = .02$, $p = .42$, $BF = 0.1$; Fig. 5.2d).

Patterns of search

Figure 5.2e shows the Manhattan distance between sequential choices, measured in either the spatial domain (distance between tiles) and in the conceptual domain (distance in feature space) for each task type (rows). Overall, we see that participants searched more locally than random chance (black line), with a high proportion of repeat clicks (distance of 0) and also locally similar options. Participants tended to make more repeat clicks in the conceptual task (distance of 0; $t(63) = 2.4$, $p = .02$, $d = 0.21$, $BF = 1.8$), whereas participants in the spatial task were more likely to search neighboring options (distance of 1; $t(63) = 4.0$, $p < .001$, $d = 0.37$, $BF > 100$). We also see the trend that participants in the conceptual task (top row) displayed a pattern of spatial locality (similar to conceptual locality), even though spatial features did not predict rewards. This was not the case on the spatial task, where the distribution of distances measured in the conceptual domain resembled the random baseline, with the exception of a high number or repeat samples (distance=0). Thus, participants displayed spatial "stickiness" (Gershman et al., 2009) in the conceptual task, but not vice versa.

Looking at the relationship between the value of a reward and the distance searched on the subsequent trial (Fig. 5.2f-g), we see that participants responded appropriately, with a stronger influence of reward value on spatial distance in the spatial task, and a stronger influence of reward on conceptual distance in the conceptual task. This suggests that participants used information from the relevant dimension (spatial or conceptual) to make their decisions, where lower rewards lead to larger jumps in the appropriate domain. Figures 5.2h-i show heatmaps of the frequency of selecting the different options, defined in terms of the conceptual features (number of leaves and berries) or spatial features (x-y coordinates on the grid). The colors of the heatmap tiles correspond to the frequency of selecting each option, with random chance (1/25) centered on yellow, while orange and red correspond to higher than chance, and green and blue correspond to lower than chance. The most pronounced trend is that participants in the conceptual task had a preference for the maximum number of leaves and berries (top right corner of heatmap), indicating a potential prior belief that larger feature values correspond to higher rewards, similar to the preference for positive linear functions found in the function learning studies (Kalish et al., 2007). All other heatmaps appear well distributed.

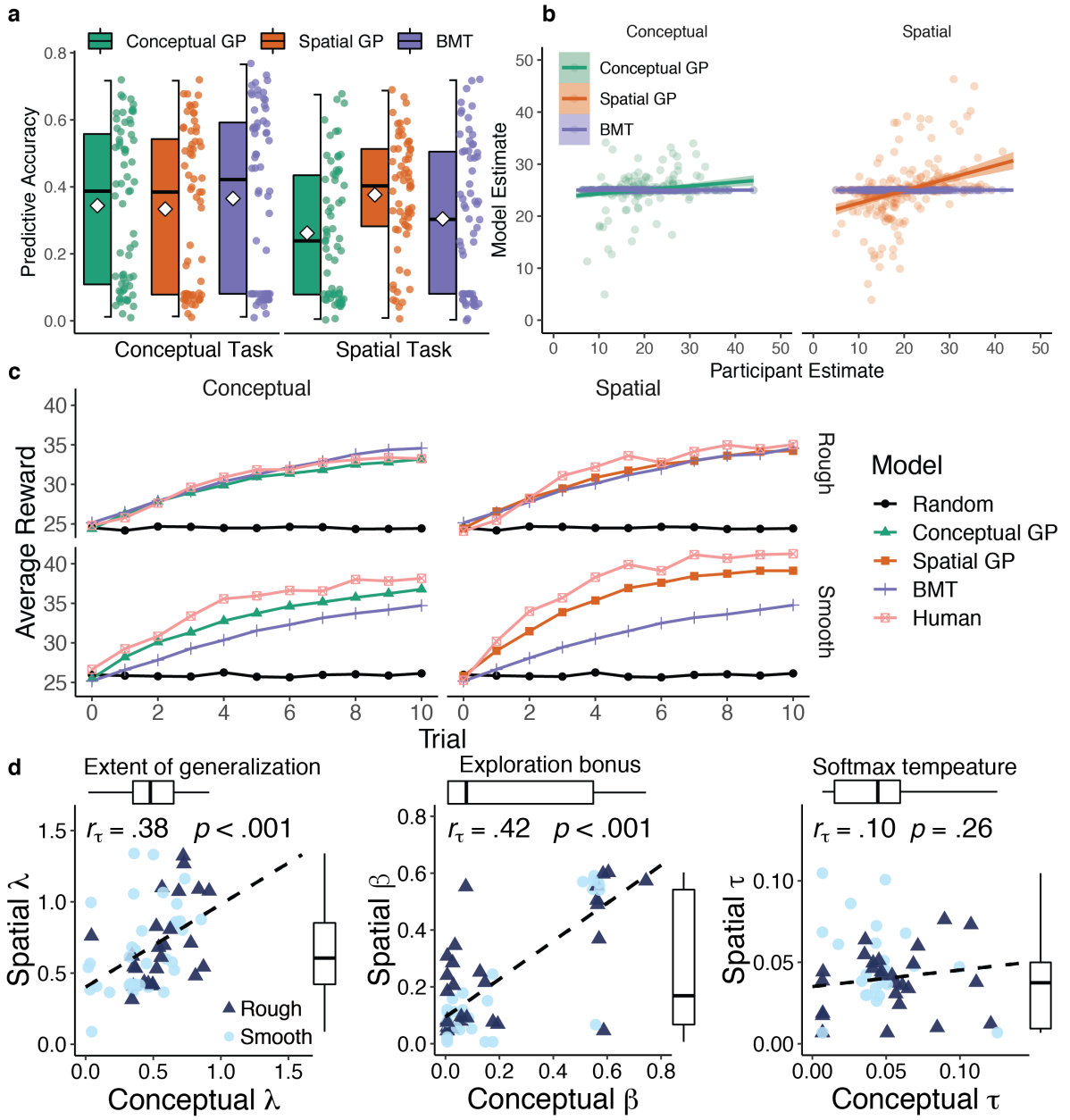

Fig. 5.3 Experiment 5 modeling results. **a**) Predictive accuracy of models on out-of-sample predictions, where 0 corresponds to random chance and 1 is a theoretically perfect model. Each dot is a single participant, and the boxplot indicate the median and 1.5 IQR, with the diamond showing group means. **b**) The relationship between participant estimates in the bonus round and model estimates (parameterized using the median participant parameter estimates). Each dot is a single participant, while the lines represent a linear regression (with ribbons showing standard error). **c**) Simulated learning curves (10k replications) using models specified with parameters sampled from participant estimates. We include comparison to a random model (black line) and human performance (pink). **d**) Median GP parameters from the spatial (y-axis) and the conceptual tasks (x-axis), where each point is a single participant and the dotted line shows a linear regression. Outliers are excluded from the plot but not from the rank correlations and boxplots (showing median and 1.5 IQR range).

5.2.4 Modeling results

How did participants use the respective spatial and conceptual features to guide search decisions? And how did they balance the exploration-exploration dilemma? We first compared models based on their ability to predict participants' behavior using leave-one-round-out cross validation (Fig. 5.3a), where the *Conceptual GP* and the *Spatial GP* utilize either conceptual or spatial features, respectively, while the BMT learns independent reward distributions for each option. In the spatial task, the Spatial GP performed better than both the Conceptual GP (i.e., using leaf and berry features; $t(63) = 8.5$, $p < .001$, $d = 0.61$, $BF > 100$) and the BMT ($t(63) = 4.6$, $p < .001$, $d = 0.36$, $BF > 100$) replicating previous findings reported in Wu, Schulz, Speekenbrink, et al. (2018). However, all models performed equally well in the conceptual task ($F(2,189) = 0.27$, $p = .76$, $BF = 0.07$).

Bonus round judgments

The correspondence between participant judgments from the bonus round and model predictions based on the median parameter estimates for each participant are shown in Figure 5.3b. The corresponding GP had lower error than the BMT in the spatial task ($t(63) = 2.2$, $p = .03$, $d = 0.2$, $BF = 1.2$), but there was no difference in the conceptual task ($t(63) = 0.9$, $p = .35$, $d = 0.05$, $BF = 0.2$). GP predictions were correlated with participant judgments in both the spatial task ($r = .38$, $p < .001$, $BF > 100$) and the conceptual task ($r = .21$, $p < .001$, $BF > 100$), whereas the BMT invariably predicted a mean of 25 for all unobserved options, making correlations undefined.

To understand how well GP uncertainty estimates predicted participant confidence judgments, we use a mixed-effects model with participant as a random effect to account for individual variability (e.g., some participants may generally have lower confidence judgments for all options). We found that high GP uncertainty estimates predicted low confidence judgments in the spatial task (standardized effect size: $\beta = -.14$, $t(312) = -2.8$, $p = .005$, $BF = 70$)[†], but not in the conceptual task ($\beta = -.02$, $t(284) = -0.6$, $p = .557$, $BF = 4$). Again the BMT invariably estimates the uncertainty of unobserved options based on the prior variance, making it equivalent to the null model that the GP is compared against. Thus, the Bayes Factors reported in the mixed-effects results also indicate the log-odds of the GP as a model of how people represent uncertainty, relative to the BMT.

[†]Note that β is the standardized effect size of a mixed effects model, and should not be confused with the exploration bonus of the UCB sampling strategy. See Appendix A for clarification on the reporting of statistical tests and Bayes Factor calculation used throughout the thesis.

Learning curves

We also simulate model performance on the task over 10,000 replications, where model parameters were sampled (with replacement) from the set of cross-validated participant estimates. Figure 5.3c) shows the aggregated learning curves together with human performance (pink) and a random baseline (black). Both the GP and BMT learning curves produced similarly good approximations of human performance in rough environments. However, the gap is much wider in smooth environments, where the GP closely trails human-like performance while the BMT simulations achieves a visibly slower rate of learning. The results of these simulations are heavily dependent on our parameter estimates, which from our model comparison (Fig. 5.3a) indicate there are some identifiability issues in the conceptual task. Thus, it is not surprising that the gap between the GP and BMT is strongest in the spatial task, but also amplified by the stronger reward correlations in smooth environments, which provide more traction for generalization.

Parameter estimates

Looking at the parameter estimates of the GP for the two different tasks (Fig. 5.3d), we find that λ (extent of generalization) and β (exploration bonus) were rank-correlated[†] within participants (λ correlation: $r_\tau = .38$, $p < .001$, $BF > 100$; β correlation: $r_\tau = .42$, $p < .001$, $BF > 100$), whereas the softmax temperature τ was not correlated ($r_\tau = .1$, $p = .26$, $BF = 0.33$) and did not reliably differ between tasks (Wilcoxon signed-rank test: $Z = 1.1$, $p = .13$, $r = .14$, $BF = 1.1$). Thus, participants who generalized more or displayed more directed exploration in one task, also did so in the other. While λ-values were correlated across the two tasks, they were significantly smaller in the conceptual task than in the spatial task (Wilcoxon signed-rank test; $Z = -5.1$, $p < .001$, $r = .64$, $BF > 100$), meaning participants generalized over relatively smaller conceptual distances than spatial distances. This may have contributed to the similar predictive accuracy of the GP and BMT models, since the GP will behave like the BMT in the limit of $\lambda \to 0$, by assuming complete independence between options. We also found that larger lambdas were correlated with higher performance across both tasks ($r_\tau = .44$, $p < .001$, $BF > 100$). Estimates for the exploration factor β did not differ between tasks ($Z = 1.1$, $p = .86$, $r = .14$, $BF = 0.2$).

[†]Kendall's tau (r_τ) is used to report rank-correlation and should not be confused with the temperature parameter of the softmax choice rule τ.

5.2.5 Discussion

We investigated whether the search for spatially or conceptually correlated rewards can be connected via common principles of generalization. Our results showed that participants performed well in both domains, with correlated performance across the two tasks. Using a Gaussian Process (GP) regression framework as a model of generalization and postulating Upper Confidence Bound (UCB) sampling as an optimistic approach to the exploration-exploitation dilemma, we made progress towards understanding the computational mechanisms of generalization and search across spatial and conceptual domains. Our model produced good out-of-sample predictions in both tasks, made predictions of unobserved rewards that correlated with participants' judgments, and produced human-like learning curves based on meaningful parameter estimates that showed levels of generalization and directed exploration that correlated across the two domains.

Although the GP predicted participants better than the BMT in the spatial task, we found no difference between models in the conceptual domain, although the behavioral data indicates successful generalization (see Fig. 5.2f-g). This could be explained by two different—not mutually exclusive—reasons. One explanation could be that by presenting both spatial and conceptual features in each task, it was simply harder to ignore the spatial features when they were irrelevant. We have some evidence of this from Figure 5.2e, where we see that participants in the conceptual task tended to sample locally in spatial distance, whereas we see an asymmetric pattern in the spatial task, where participants tended to resemble the random baseline in terms of conceptual distance, with the exception of repeat clicks. Another explanation could be that conceptual stimuli induce different priors over features than in the spatial domain. The heatmap in Figure 5.2h indicates that participants may have had linear priors for conceptual features (e.g., more berries or more leaves lead to higher rewards), which was not the case in the spatial task (see 5.2h-i). We address both issue in the next experiment, where we change the task to only present either the relevant spatial or conceptual features and replace the leaf and berry stimuli with Gabor patches, which are commonly used in psychophysical and vision research (Polat, Mizobe, Pettet, Kasamatsu, & Norcia, 1998; Rolfs, Lawrence, & Carrasco, 2013) and are less likely to induce directional priors on reward.

5.3 Experiment 6: Gabor search

In order to avoid an overlap between conceptual and spatial features, we constructed a new experiment where participants used the arrow keys to select a single option at a time, where each option had only spatial features or conceptual features, but not both (Fig. 5.4). Options were displayed as a single highlighted tile in the spatial task, or as a unique Gabor patch in the

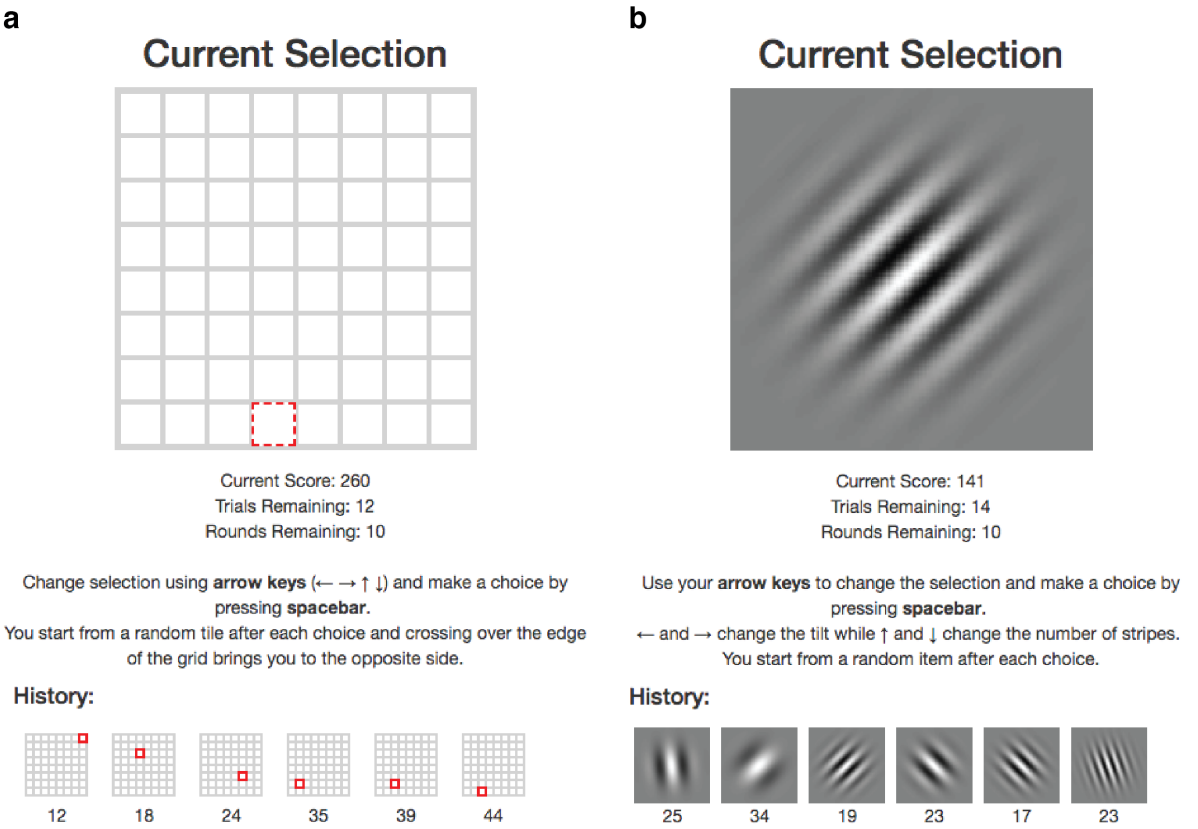

Fig. 5.4 Experiment 6 design. **a**) In the spatial task, options were defined as a highlighted tile on a 8x8 grid, where the arrow keys could be used to move the highlighted square around the grid. **b**) In the conceptual task, each option was represented as a Gabor patch, where the arrow keys changed the tilt and the number of stripes.

conceptual task. Pressing one of the arrow keys would move the highlighted option (spatial) or alter the features of the Gabor patch (conceptual) by changing the tilt or the number of stripes. The tilt and stripes of the Gabor patch stimuli are not as clearly countable as the leaf and berry stimuli, and were chosen to reduce the possibility of directional prior beliefs about reward (e.g., more leaves or berries produce higher rewards). The environment was designed as a torus, such that moving to the extreme of one feature direction would result in starting from the opposite end (*a la* Mrs. Pacman).

Selections were made by pressing the space bar, at which point the reward value for the selected option was displayed for 800 milliseconds, and then subsequently added to the history at the bottom of the page. The history included all previous selections and reward values for the current round. After making a selection, a randomly selected option was displayed as the new initial position, at which point the participant could again use the arrow keys to navigate to the desired selection.

5.3.1 Methods

Participants and design

120 participants were recruited through Amazon Mechanical Turk (requiring 95% approval rate and 100 previously approved HITs, while excluding those who participated in previous experiments) for a two part experiment, where only those who completed part one were invited for part two. In total, 95 participants completed both parts of the experiment and were included in the analyses (31 Female; mean age=36, SD=10). Participants were paid $2.00 for each part of the experiment, with those completing both parts being paid an additional performance-contingent bonus of up to $6.00. Participants earned $8.22 $\pm$ 0.57 and spent 36 $\pm$ 12 minutes completing both parts. There was an average gap of 17 $\pm$ 7.8 hours between the two parts of the experiment.

We varied the task order between subjects, with participants completing the Spatial and Conceptual task in counterbalanced order in separate sessions. We also varied between subjects the extent of reward correlations in the search space by randomly assigning participants to one of two different classes of environments (*Smooth* vs. *Rough*), with smooth environments corresponding to stronger correlations. The same environment class was used for both tasks.

Materials and procedure

Each search task comprised 10 rounds, with a different reward function sampled without replacement from the set of assigned environments (see below). The reward function specified how rewards mapped onto either the spatial or conceptual features. In each round, participants had a limited search horizon of 20 available actions (i.e., choices), which could be used to either explore unrevealed options or to exploit previously selected options. Participants were instructed to accumulate as many points as possible, which were later converted into monetary payoffs.

Participants were told that options with either similar spatial features (spatial task) or similar conceptual features (conceptual task) would yield similar rewards. Three comprehension questions (different for spatial and conceptual tasks) were used to ensure full understanding of the task (specifically whether spatial or conceptual features predicted reward) before participants were allowed to continue. The last round was the "bonus round", where participants made explicit judgments about the expected rewards and their estimated uncertainty of five unrevealed tiles in the middle of the round (i.e., after 15 choices), in order to tap into beliefs about unobserved options supported by generalization. All behavioral and computational modeling analyses exclude the last round, except for the analysis of the bonus round judgments.

As in the previous experiment, each observation included normally distributed noise, $\varepsilon \sim \mathcal{N}(0,1)$, where the rewards for each round were scaled to a uniformly sampled maximum value in the range of 35 to 45, so that the value of the global optima could not be easily guessed. All environments were sampled as a 8 x 8 bivariate function from a GP prior parameterized with a RBF kernel. We again generated 40 samples of each type of environments, but this time using $\lambda_{Smooth} = 4$ and $\lambda_{Rough} = 1$, since the smoothness of λ is defined relative to the size of the environment. Participants performed the task on 10 environments (one per round) that were sampled without replacement from the set of 40 generated environments in each task.

After each round, participants were given feedback about how well they performed as a percentage of the best possible score (i.e., compared to always selecting the global optimum). The performance bonus (up to $3.00 for each task) was also calculated based on this percentage, averaged over all rounds excluding the tutorial round.

Bonus round judgments

In both tasks the last round was a "bonus round", which solicited judgments about the expected reward and estimated uncertainty of five unrevealed options. Participants were informed that the goal of the task remained the same (maximize cumulative rewards), but that after 15 selections, they would be asked to provide judgments about five randomly selected options, which had not yet been explored (sampled uniformly from unexplored options). Judgments about expected rewards were elicited using a slider, which changed the displayed value of the selected option from 0 to 50 (in increments of 1). Judgments about uncertainty were elicited using a slider from 0 to 10 (in increments of 1), with the endpoints labeled 'Not at all' and 'Highly confident'. After providing the five judgments, participants were asked to choose one of the five selected options by clicking with their mouse. The round was subsequently completed the same as all others.

5.3.2 Behavioral results

Overall, participants performed better in the spatial task than the conceptual task ($t(94) = 11.2$, $p < .001$, $d = 1.1$, $BF > 100$; Fig. 5.5a), and better in smooth than in rough environments ($t(93) = 5.7$, $p < .001$, $d = 1.2$, $BF > 100$), with performance correlated across the two tasks ($r = .59$, $t(93) = 7.0$, $p < .001$, $BF > 100$; Fig. 5.5b). In contrast to Experiment 5, we found no evidence of any order effect, with equivalent performance in both spatial ($t(93) = 1.43$, $p = .16$, $d = 0.29$, $BF = 0.5$) and conceptual tasks ($t(93) = 0.39$, $p = .70$, $d = 0.08$, $BF = 0.23$), no matter the order in which they were performed. Figure 5.5c-d show the learning curves for trials and rounds, respectively, where we see a strong positive correlation between trials and

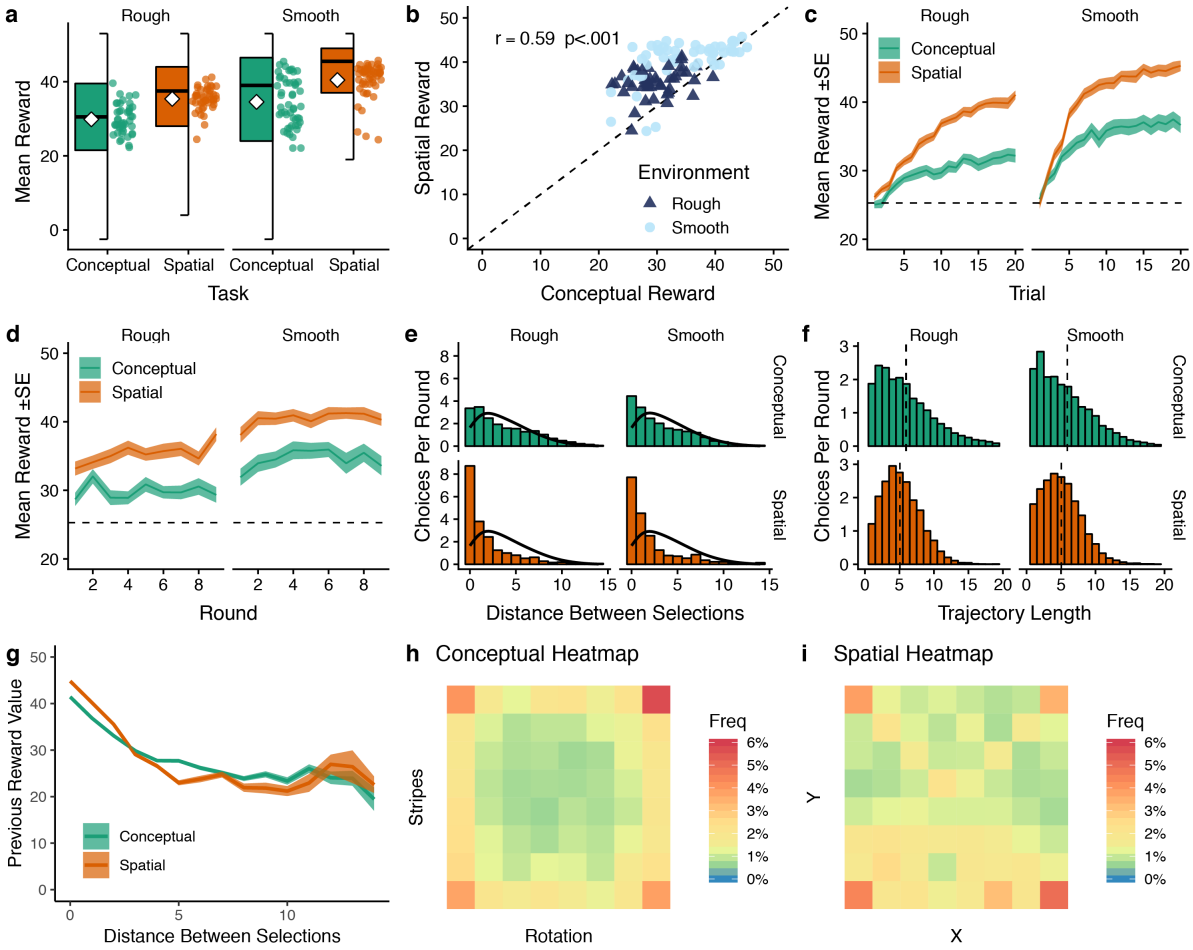

Fig. 5.5 Experiment 6 behavioral results. **a**) Mean participant performance in each task, where each dot is a participant and boxplots indicate the median and 1.5 IQR across all search decisions, with the diamond showing the group mean. **b**) Correlation of performance across the two tasks, where each dot is a participant and the dashed line indicates $y = x$. **c**) Learning over trials and **d**) learning over rounds, where each line is the aggregate mean and the ribbons show the standard error. The dashed line provides a comparison to a rnadom baseline. **e**) The Manhattan distance between selections compared to a random baseline (black line), where participants made more repeat selections in the spatial condition (distance=0) compared to the conceptual condition. **f**) Distribution of trajectory length (i.e., number of steps taken before a selection was made), where the dashed line indicates the median in each condition. Participants had longer spatial trajectories on average, although the distribution for the conceptual task has a longer tail. **g**) The correspondence between previous reward value and the Manhattan distance to the next selected option. Lines indicate aggregate mean, with ribbons indicating the standard error. Lower rewards led to participants moving further in the relevant feature space. **h-i**) Heatmaps of the distribution of choices, aggregated over all participants. Colors are centered on yellow representing random chance (1/64), with orange and red indicating higher than chance, while green and blue were lower than chance.

rewards ($r = .91$, $p < .001$, $BF > 100$), but only anecdotal evidence of a correlation between rounds and rewards ($r = .63$, $p = .07$, $BF = 1.8$).

Patterns of search

Looking at the distance between sequential selections, participants searched more locally in the spatial than the conceptual task ($t(94) = 9.3$, $p < .001$, $d = 1.5$, $BF > 100$), but there was no difference across environments ($t(93) = 1.0$, $p = .32$, $d = 0.2$, $BF = 0.3$). The driving factor behind this increased locality of choices in the spatial task can be seen in Figure 5.5e, where participants made more repeat selections in the spatial task (distance=0; $t(94) = 4.27$, $p < .001$, $d = 0.62$, $BF > 100$) than in the conceptual task. The tendency towards locality was not due to lack of effort, because recall that each trial began with a random initialization of the stimuli. Figure 5.5f shows the average trajectory length (i.e., number of steps) before a selection was made, where overall, participants had an average trajectory length of 5.5 steps. Participants in the conceptual task produced longer trajectories than in the spatial task ($t(94) = 3.4$, $p < .001$, $d = 0.4$, $BF = 24$), but with no difference across environments ($t(93) = 0.1$, $p = .89$, $d = 0.03$, $BF = 0.2$).

Figure 5.5g shows the relationship between the value of a reward and the Manhattan distance to the next selection, measured in the relevant feature space. These response curves indicate that participants responded similarly in both task, with a tendency to jump further in space when encountering low rewards. Figures 5.5h-i show the frequency of selecting the various options. Participants had a preference for sampling the four extremities of the feature space (i.e., corners), although the conceptual task also produced a sampling pattern along the outer perimeter that was not found in the spatial task. However, compared to Experiment 5, we see less of a tendency to sample a single value in the conceptual space (i.e., selecting five leaves and five berries; Fig. 5.1h).

5.3.3 Model Comparison

We again use leave-one-round-out cross validation to estimate model parameters and compare models based on the out-of-sample prediction accuracy, where $R^2 = 0$ is equivalent to a random model and $R^2 = 1$ represents a theoretically perfect model. The GP made small but meaningfully better predictions than the BMT in both the conceptual task ($t(94) = 3.75$, $p < .001$, $d = 0.06$, $BF = 67$; Fig. 5.6a) and in the spatial task ($t(94) = 4.59$, $p < .001$, $d = 0.03$, $BF > 100$). Overall, the GP best predicted 72 out of 95 participants in the conceptual task and 67 out of 95 participants in the spatial task.

Bonus round judgments

Figure 5.6b shows the correspondence between participant and model predictions in the bonus round. Even though the BMT invariable makes the same prediction of 25 for all unobserved

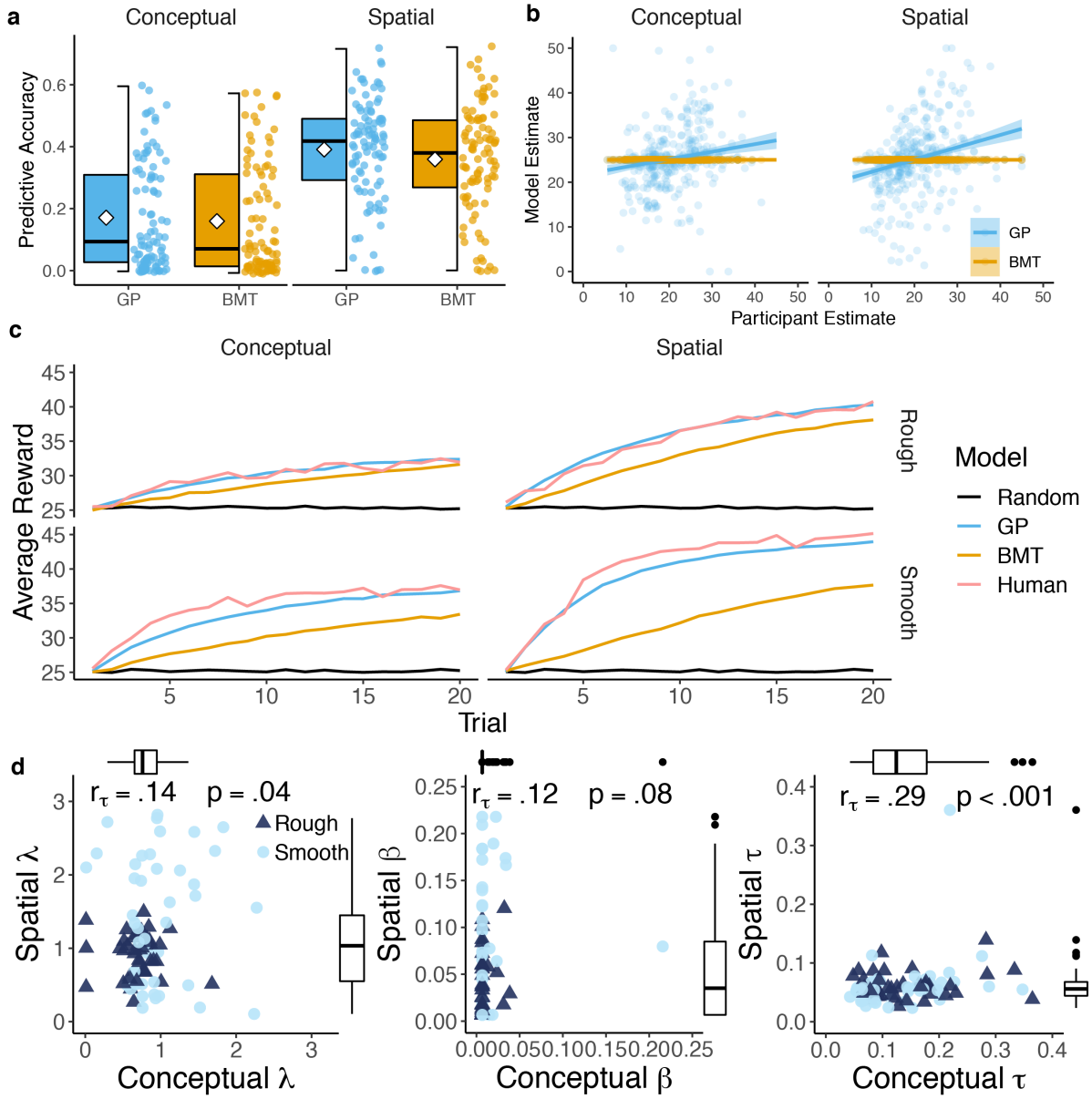

Fig. 5.6 Experiment 6 modeling results. **a**) Predictive accuracy of models on out-of-sample predictions, where 0 corresponds to random chance and 1 is a theoretically perfect model. Each dot is a single participant, with boxplots showing the median and 1.5 IQR, and the diamond indicating group means. **b**) The relationship between participant estimates in the bonus round and model estimates (parameterized using the median participant parameter estimates). Each dot is a single estimate, while lines represent a linear regression (with ribbons showing standard error). **c**) Simulated learning curves (10k replications) using models specified with parameters sampled from participant estimates. We include comparison to a random model (black line) and human performance (pink). **d**) Median GP parameters from the conceptual (x-axis) and spatial tasks (y-axis), where each point is a single participant. Outliers (using the Tukey criteria) are excluded from the plot but not from the rank correlations.

options (based on the prior mean), it still managed to make marginally better predictions of participant judgments than the GP in both the conceptual task (comparing mean absolute error: $t(94) = 2.10, p = .039$, $d = 0.20$, $BF = 0.92$) and in the spatial task ($t(94) = 2.48, p = .015$,

$d = 0.25$, $BF = 2.07$). While not necessarily accurate, GP predictions were correlated with participant judgments in both the conceptual task ($r = .20$, $p < .001$, $BF > 100$) and in the spatial task ($r = .28$, $p < .001$, $BF > 100$), with correlations stronger in the latter. Correlations for the BMT are undefined, since it invariably makes the same predictions for all unobserved options (based on the prior mean of 25).

We again use mixed-effects regression to compare model predictions about uncertainty to confidence ratings, in order to account for individual variability. We find that higher GP uncertainty estimates predicted lower confidence ratings in the spatial task ($\beta = -.16$, $t(466) = -3.8$, $p < .001$, $BF > 100$), and in the conceptual task ($\beta = -.10$, $t(473) = -2.1$, $p = .039$, $BF = 9$), albeit with lower predictive power in the latter. In both cases, the Bayes Factor is computed based on a comparison to a null intercept only model. This is also equivalent to a comparison against the BMT, since it invariantly makes the same uncertainty predictions about unobserved options.

Learning curves

We again simulated learning curves for each model over 10,000 replications by sampling (with replacement) from the set of cross-validated participant parameter estimates. Figure 5.6c shows the aggregated learning curves together with a random baseline (black) and human performance (pink). We see a very strong correspondence between the GP and human learning curves, while the BMT simulations performed visibly worse. The relatively better performance of human participants in the spatial vs. conceptual task and in smooth vs. rough environments is also mirrored by the simulated GP curves.

Parameter estimates

Figure 5.6d shows the correspondence between GP parameters across the conceptual and spatial tasks. We find a marginal correlation between λ estimates across the two tasks ($r_\tau = .14$, $p = .04$, $BF = .99$), with participants tending to have lower λ estimates in the conceptual task ($Z = -2.6$, $r = -.27$, $p = .005$, $BF = 1.6$). This indicates that participants tended to generalize less in the conceptual domain compared to the spatial domain. Environment type also influenced the level of generalization, with participants assigned to the smooth condition generalizing more than those in the rough condition (Mann-Whitney U test: $U = 3451$, $p = .006$, $r_\tau = .16$, $BF = 2.4$).

We find no correlation between β across the two tasks ($r_\tau = .12$, $p = .08$, $BF = .58$), in contrast to Experiment 5. This is potentially due to the substantially lower estimates for β in the conceptual task ($Z = -5.29$, $r = -.54$, $p < .001$, $BF > 100$), where there was close to

negligible levels of directed exploration. Participants had marginally higher levels of directed exploration (β) in smooth environments compared to rough ($U = 3737$, $p = .047$, $r_\tau = .12$, $BF = 0.83$), but more so when only comparing estimates from the spatial task ($U = 726$, $p = .003$, $r_\tau = .25$, $BF = 2.9$)

Looking at the correspondence between levels of undirected, random exploration as captured by the softmax temperature parameter τ, we find correlations across the two tasks ($r_\tau = .29$, $p < .001$, $BF > 100$). Additionally, participants had substantially higher levels of τ in the conceptual task than in the spatial task ($Z = 7.12$, $r = .73$, $p < .001$, $BF > 100$). Thus, while directed exploration was reduced in the conceptual task, random exploration increased.

5.3.4 Discussion

The results of Experiment 6 show that by only showing the relevant spatial of conceptual features of the task (thereby removing potential interference), we eliminated the order effect found in Experiment 5. The use of Gabor patches as conceptual stimuli also reduced preferential sampling towards a single extreme region of the conceptual space. By accounting for these potential confounds, we find that the GP function learning model is able to best predict behavior compared to a non-generalizing BMT model in both domains and produce simulated learning curves that closely match the differences across tasks and environments of human performance. This provides evidence of a common search process across domains, although the comparison to Experiment 5 informs us that the choice of stimuli and the design of the task play an important role in influencing behavior.

While there are similarities, we also find differences in how people search in spatial and conceptual domains. Participants performed systematically better in the spatial task than the conceptual task, indicating a difference in difficulty. One possibility is that the Gabor patches induced more perceptual error than the grid locations. Participant may have been more likely to confuse a a selected Gabor patch for an intended target, compared to the likelihood of confusing two separate grid locations. We see some evidence for this possibility in the higher proportion of repeat samples in the spatial domain, although this may also be due to differences in generalization and sampling strategies.

While the GP was the best predictive model in both domains, the overall accuracy was lower in the conceptual task for both models. Indeed, the parameter estimates of the GP show that there was a shift in exploration strategy, leading to less directed exploration and more random exploration in the conceptual task. Based on the lower performance and lower estimates of λ in the conceptual task, participants may have put less confidence in the effectiveness of directed exploration as a function of their ability to make accurate generalizations. This could explain

why they favored random instead of directed exploration. When your model of the world is less accurate, a more simple and random strategy may become adaptive.

A different (and non-mutually exclusive) explanation could be that because participants were only shown a single target stimuli in the conceptual task (i.e., a single variant of the Gabor patch that could be altered using the arrow keys), it may have been more difficult to make predictive generalizations about other unseen stimuli. In contrast, all potential grid locations were visible at all times in the spatial domain. Thus, the entire map of possibilities were available at all times in the spatial task, whereas participants could only see a single point in the conceptual search space. While the modeling results and bonus round judgments indicate that participants were capable of generalizing in both domains, the presentation of the task may have played a role in how people organized a cognitive map of the search space. Indeed, theories about a generalized search processes are typically unidirectional, with the argument that a spatial organization of knowledge is more primary and has been adapted through evolution for representing more complex conceptual information (Hills et al., 2008; Todd et al., 2012)

5.4 General Discussion and Conclusion

Humans search for rewards across a multitude of different domains, using effective generalization and clever exploration to great success. Historic psychological findings explained adaptive generalization in spatial domains by evoking the concept of a cognitive map, whereas more recent neuroscientific evidence suggests cognitive maps can be found in both spatial and conceptual domains. We investigated whether the search for spatially or conceptually correlated rewards can be connected via common principles of generalization.

Our results showed correlated performance in both domains, with evidence for common computational principles of generalization and search. However, the choice of stimuli and the design of the task played a consequential role in influencing behavior. In Experiment 5, where both spatial and conceptual features were simultaneously presented, we found that irrelevant spatial features were more difficult to ignore, leading to a task order effect and less effective generalization in the conceptual domain. In Experiment 6, we adapted the design to show only the relevant spatial or conceptual features, where we found that a Gaussian Process (GP) model of generalization combined with Upper Confidence Bound (UCB) sampling was the best model of behavior in both tasks. Our model produced the best out-of-sample predictions in both tasks, made predictions of unobserved rewards that correlated with participants' judgments, and produced human-like learning curves based on participant parameter estimates.

Nevertheless, we also find notable differences between how people search in spatial and conceptual domains. The parameter estimates from Experiment 6 indicate that participants

swapped directed exploration in favor of more random exploration in the conceptual domain. Consistent with the lower level of correspondence between participant and GP predictions about the expected reward and the underlying uncertainty of unobserved options, this suggests that generalization was more difficult in the conceptual domain. We present two potential explanations in the Discussion section of Experiment 6 related to the design of the experiment (perceptual error and seeing only a single conceptual stimuli at a time), but there may also be more general domain differences. Spatial relationships may in principle be easier to learn due to the extended nature of our awareness of the world around us, whereas conceptual stimuli are more commonly experienced one at a time (e.g., a single idea in a sequential train of thought). Currently our model does not account for attentional mechanisms (Radulescu, Niv, & Ballard, 2019) or working memory (Collins & Frank, 2018; Ohl & Rolfs, 2018) constraints on how people integrate information, although this may be a crucial source of domain differences. Indeed, the "stretchy birds" paradigm used by Constantinescu et al. (2016) as evidence for a common neural representation of spatial and conceptual knowledge required several hours of training before being measured in the scanner. Thus, while both spatial and conceptual knowledge are capable of being organized into a common map-like representation, there may be domain differences in terms of the ease of learning such a map, owing to different demands on cognitive resources.

In summary, we have presented a detailed set of experimental data studying how people generalize and search in both spatial and conceptual domains. Our behavioral and computational modeling results enrich our understanding of the similarities and differences in how people represent and perform search across domains. While people employ similar principles of generalization and search, we also find domain differences that open up an intriguing avenue for future research.

Chapter 6

Generalization in Structured Spaces

Life is not like water. Things in life don't
necessarily flow over the shortest possible route.
—Haruki Murakami

This chapter extends theories of spatial and conceptual generalization into the domain of structured representations, where transitions rather than singular features define relationships in the environment. We use a diffusion kernel to model structural similarity based on transition properties of discrete graphs, and show that the RBF kernel emerges as a special case when the environment is perfectly symmetrical and transitions are unrestricted.

In two experiments, we demonstrate that people are able to use structural similarity to make inferences about unobserved outcomes and to guide the search for graph-correlated rewards. Our results provide evidence for the theory that generalization is supported by a predictive map of the world that is sensitive to the transition dynamics that define structured environments.

This chapter is based on the following publication in addition to new unpublished research:

Wu, C. M., Schulz, E., & Gershman, S. J. (2019). Generalization as diffusion: human function learning on graphs. In A.K. Goel and C.M. Seifert & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society*. (pp. 3122–3128). Montreal, QB: Cognitive Science Society.

6.1 Introduction

From encountering a new move in a game of chess, to finding your way through a foreign airport, representing and reasoning about structured environments is a pervasive feature of intelligence. Any adaptive agent must be able to respond and act appropriately in novel situations (Lake et al., 2017), where leveraging the structure of the environment is an essential ingredient for intelligent behavior (Gigerenzer & Todd, 2012; Simon, 1990). Specifically, we focus on the problem of how people generalize over structured spaces, by making predictive inferences to guide efficient learning. For instance, by predicting which sequence of moves will lead to a checkmate or which route through the airport will lead you to the gate the fastest.

Generalization as a means for adapting to new situations has featured prominently in research on classification (Erickson & Kruschke, 2002; Nosofsky, 1988b) and reinforcement learning (Wu, Schulz, Garvert, et al., 2018; Wu, Schulz, Speekenbrink, et al., 2018), where a common assumption is that similarity is a guiding principle (Tenenbaum & Griffiths, 2001). Shepard (1987) famously showed that once trained to react to a stimulus, an animal's probability of showing the same response to new stimuli decays exponentially as the similarity between stimuli decreases.

Here, we examine whether principles of similarity-based generalization can be used to perform inference and guide exploration in structured environments, where transitions rather than stimuli features define the distribution of rewards. We use a function learning framework to describe generalization, where previous observations can be interpolated or extrapolated to make predictions about unobserved stimuli. We show that traditional models of function learning in continuous domains can be subsumed under a model of structure-based generalization, offering a unifying framework for understanding adaptive behavior across spatial, conceptual, and structured domains.

6.1.1 Generalization as function learning

How much hot sauce should you add to a meal to enhance the flavor? How hard should you push a child on a swing? There is a wealth of literature devoted to studying how humans learn functions in the world, which has traditionally focused on how people learn an explicit relationship between two continuous variables (Brehmer, 1974; Busemeyer et al., 1997; Carroll, 1963; Koh & Meyer, 1991).

While the majority of function learning research has studied continuous spaces, many real-world problems are better represented using discrete graph structures, where transition dynamics rather than Cartesian distance between features define the relationship between inputs and outputs. For example, the domain of Reinforcement Learning (RL; Sutton & Barto,

2018) assumes that the problem of learning through interactions with the environment can be described as a Markov Decision Process (MDP). MDPs are graph structures representing each state of the world as a node and transitions between states as edges. Function learning in RL often means learning a value function over possible states, in order to train a policy that will guide the agent to rewarding states. In vast problems where not all states can be observed, modern implementations of RL often use neural network to learn an approximation of the value function (Mnih et al., 2015; Silver et al., 2016).

An early and influential approach to generalization in RL was the *Successor Representation* (SR; Dayan, 1993). The SR describes a similarity metric based on the state dynamics (i.e., the transition structure between states), which allows the value function to generalize observations of rewards to unobserved states. This can also be understood as a similarity-based approach for function learning on discrete spaces, using similarity between future expected state transitions as the basis for generalization. A simple example is that if one's goal is to get on an airplane, one should not drive directly to where the plane is parked on the tarmac and try to board. Rather, it would be better to head towards into terminal, because from there the chain of state transitions (e.g., check-in, security screening, then boarding) is more likely to lead to a pleasant flight. Thus, function learning in discrete spaces can also be guided by similarity, but one based on transition structure as opposed to distance.

6.1.2 Goals and scope

We describe a method for learning functions and performing Bayesian inference over graph structures using a diffusion kernel. The diffusion kernel provides a similarity metric of a graph based on its transition structure, and when combined with the Gaussian Process (GP) framework, allows us to make Bayesian predictions about unobserved nodes. We show that a prominent model of function learning in continuous spaces is a special case of this model, and that both have theoretical equivalencies to the Successor Representation. To put our theory into practice, we present two experiments studying how people make inferences and search for rewards on graph structures. Experiment 7 is a function learning task where participants are shown a series of generated subway maps and asked to predict the number of passengers at unobserved stations. Our GP model using a diffusion kernel is able to capture both participant prediction and confidence judgments. Experiment 8 is a multi-armed bandit tasks, where nodes on a graph yield normally distributed rewards with a mean that is correlated across the graph structure (i.e., similarly connected nodes have similar rewards). We show that the GP model best predicts choices, produce human-like learning curves, and also describes judgments about unobserved nodes. Overall, these results extend the scope of all previous chapters in this

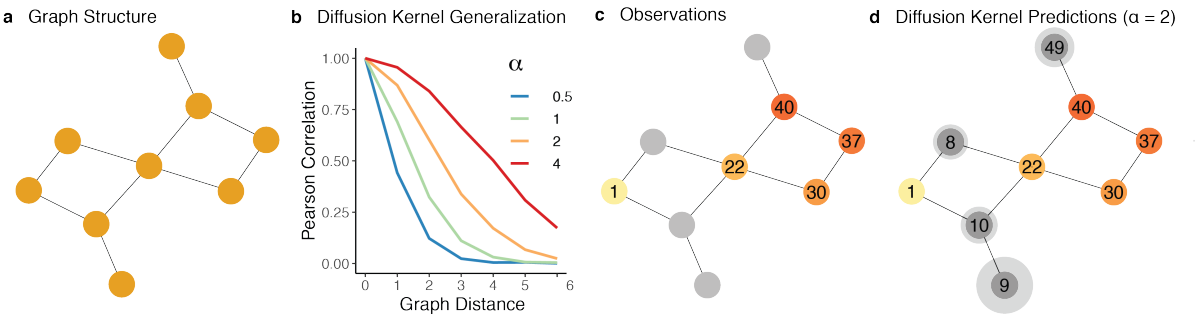

Fig. 6.1 Inference over graphs. **a**) An example of a graph structure, where nodes represent states and edges indicate the transition structure. **b**) A diffusion kernel is a similarity metric between nodes on a graph, allowing us to generalize to unobserved nodes based on the assumption that correlations of rewards decays as an exponential function of the distance between two nodes. The diffusion parameter (α) governs the rate of decay. **c**) Given some observations on the graph (colored nodes), we can use the diffusion kernel combined with the Gaussian Process framework to make predictions (**d**) about expected rewards (numbers in grey nodes) and the underlying uncertainty (size of halo) for each of the unobserved nodes.

thesis to structured spaces, opening a rich connection to the neurological basis for how people organize structured knowledge and perform generalization.

6.2 Generalization on graph structures

We can specify a graph $G = (\mathcal{S}, \mathcal{E})$ with nodes $s_i \in \mathcal{S}$ and edges $e_i \in \mathcal{E}$ to represent a structured state space (Fig. 6.1a). Nodes represent states and edges represent allowed transitions. For now, we assume that all edges are undirected (i.e., if $x \to y$ then $y \to x$). The connectivity structure of the graph determines which states are accessible from a given prior state, and is often described using the graph Laplacian L:

$$L = D - A \tag{6.1}$$

where A is the adjacency matrix and D is the degree matrix. Each element $a_{ij} \in A$ is 1 when nodes i and j are connected, and 0 otherwise, while the diagonals of D describe the number of connections of each node[†].

6.2.1 The diffusion kernel

The diffusion kernel (DF; Kondor & Lafferty, 2002) defines a similarity metric $k(s, s')$ between any two nodes based on the matrix exponentiation of the graph Laplacian:

$$k(s, s') = \exp^{\alpha L} \tag{6.2}$$

[†]The graph Laplacian can also describe graphs with weighted edges, where we substitute the weighted adjacency matrix W for A and the degree matrix describes the weighted degree of each node. All analyses apply in the weighted case.

Intuitively, the diffusion kernel assumes that rewards diffuse along the graph similar to a heat diffusion process (i.e., by assuming a continuous random walk). Thus, closely connected nodes are assumed to be more similar and to have similar rewards. The free parameter α models the level of diffusion, where $\alpha \to 0$ assumes complete independence between nodes, while $\alpha \to \infty$ assumes all nodes are perfectly correlated. In practice, the matrix exponentiation described in Eq. 6.2 can be accomplished by first decomposing L into its eigenvectors $u_i \in U$ and eigenvalues $\lambda_i \in \Lambda$, and then substituting matrix exponentiation with real exponentiation using $k(s,s') = \exp^{\alpha L} = \sum_i u_i \exp^{\alpha\lambda_i} u_i^\top$.

Gaussian Process regression

From the similarity metric created by the diffusion kernel, we can use Gaussian Process (GP) regression (Rasmussen & Williams, 2006) to perform Bayesian inference over observations on structured graphs. A GP defines a distribution over functions $f : \mathcal{S} \to \mathbb{R}^n$ that map the state space $\mathcal{S}$ to real-valued scalar outputs (e.g., rewards). Functions are modeled as a random draw from a multivariate normal distribution:

$$f \sim \mathcal{GP}\left(m(s), k(s,s')\right), \tag{6.3}$$

where $m(s)$ is a mean function specifying the expected output of s, and $k(s,s')$ encodes prior assumptions about the underlying function. We use the diffusion kernel (Eq. 6.2) to represent the covariance based on the connectivity structure of the graph, and follow the convention of setting the mean function to zero, such that the GP prior is fully defined by the kernel.

Given some observations $\mathcal{D}_t = \{\mathbf{s}_t, \mathbf{y}_t\}$ of observed rewards $\mathbf{y}_t$ at states $\mathbf{s}_t$, we can compute the posterior distribution $p(f(s_*)|\mathcal{D}_t)$ for any target state s_*. The posterior is a normal distribution with mean and variance defined as:

$$m(s_*|\mathcal{D}_t) = \mathbf{k}_{*,t}^\top (\mathbf{K} + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{y}_t \tag{6.4}$$

$$v(s_*|\mathcal{D}_t) = K(s_*, s_*) - \mathbf{k}_{*,t}^\top (\mathbf{K}_t + \sigma_\varepsilon^2 \mathbf{I})^{-1} \mathbf{k}_{*,t} \tag{6.5}$$

where $\mathbf{K}_t$ is the $t \times t$ covariance matrix evaluated at each pair of observed inputs, and $\mathbf{k}_{*,t} = [k(s_1, s_*), \ldots, k(s_t, s_*)]$ is the covariance between each observed input and the target input s_*, and σ_ε^2 is the noise variance. Thus, for any node in the graph, we can make Bayesian predictions about the expected reward $m(s_*|\mathcal{D}_t)$ and also the level of uncertainty $v(s_*|\mathcal{D}_t)$ (Fig. 6.1e).

The posterior mean function of a GP can be rewritten as:

$$m(s) = \sum_{i=1}^{t} w_i k(s_i, s) \tag{6.6}$$

where each s_i is a previously observed state and the weights are collected in the vector $\mathbf{w} = \left[k(\mathbf{s}_t, \mathbf{s}_t) + \sigma_\varepsilon^2 \mathbf{I}\right]^{-1} \mathbf{y}_t$. Intuitively, this means that GP regression is equivalent to a linearly-weighted sum using basis functions $k(s_i, s)$ to project observed states onto a feature space (E. Schulz, Speekenbrink, & Krause, 2018). To generate new predictions for an unobserved state s, each output y_t is weighted by the similarity between observed states s_t and the target state s.

6.2.2 Connecting spatial and structured generalization

The Gaussian Process framework allows us to relate similarity-based generalization on graphs to theories of generalization in continuous domains. Consider the case of an infinitely fine lattice graph (i.e., a grid-like graph with equal connections for every node and with the number of nodes and connections approaching continuity). Following Kondor and Lafferty (2002) and using the diffusion kernel defined by Eq. 6.2, this limit can be expressed as

$$k(s, s') = \frac{1}{\sqrt{(4\pi\alpha)}} \exp\left(\frac{-|s - s'|^2}{4\alpha}\right), \tag{6.7}$$

which is equivalent to the Radial Basis Function (RBF) kernel that has been used throughout this thesis. Thus, the RBF can be understood as a special case of the diffusion kernel, offering a rich set of connections to theories of generalization over spatial and conceptual features (E. Schulz et al., in press; Wu, Schulz, Garvert, et al., 2018; Wu, Schulz, et al., 2017; Wu, Schulz, Speekenbrink, et al., 2018).

This equivalency between the RBF and diffusion kernel holds when knowledge is represented as a continuous Cartesian coordinate system, where all neighboring points are symmetrically reachable and movement through this representational space is unrestricted. In this case, relationships between stimuli are defined purely in terms of geometry, such that any distance or feature-based assessment of similarity is equivalent to the transition-based similarity metric of the diffusion kernel. However, once we introduce restrictions on transitions (such as a wall preventing movement or sequential dynamics where past actions restrict the set of future possible actions), this equivalency fails. Thus, one could consider spatial representations of similarity to be subsumed under a broader representation of similarity based on the transition structure of the environment.

While we have presented evidence for a spatial organization of knowledge (Constantinescu et al., 2016; Hills, 2006; Theves et al., 2019; Todd et al., 2012; Wu, Schulz, Garvert, et al., 2018; Wu, Schulz, Speekenbrink, et al., 2018), there is more to the story. The neural encoding of spatial information in the hippocampus is sensitive to the transition structures of the environment (Stachenfeld et al., 2014, 2017), such that representations of spatial location become skewed along commonly travelled directions. This suggests that the organization of knowledge in the brain produces a predictive rather than a static map of the world. Specifically, our cognitive map (Tolman, 1948) is organized for predicting future outcomes, such that our representation of similarity captures the transition structures of the environment.

6.2.3 The Successor Representation

The Successor Representation (SR) originated as a method for improving the generalization of Temporal Difference learning (TD learning; Dayan, 1993), but recent work has discovered striking similarities to the neural basis for how humans encode state transitions (Momennejad & Howard, 2018; Momennejad et al., 2017; Stachenfeld et al., 2017).

Dayan (1993) showed that the value function of a TD-learning agent (Eq. 1.5) can be decomposed into a linear combination of state transitions $M(s,s')$ and learned reward representation $R(s')$:

$$V(s) = \sum_{s'} M(s,s')R(s'). \tag{6.8}$$

This matrix of state representations $M(s,s')$ is the SR, where each element m_{jk} encodes the similarity of successor states for states s_j and s_k (Dayan, 1993; Gershman, 2018b). Intuitively, the SR can be understood as a similarity measure based on expectations of future state transitions, rather than the singular features of each state.

When the transition structure of the task is known *a priori*, the SR can be computed in closed form:

$$M(s,s') = (I - \gamma T)^{-1}, \tag{6.9}$$

where I is the identity matrix, γ is the TD discount factor, and T is the transition matrix where $t_{jk} = P(s' = k|s = j)$. A common approach is to assume a random walk over the state space, and substitute T with a modified form of the graph Laplacian:

$$T = I - D^{-1}L \tag{6.10}$$

As suggested by Eq. 6.2 and Eq. 6.10, both the diffusion kernel and the SR provide a similarity metric for generalization based on the transition structure captured by the graph Laplacian. Indeed, in the limit of latent function space, both methods are exactly equivalent

(originally shown in Stachenfeld et al., 2014, but also see Stachenfeld et al., 2017). Given that the SR and the GP parameterized by a diffusion kernel both predict the expected value of a state using a similarity-weighted sum of the rewards of other states (Eqs. 6.6 and 6.8), it can be shown that both similarity measures are the same by an equivalency of their eigenvectors (see Machado et al., 2018).

If $M(s,s') = (1-\gamma T)^{-1}$ and given a normalized graph Laplacian $\tilde{L} = D^{-1/2}LD^{-1/2}$, the i-th eigenvalue λ_i of the SR and the j-th eigenvalue λ_j of the normalized Laplacian can be equated as

$$\lambda_j = \left(1-(1-\lambda_i^{-1})\gamma^{-1}\right), \tag{6.11}$$

and the i-th eigenvector u_i of the SR and the j-th eigenvector u_j of the normalized Laplacian are related by

$$u_j = (\gamma^{-1}D^{1/2})u_i \tag{6.12}$$

Therefore, since both the SR and the GP make predictions as a function of their eigendecompositions, these two models are equivalent in the limit of the latent function space (Stachenfeld et al., 2014). Thus, popular models of human function learning, theories of generalization in reinforcement learning, and historic accounts of psychological laws of generalization laws can all be linked via Gaussian Process inference over graph structures.

This chapter proceeds by presenting two experiments. Experiment 7 is a function learning task, where participants are asked to make predictions about the number of passengers on generated subway maps. We use a GP with the diffusion kernel to describe participant predictions, which we find to be better than several nearest-neighbor averaging rules. Additionally, the uncertainty estimates of the GP also predict participant confidence judgments. Experiment 8 uses a graph-structured bandit task, where participants search for rewards by clicking the nodes of a graph.

6.3 Experiment 7 - Subway prediction task

In this task, we assess how well a GP using the diffusion kernel corresponds to human intuitions about structured spaces. Participants were shown various graph structures (Fig. 6.2) and asked to make a prediction about unobserved nodes by generalizing from the values of other observations. Participants also gave confidence judgments for each judgment, which we also predict using the uncertainty estimates of the GP.

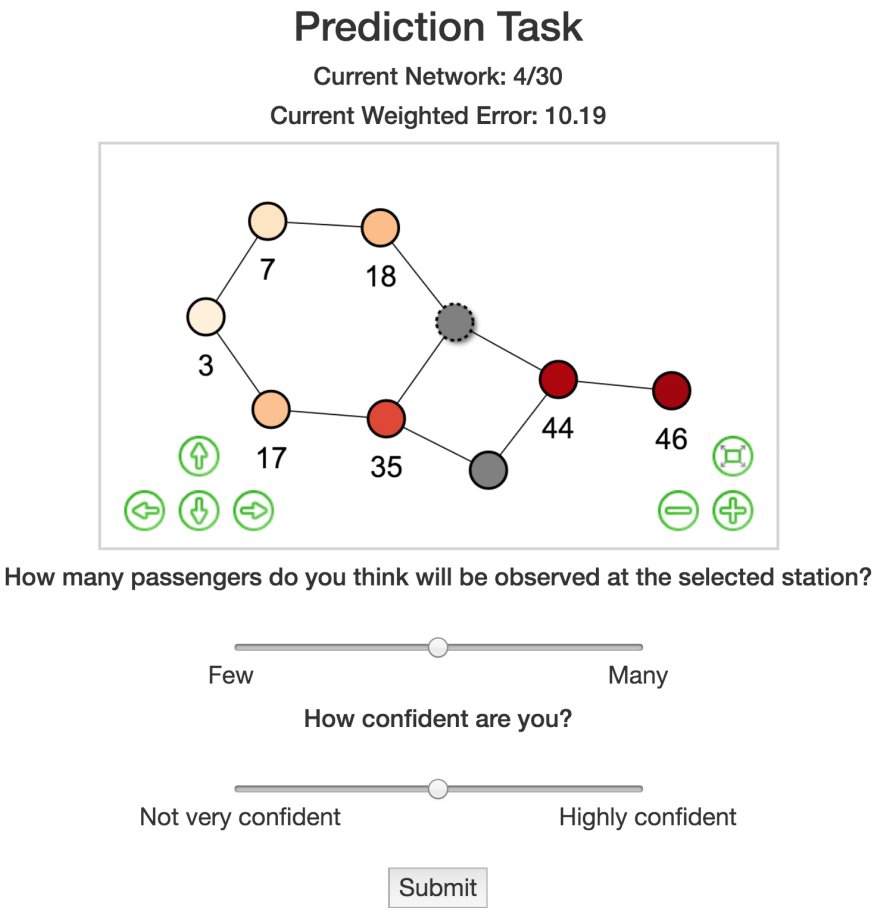

Fig. 6.2 Screenshot from Experiment 7. Observed nodes (3, 5, or 7 depending on the information condition) are shown with a numerical value and a corresponding color aid (darker indicates larger values. The target node is indicated by the dashed line, and would dynamically change color and display a numerical value when participants moved the top slider. Confidence judgments were used to compute a weighted error (i.e., more confident answers having a larger contribution), which was used to determine the performance contingent bonus.

6.3.1 Methods and procedure

We recruited 100 participants ($M_{age} = 32.7$; $SD = 8.4$; 28 female) on Amazon MTurk (requiring 95% approval rate and 100 previously completed HITs) to perform 30 rounds of a graph prediction task. On each graph, numerical information was provided about the number of passengers at 3, 5, or 7 other stations (along with a color aid), from which participants were asked to predict the number of passengers at a target station and provide a confidence judgment (Likert scale from 1 - 11). The subway passenger cover story was used to provide intuitions about graph correlated functions. Additionally, participants observed 10 fully revealed graphs to familiarize themselves with the task and completed a comprehension check before starting the task. Participants were paid a base fee of $2.00 USD for participation with an additional performance contingent bonus of up to $3.00 USD. The bonus payment was

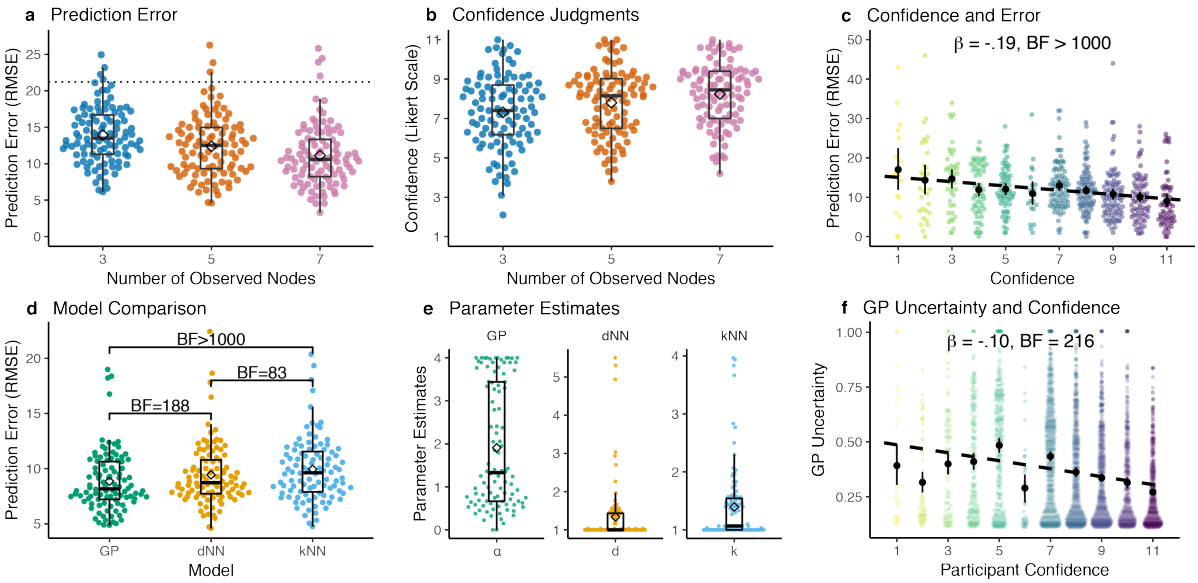

Fig. 6.3 Experiment 7 results. **a-b**) Participant judgment errors and confidence estimates. Each dot is a single participant (averaged over each number of observed nodes), with Tukey boxplots and diamonds indicating group means. The dotted line in **a**) is a random baseline. **c**) Judgment error and confidence. Each colored dot is a participant (averaged over each confidence level), dashed line is a linear regression, with black dots and error bars indicating group means and 95% CI. We report the mixed-effects regression coefficient and Bayes Factor above. **d**) Cross-validated model comparison between the Gaussian Process with diffusion kernel (GP), d-nearest neighbors (dNN), and k-nearest neighbors (kNN). Each point is a single participant with a Tukey boxplot overlaid and diamonds indicating group means. Comparisons are for a Bayesian one-sample t-test. **e**) Parameter estimates, where each dot is the mean cross-validated estimate for each participant, with Tukey boxplots and diamonds indicating group means. **f**) GP uncertainty estimates and participant confidence judgments. Dotted line is a linear regression, with black dots and error bars indicating mean and 95% CI.

based on the mean absolute judgement error weighted by confidence judgments: $R_{bonus} = \$3.00 \times (25 - \sum_i \tilde{c}_i \varepsilon_i)/25$ where $\tilde{c}_i$ is the normalized confidence judgment $\tilde{c}_i = \frac{c_i}{\sum c_j}$ and ε_i is the absolute error for judgment i. On average, participants completed the task in 8.09 minutes ($SD = 3.7$) and earned \$3.87 USD ($SD = \$0.33$).

All participants observed the same set of 40 graphs that were sampled without replacement for the 10 fully revealed examples in the familiarization phase and for the 30 graphs in the prediction task. We generated the set of 40 graphs by iteratively building 3×3 lattice graphs (also known as mesh or grid graphs), and then randomly pruning 2 out of the 12 edges. In order to generate the functions (i.e., number of passengers), we fit a diffusion kernel to the graph and then sampled a single function from a GP prior, where the diffusion parameter was set to $\alpha = 2$ (see Fig. 6.2).

6.3.2 Results

Figure 6.2 shows the behavioral and model-based results of the experiment. We applied linear mixed-effects regression to estimate the effect of the number of observed nodes on participant prediction errors, with participants as a random effect. Participants made systematically lower error predictions as the number of observations increased ($\beta = -.11, t(2899) = -6.3, p < .001$, $BF > 100$[†]; Fig. 6.3a). Repeating the same analysis but using participant confidence judgments as the dependent variable, we found that confidence increased with the number of available observations ($\beta = .16, t(2899) = 11.3, p < .001, BF > 100$; Fig. 6.3b). Finally, participants were also able to calibrate confidence judgments to the accuracy of their predictions, with higher confidence predictions having lower error ($\beta = -.19, t(1637) = -9.0, p < .001, BF > 100$; Fig. 6.3c). There were no substantial effects of learning over rounds ($\beta = .01, t(2899) = 0.5$, $p = .614, BF = 0.4$), suggesting the familiarization phase and cover story were sufficient for providing intuitions about graph correlated structures.

6.3.3 Model comparison

We compare the predictive performance of the GP with two heuristic models that use a nearest-neighbors averaging rule (see below). We fit our models to individual participants using leave-one-graph-out cross-validation, and then make out-of-sample predictions for the left-out judgment. We repeat this procedure for all trials and compare predictive performance using Root Mean Squared Error (RMSE) over all left-out trials.

Heuristic models

We compare the GP model to two heuristic strategies for function learning on graphs, which make predictions about the rewards of a target state s_* based on a simple nearest neighbors averaging rule. The *k-Nearest Neighbors* (kNN) strategy averages the values of the k nearest nodes (including all nodes with same shortest path distance as the k-th nearest), while the *d-Nearest Neighbors* (dNN) strategy averages the values of all nodes within path distance d. Both kNN and dNN default to a prediction of 25 when the set of neighbors are empty (i.e., the median value in the experiment).

Both the dNN and kNN heuristics approximate the local structure of a correlated graph structure with the intuition that nearby states have similar function values. While they sometimes make the same predictions as the GP model and have lower computational demands, they

[†] β is the standardized effect size and we approximate the Bayes Factor using bridge sampling (Gronau, Singmann, & Wagenmakers, 2017) to compare our model to an alternative intercept only null model. See Appendix A for further details.

fail to capture the connectivity structure of the graph and are unable to learn directional trends. Additionally, they only make point-estimate predictions, and thus do not capture the underlying uncertainty of a prediction (which we use to model confidence judgments).

Model results

Figure 6.3d shows that the GP made better predictions than both the dNN ($t(99) = -4.06$, $p < .001$, $d = 0.41$, $BF > 100$) and kNN models ($t(99) = -7.19$, $p < .001$, $d = 0.72$, $BF > 100$). Overall, 58 out of 100 participants were best predicted by the GP, 31 by the dNN, and 11 by the kNN. Figure 6.3e shows individual parameter estimates of each model. The estimated diffusion parameter α was not substantially different from the ground truth of $\alpha = 2$ ($t(99) = -0.66$, $p = .51$, $d = 0.07$, $BF = 0.14$), although the distribution appeared to be bimodal, with participants often underestimating or overestimating the correlational structure. Estimates for d and k were highly clustered around the lower limit of 1, suggesting that averaging over larger portions of the graph were not consistent with participant predictions.

Lastly, an advantage of the GP is that it produces Bayesian uncertainty estimates for each prediction. While the dNN and kNN models make no predictions about confidence, the GP uncertainty estimates correspond to participant confidence judgments ($\beta = -.10$, $t(2043) = -3.4$, $p < .001$, $BF > 100$; linear mixed-effects model with participant as a random effect).

6.3.4 Discussion

How do people generalize on structured spaces? We show how a GP with a diffusion kernel can be used as a model of generalization that produces Bayesian predictions about unobserved nodes. Our model integrates existing theories of human function learning in continuous spaces, where the RBF kernel (commonly used in continuous domains) can be seen as a special limiting case of the diffusion kernel. Using a virtual subway task, we show that the GP was able to capture how people make judgments about unobserved nodes and is also able to generate uncertainty estimates that correspond to participant confidence ratings.

Next, we assess the suitability of the diffusion kernel as a model for more complex problems, by transitioning into a choice paradigm using a multi-armed bandit task with structured rewards E. Schulz, Franklin, and Gershman (2018). One advantage of the GP diffusion kernel model is that it makes prediction with estimates of the underlying uncertainty. In comparison, the SR model only makes point-estimates about the value of a state. Thus, the GP framework offers opportunities for uncertainty-guided exploration strategies (e.g., Auer, 2002).

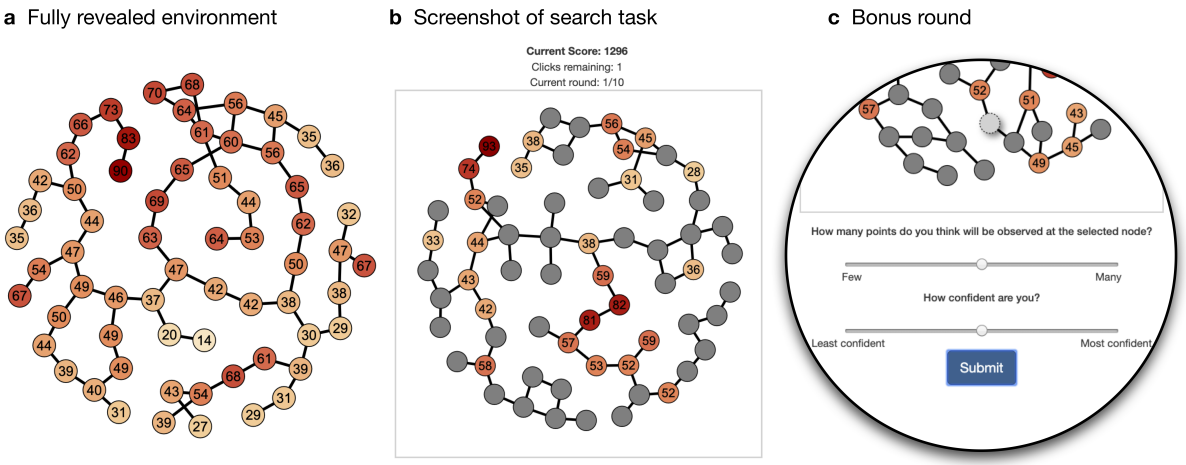

Fig. 6.4 Experiment 8 screenshots. **a**) Four fully revealed environments were shown to participants prior to beginning the task. **b**) During the task participants were instructed to click nodes to earn as much reward as possible. Clicked nodes displayed the numeric value of the earned reward and a color guide (darker colors indicate higher rewards). **c**) Zoomed in screenshot of the bonus round, which activated after the 20th trial on the last round. 10 unclicked nodes were uniformly sampled and participants were sequentially asked to make judgments about expected rewards and their confidence rating. The expected reward slider was mapped to the selected node such that the color and numerical value dynamically changed as the slider was moved.

6.4 Experiment 8: Graph bandit

As an extension of previous work on spatially (E. Schulz, Wu, Huys, Krause, & Speekenbrink, 2018; E. Schulz et al., in press; Wu, Schulz, et al., 2017; Wu, Schulz, Speekenbrink, et al., 2018) and conceptually correlated bandits (Wu, Schulz, Garvert, et al., 2018), we constructed a task where rewards are defined by the connectivity structure of a graph. Participants search for rewards by clicking the various nodes of a graph, where the connections between nodes influence rewards. This provides a reward structure allowing for similarity-based generalization to aid in search, but where similarity is defined based on connectivity rather than perceptual features.

6.4.1 Methods

Participants and design

We recruited 100 participants on Amazon MTurk (requiring 95% approval rate and 100 previously completed HITs). Two participants were excluded because of missing data, making the total sample size $N = 98$ ($M_{age} = 34.3$; $SD = 8.7$; 32 female). Participants were paid \$2.00 for completing the task and earned an additional performance contingent bonus of up to \$3.00. Overall, the task took 7.2 ± 3.3 minutes and participants earned \$4.32 ± \$0.24 USD.

Materials and procedure

Participants were instructed to earn as many points as possible by clicking on the nodes of a graph. Each node represented a reward generating arm of the bandit, where connected nodes yielded similar rewards, such that across the whole graph the expected rewards were defined by a graph-correlated structure (see Fig. 6.4a). Along with the instructions indicating the correlated structure of rewards, participants were shown four fully revealed graphs to familiarize them with the reward structure and had to correctly answer three comprehension questions before starting the task.

After completing the comprehension questions, participants performed a search task over a 10 rounds, each corresponding to a different randomly generated graph structures. In each task, participants were initially shown a single randomly revealed node, and had 25 clicks to either explore unrevealed nodes or to reclick previously observed nodes, where each observation included normally distributed noise $\varepsilon \sim \mathcal{N}(0,1)$. Each clicked node displayed the numerical value (most recent observation if selected multiple times) and a color aid, where darker colors corresponded to larger rewards (Fig. 6.4b). After finishing each round, participants were informed about their performance as a percentage of the best possible score (compared to selecting the global optimum each trial). The final performance bonus (up to $3.00) was also calculated based on this percentage, averaged over all rounds.

In total we generated 40 different graphs by building 8x8 lattice graphs and then randomly pruning 40% of the edges, with the constraint that the resulting graph be comprised of a single connected component. We then sampled a single reward function for each graph from a GP prior, parameterized by a diffusion kernel fit on the graph (with $\alpha = 2$). The layout for each graph was pre-generated using the Fruchterman-Reingold (1991) force-directed graph placement algorithm, such that a single canonical layout for each graph was observed by all participants. For each participant, we sampled (without replacement) from the same set of 40 pre-generated graphs to build the set of 4 fully revealed graphs shown in the instructions and the 10 graphs used in the main experiment.

Judgments

Prior to beginning the very last round, participants were informed that it was a “bonus round”. The goal of acquiring as many points as possible remained the same, but after 20 clicks, participants were shown a series of 10 unrevealed nodes and asked to make judgments about the expected reward and their confidence (Fig. 6.4c). After all 10 judgments were completed, participants were forced to choose one of the 10 options, and then the task was completed as

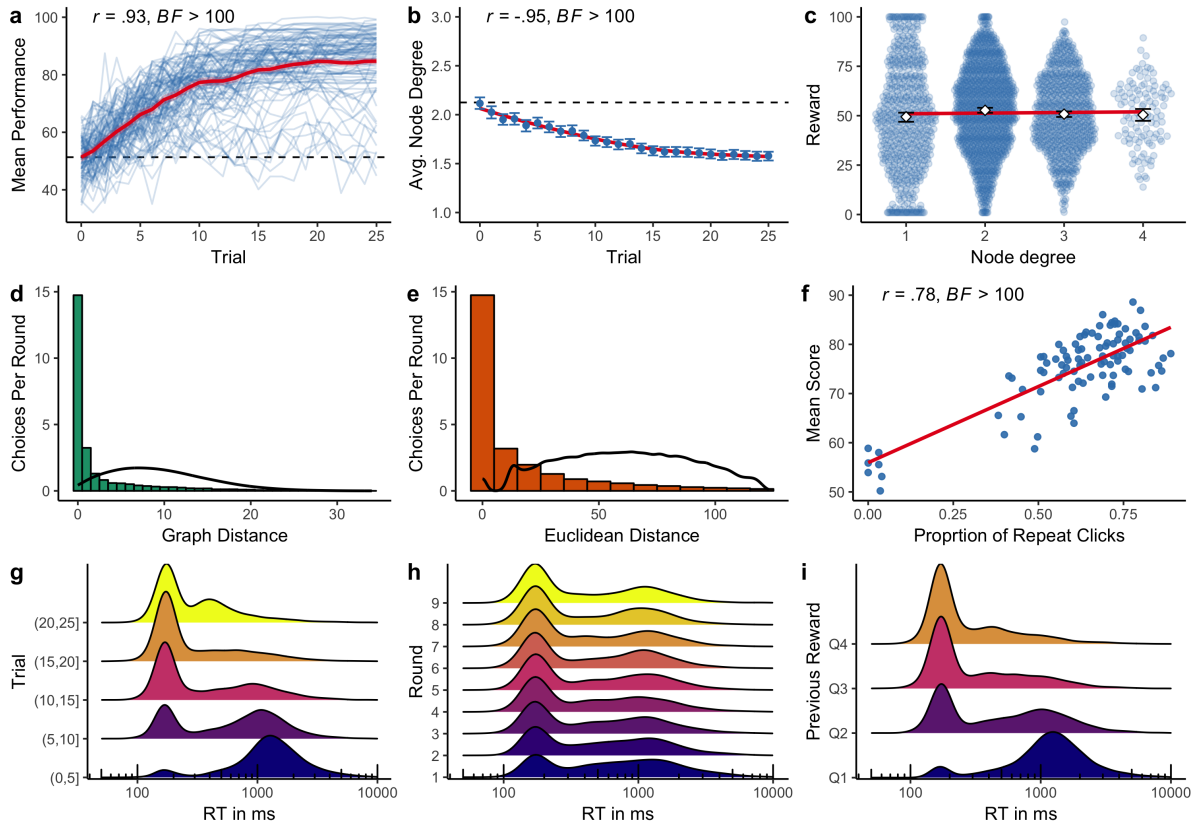

Fig. 6.5 Experiment 8 results. **a**) Mean performance over trials, where each blue line is a single participant and the red line is the group mean. The dotted line provides a comparison to a random baseline. **b**) Average degree (i.e,. number of connected nodes) of selected nodes over trials, where the dotted line indicates the mean node degree over the set of 40 environments. Each blue dot indicates the mean with 95% CI error bars with the red line showing the trend towards lower degree nodes over trials. **c**) Reward distribution of the 40 environments separated by node degree, where the white diamond indicates the mean with black 95% CI error bars. The red line is a linear regression. **d**) The distribution of graph distances (shortest path length) between choices, where the black line is a random baseline. **e**) The distribution of spatial distances between choices (based on the Fruchterman-Reingold projected coordinates), where the black line is a random baseline. **f**) The correlation between the overall proportion of repeat clicks and mean score, where each dot is a single participant and the red line is a linear regression. **g-i**) Reaction times (RT) in milliseconds (ms), over trials, rounds, and as a function of the previous reward value (split into four quartiles per participant, where Q4 represents the highest set of rewards).

normal. Behavioral and modeling results exclude the bonus round, except for the analyses of the judgment data.

6.4.2 Results

Participants performed well in the task, achieving higher rewards over successive trials ($r = .93$, $p < .001$, $BF > 100$; Fig. 6.5a) and decisively outperformed a random baseline ($t(97) = 29.6$, $p < .001$, $d = 3.0$, $BF > 100$). We find no influence of round number on performance ($r = .49$, $p = .182$, $BF = 1$), indicating that the fully revealed environments in the instructions (Fig.

6.4a) and comprehension questions were sufficient for conveying the goal of the task and the correlational structure.

Participants increasingly selected lower degree nodes over successive trials ($r = -.95$, $p < .001$, $BF > 100$; Fig. 6.5b). While this does not correspond to differences in expected rewards for the underyling reward distribution (Fig. 6.5c), lower degree nodes did tend to have more eccentric reward values. Indeed, the maximum reward across environments had an average degree of 1.28. However, the minimum reward was also more likely to be found on a low degree node, with an average degree of 1.2. This trend of preferentially sampling lower degree nodes may reflect a high-risk high-reward heuristic (Leuker, Pachur, Hertwig, & Pleskac, 2018), although the causality may also be reversed, with participants ceasing exploration once a suitably high reward has been found, which is also likely to correspond to a low degree node.

Participants also had a strong tendency to search locally, in terms of both graph distance (Fig. 6.5d) and spatial distance (Fig. 6.5e) measured using the Fruchterman-Reingold projection coordinates of the graph. Both spatial and graph distance are highly rank correlated ($r_\tau = .92$, $p < .001$, $BF > 100$), since the purpose of the projection algorithm is to accurately represent the connection structure of the graph on a 2-dimensional plane. Overall, participants had a mean rate of 62% repeat selections, with the proportion of repeats correlating with mean performance ($r = .78$, $p < .001$, $BF > 100$; Fig 6.5f). This seems to suggest that much of the improvement in the learning curves (Fig. 6.5a) are owed to few but effective exploration decisions.

Figures 6.5g-i show the reaction times (RT) of participants. Participants sped up over trials (correlated with log(RT): $r = -.64$, $p < .001$, $BF > 100$) and to a lesser extent over rounds ($r = -.20$, $p < .001$, $BF > 100$). Additionally, participants responded faster when the previous reward value was high ($r = -.59$, $p < .001$, $BF > 100$).

6.4.3 Modeling Comparison

In order to understand how participants search for rewards, we used computational modeling to make predictions about choices and the judgments from the bonus round. Models were fit using leave-one-round-out cross validation, and then compared using the summed out-of-sample prediction accuracy of the left out rounds. Altogether, we compared five different models corresponding to different strategies for generalization and exploration (see below).

Each model computes a value for each option $q(s)$, which is then transformed into a probability distribution using a softmax choice rule:

$$P(s_i) = \frac{\exp(\frac{q(s_i)}{\tau})}{\sum_j \exp(\frac{q(s_j)}{\tau})}, \tag{6.13}$$

where the temperature parameter τ is a free parameter controlling the level of random exploration. In addition, all models also use a stickiness parameter ω that adds a bonus onto the value of the most recently chosen option. This is a common feature of reinforcement learning models (Christakou, Gershman, et al., 2013; Gershman et al., 2009), which we include here to account for the high proportion of repeat clicks.

Gaussian process with diffusion kernel

The Gaussian process (GP) model uses the diffusion kernel (Eq. 6.2) to make predictive generalizations about reward, where we fit α as a free parameter defining the extent to which generalizations diffuse along the graph structure. For each node s, the GP produces normally distributed predictions that can be summarized in terms of an expected value $m(s)$ and the underlying uncertainty $v(s)$. In order to model how participants balance between exploiting high value rewards and exploring highly uncertain options, we use upper confidence bound (UCB) sampling (Auer, 2002) to produce a valuation of each node:

$$q_{\text{UCB}}(s) = m(s) + \beta\sqrt{v(s)}, \tag{6.14}$$

where the exploration bonus β is a free parameter that governs the level of exploration directed towards highly uncertain options.

Bayesian mean tracker

The Bayesian mean tracker (BMT) is a prototypical reinforcement learning model that can be interpreted as a Bayesian variant of the traditional Rescorla-Wagner (1972) model (Gershman, 2015). The BMT also produces normally distributed predictions of reward $m(s)$ and $v(s)$ for each node, but without generalization (see subsection 2.3.1). The rewards of each node are learned independently, with predictions of unobserved nodes defaulted to a prior of $m_0 = 50$ and $v_0 = 500$. The BMT has the error variance σ_ε^2 as a free parameter, which can be interpreted as inverse sensitivity. Smaller values result in larger updates to the learned mean $m(s)$ and larger reductions of uncertainty $v(s)$. The BMT also uses UCB as a sampling strategy, along with stickiness and a softmax choice rule.

Successor representation

The successor representation (SR; Dayan, 1993) is a reinforcement learning model that performs generalization based on building a predictive map of the connection structure

$$M(s,s') = (I - \gamma T)^{-1}. \tag{6.15}$$

$M(s,s')$ is the success representation matrix that models the similarity of node s to node s' based on future expected state occupancy, where we assume a random walk policy by setting the transition matrix T to the row normalized graph Laplacian $T = I - D^{-1}L$. The extent of generalization is governed by the temporal discount parameter γ, which we treat as a free parameter.

While the SR has theoretical equivalencies to the diffusion kernel (Stachenfeld et al., 2014), there are practical differences when computed on finite graphs and also by modeling the extent of generalization using the temporal discount rate γ rather than the diffusion parameter α. Additionally, the SR only makes predictions about expected value

$$m(s) = \sum_{s'} M(s,s')R(s'), \tag{6.16}$$

where $R(s')$ is the observed reward at state s'. Because there are no uncertainty estimates, the SR does not implement any directed sampling using UCB. Instead, we set $q(s) = m(s)$ and apply stickiness along with a softmax choice rule.

Nearest neighbors models

In additional to reinforcement learning models, we also consider two simple nearest neighbor averaging models. The d-nearest neighbors (dNN) model estimates expected reward for unobserved node by averaging the rewards of all observed nodes within a distance of d. The k-nearest neighbors (kNN) model estimates expected reward by averaging the observed rewards for the k nearest nodes, including all ties. Both d and k are estimated as free parameters. For predictions where no observed nodes satisfied the averaging rule (i.e., all observations were too far away), we defaulted to an expected value of $m(s) = 50$. Both dNN and kNN apply stickiness and use a softmax choice rule.

6.4.4 Model results

Figure 6.6a shows model performance in terms of predictive accuracy, which is an R^2 measure comparing the out-of-sample loss $log\mathcal{L}(\mathcal{M}_k)$ of each model k to a random model $\mathcal{M}_{\text{rand}}$:

$$R^2 = 1 - \frac{\log \mathcal{L}(\mathcal{M}_k)}{\log \mathcal{L}(\mathcal{M}_{\text{rand}})}, \tag{6.17}$$

Intuitively, $R^2 = 0$ represents random chance (i.e., predicting each of the 64 nodes with equal probability) and $R^2 = 1$ corresponds to a theoretically perfect model.

We evaluate the relative performance of models using the protected exceedence probability (pxp) as a Bayesian model selection framework for estimating the prevelance of each model in the population, corrected for chance (Rigoux et al., 2014; Stephan et al., 2009). Overall, the GP had the highest predictive accuracy ($R^2 = .56$) and produced an estimated prevalence of $pxp(GP) = .90$, with the other models having $pxp(BMT) = .07$, $pxp(SR) = .001$, $pxp(dNN) = .02$, and $pxp(kNN) = .001$. It should be noted that the high proportion of repeats contributed towards all models having high predictive accuracy levels. As a benchmark, we also fit a null model that makes the same prediction for every node, which combined with stickiness and the softmax choice rule achieved a mean predictive accuracy of $R^2 = .45$.

We also simulate the behavior of each model by sampling (with replacement) from the set of participant parameter estimates (10k samples) and computing the average learning curves (Fig. 6.6b). Although all models underperform the human curves, the GP achieves the closest levels of performance, with the BMT performing next best.

Parameter Estimates

We now turn to the parameter estimates to interpret characteristics of participant behavior captured by the models (Fig. 6.6c). The GP uses the diffusion parameter (α) to define the extent of generalization, where larger values of α imply wider influences of observed rewards over the graph structure. While the model provides evidence that generalization occurs, α estimates were systematically lower than the underlying value of $\alpha = 2$ used to generate the environments ($t(97) = -13.5$, $p < .001$, $d = 1.4$, $BF > 100$). Undergeneralization rather than overgeneralization is the norm, consistent with the results of Chapter 3, where we found potential benefits of a bias towards undergeneralization in a similar search context. Additionally, the GP not only makes predictions of reward, but also of the underlying uncertainty, where the exploration bonus (β) is used to balance the exploitation of expected rewards with the exploration of the most uncertain options. Thus, the results of the GP model capture both forms of exploration in participant behavior, suggesting a combination of both directed and random exploration describes participant behavior. Lastly, we also define a stickiness parameter (ω), which captures an aspect of the high rates of repeat clicks by adding an additional bonus to the value of the last selected options.

In comparison, while the BMT also makes uncertainty estimates, these are defaulted to the prior variance ($v_0 = 500$) for all unobserved options, making the same estimate for nodes near and far from previous observations. In contrast to the GP model, the BMT shows little evidence of directed exploration, with β estimates only marginally different from the lower bound of .007 ($t(97) = 2.1$, $p = .038$, $d = 0.2$, $BF = .92$). The BMT also made similar use of the stickiness parameter compared to the GP ($t(97) = 0.7$, $p = .460$, $d = 0.1$, $BF = .15$).

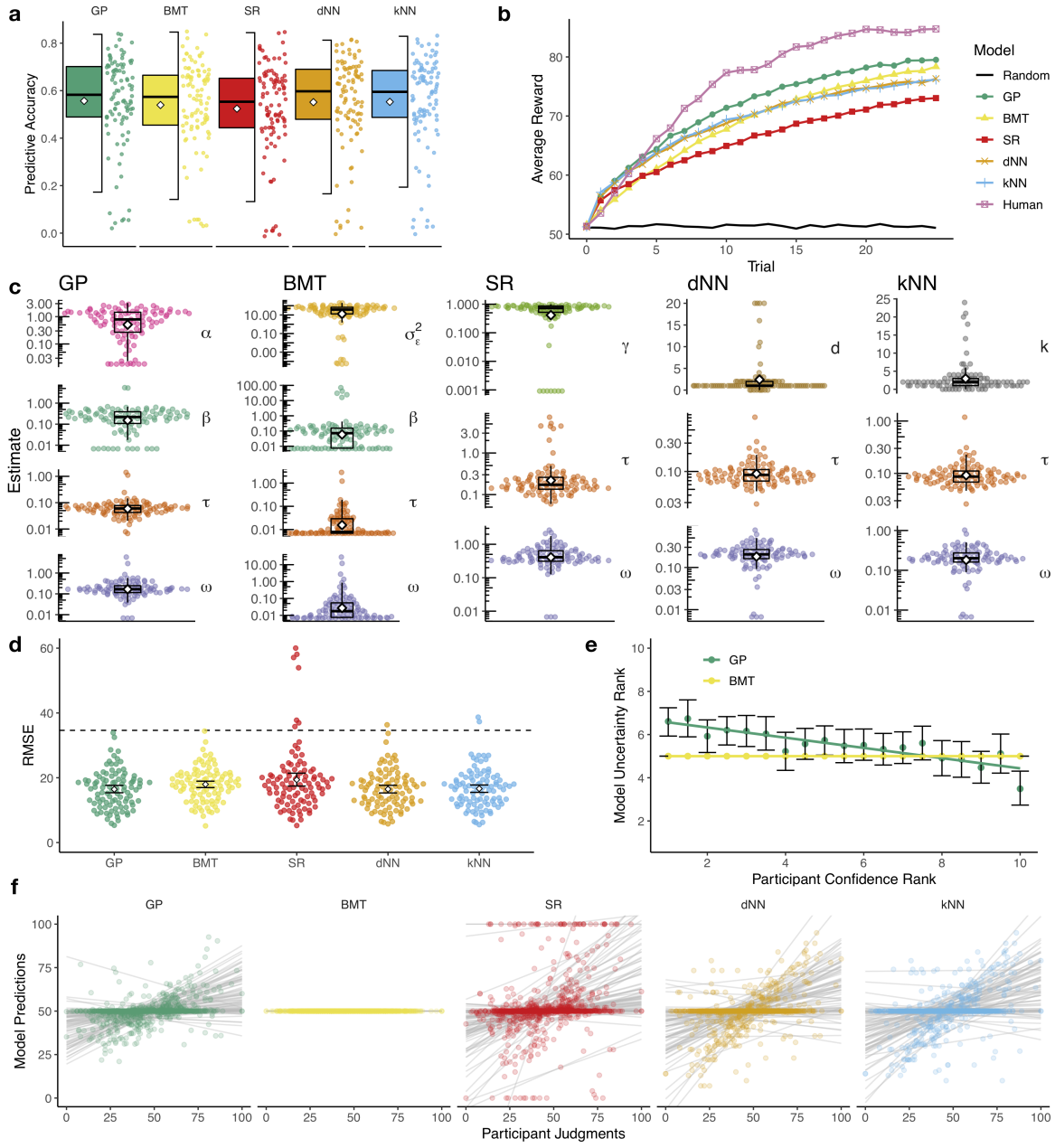

Fig. 6.6 Experiment 8 modeling results. **a**) Predictive accuracy (R^2). Each dot is a single participant with the diamond indicating group means. Tukey boxplots indicate median and 1.5 IQR. **b**) Simulated learning curves by sampling (with replacement) from participant parameter estimates (10k replications), where the black line shows a random baseline and the pink line shows mean participant performance. **c**) Participant parameter estimates, where each dot is the median cross-validated estimate, the diamond indicates group mean, and Tukey boxplots show the median and 1.5 IQR. Note that the y-axis is log-scaled for parameters except d and k, which are natural numbers. **d**) Model predictions of bonus round judgments, using the median parameter estimates from the other rounds. Performance is reported in root mean squared error (RMSE). Each dot is a single participant, with diamonds indicating group mean and the error bars showing 95% CI. The dotted line indicates a random baseline. **e**) Only the DF and BMT make uncertainty estimates. Here we show the correspondence between rank ordered (per participant) confidence judgments and model uncertainty estimates. Dots indicate means with error bars showing 95% CI, and colored lines represent a linear regreesion. **f**) The correspondence between each bonus round judgment and model predictions, where each dot is one data point and each line is a linear regression at the individual level. Model predictions were limited to the same 0-100 range as participant responses.

While the SR did not perform as well in terms of predictive accuracy or simulated learning curves, we still find sensible parameter estimates. It should be noted that the SR is better suited for sparse reward environments—in contrast to the task here—since lacking a prior over rewards, it is prone to either over or under inflation of predictions. Nevertheless, we still find estimates of the temporal discount parameter (γ) within the range commonly used in RL tasks (near the upper limit of 1). Because the SR does not make predictions about uncertainty, it only has access to random exploration, showing larger values of τ than both the GP ($t(97) = 3.6$, $p < .001$, $d = 0.5$, $BF = 49$) and BMT models ($t(97) = 4.2$, $p < .001$, $d = 0.5$, $BF > 100$), perhaps as a means to compensate for lacking directed exploration.

Turning to the nearest-neighbor models, the dNN generated predictions by averaging the rewards of observed nodes within a distance of d. The mean estimate of distance was $d = 2.4$, although the mode and median were both 1. Thus, the dNN predominately made predictions solely based on the observations of directly connected nodes. Nonetheless, it was still able to predict participant choices fairly accurately. While the dNN had no access to directed exploration, we nevertheless find similar levels of random exploration (τ: $t(97) = 1.2$, $p = .230$, $d = 0.1$, $BF = .23$) and stickiness (ω: $t(97) = -1.0$, $p = .317$, $d = 0.1$, $BF = .18$) compared to the GP. Thus, one potential source of the gap in simulated learning performance compared to the GP (Fig. 6.6b), could be the due to the dNN lacking a form of directed exploration. The kNN model also performed similar to the dNN, by averaging the k nearest nodes rather than selecting nodes at a fixed distance. The mean number of neighbors was $k = 3$, and with a mode and median of 2. Thus, like the dNN, generalizations were on the basis of integrating a small number of other observations. The dNN and kNN also shared similar levels of both undirected exploration ($t(97) = 1.8$, $p = .077$, $d = 0.3$, $BF = .52$), and stickiness ($t(97) = 1.1$, $p = .290$, $d = 0.2$, $BF = .19$).

Bonus Round

To provide additional support for our modeling results, we also predicted participant judgments in the bonus round using participant parameter estimates. Since model parameters were estimated through cross-validation on all rounds except the bonus round, we used each participant's median parameter estimates (over rounds 1 to 9) to make out-of-sample predictions about the bonus round judgments. Figure 6.6d shows the root mean squared error (RMSE) of these predictions. The GP had the lowest prediction error on average, although there was no difference in comparison to the dNN, which had the second lowest prediction error ($t(97) = -0.1$, $p = .897$, $d = 0.01$, $BF = .11$). However, accuracy alone does not indicate a good correspondence between model predictions and participant judgments. The BMT makes the same prediction for all unobserved nodes (based on the prior mean of $m_0 = 50$), yet still achieves a substantial

margin below random chance (dotted line) and comparable performance to the other models. Looking more closely at the individual correlation between participant judgments and model predictions (Fig. 6.6f), we find that overall, the GP had the highest average correlation ($r = .41$), which was better than the dNN (comparing Z-transformed correlation coefficients: $t(97) = 3.0$, $p = .004$, $d = 0.2$, $BF = 7$), and kNN models ($t(97) = 3.0$, $p = .003$, $d = 0.2$, $BF = 8$), but equally good as the SR ($t(97) = 0.1$, $p = .901$, $d = 0.0$, $BF = .11$). Correlations are undefined for the BMT, since it invariably makes the same prediction.

Lastly, we look at how predictions of uncertainty correspond to participant confidence ratings, where we interpret confidence to be the inverse of uncertainty. In this analysis, we consider only the GP and BMT models, since no other models make uncertainty estimates. Figure 6.6e shows a comparison between the (per participant) rank-ordered confidence ratings and rank-ordered uncertainty estimates of the models. While the BMT estimates the same level of uncertainty for all unobserved nodes (making correlations undefined), we find that the GP uncertainty estimates correspond to participant confidence ratings. In order to test this relationship by accounting for individual differences in subjective ratings of confidence, we fit a mixed effects model to predict the raw confidence judgment (Likert scale 1-11) using the GP uncertainty estimate as a fixed effect and participant as a random effect. These results show a strong correspondence between lower confidence ratings and higher GP uncertainty estimates ($\beta = -.30$, $t(414) = -5.7$, $p < .001$, $BF > 100$).

6.4.5 Discussion

Experiment 8 provides a rich set of behavioral and modeling results analyzing how people perform inference and search in structured environments. Participants learned efficiently, searched locally, and may have adopted the high-risk high-reward heuristic of preferentially clicking terminal nodes, which tended to have more extreme reward values. Additionally, the proportion of repeat clicks correlated strongly with performance, suggesting that lower scoring participants suffered from over-exploration rather than over-exploitation of the environment.

Our modeling results provide evidence that a Gaussian process (GP) model using the diffusion kernel is able to capture how people use generalization to guide search in structured environments. The GP provides the best predictive accuracy of choices, produces similar learning curves to human performance, and can robustly predict judgments about expected reward and confidence. Nevertheless, we also find that the dNN and kNN heuristic models are able to closely match the GP in terms of predicting choices and judgments (expected reward but not confidence). However, lacking access to directed exploration, dNN and kNN fail to produce simulated learning curves at the same level as the GP. Yet the nearest neighbor averaging rule used by the dNN and kNN models operated largely at low distances d or with small numbers

of nodes k. This highly local averaging strategy is able to effectively mimic the generalizations of the GP, where we also found lower levels of α compared to the ground truth. In both cases, overgeneralization is likely prone to larger errors than undergeneralization. Additionally, local averaging is a sensible heuristic given the local search patterns of participants. While the SR matches the GP in terms of the correspondence between participant judgments and model predictions, it performed poorly in predicting choices and in simulating human-like learning curves. Thus, while there is a theoretical equivalency between the SR and the diffusion kernel, the ability to estimate uncertainty within the GP framework gives it a clear advantage in describing search behaviour.

6.4.6 General discussion

We studied how people generalize in structured spaces, where transition structure rather than the singular stimuli features define the distribution of rewards in the environment. We use a diffusion kernel as a similarity metric capturing the transition dynamics of a graph, which offers a unifying framework subsuming distance or feature-based generalization (Shepard, 1987; Tenenbaum & Griffiths, 2001; Tversky, 1977; Wu, Schulz, Garvert, et al., 2018; Wu, Schulz, Speekenbrink, et al., 2018) as a special case where transitions are unrestricted. Across both inference and choice paradigms, we find that a Gaussian process (GP) implementation of the diffusion kernel describes how participants generalize and search for rewards, capturing intuitions about both expected rewards and underlying uncertainty.

These results provide insights into how people organize knowledge and represent similarity in structured environments. Evidence for a common encoding of both spatial and conceptual knowledge (Behrens et al., 2018; Constantinescu et al., 2016; Garvert et al., 2017) in the grid cells of the entorhinal cortex has provided neurological support for Tolman's (1948) theory of a cognitive map. Yet these representations are not only spatial, but also responsive to the transition structures of the environment. Spatial encodings are skewed based on how people move through the environment (Gustafson & Daw, 2011; Momennejad & Howard, 2018; Stachenfeld et al., 2017), creating a predictive map of the transition dynamics. Our findings support this hypothesis, by showing people are able to use a representation of structural similarity to guide exploration

We also find that variants of a simple nearest-neighbor averaging rule are surprisingly effective heuristics in environments with graph-correlated rewards. Both the dNN and kNN can be understood as binarized simplification of the similarity metric used by the GP. While the GP predicts expected rewards using a similarity-weighted sum of previous observations (Eq. 6.6), the dNN and kNN use either a distance or count-based threshold of similarity, such that nodes are either similar if considered a neighbor, or dissimilar otherwise. Similar nodes

are then averaged, equivalent to an equal-weight regression model (Lichtenberg & Simsek, 2016; Wesman & Bennett, 1959). Although these heuristics are able to efficiently capture many aspects of judgments and choices, our results also show that human behavior is more sensitive to the transition structure of the environment, as evidenced by the better predictive power of the GP—particularly in its ability to represent uncertainty, which is an essential ingredient for directed exploration.

One limitation of the diffusion kernel is that it assumes *a priori* knowledge of the graph structure. While this may be a reasonable assumption in problems such as navigating a a subway network where a map is readily available, this is not always the case. In contrast, the SR can learn the graph structure through experience, using prediction-error learning to update $M(s, s')$. Thus, the connection between the SR and the diffusion kernel presents a promising avenue for integrating a method for structure learning into the GP framework (but also see Kemp & Tenenbaum, 2008, 2009). The benefits of this equivalency go both ways, because the results of Experiment 8 show that that the uncertainty estimates of the GP play a key role in how participants actively explore the environment and estimate confidence about predictions. Thus, one remaining challenge for theories linking the SR to the predictive coding of grid cells is to explain how people represent uncertainty.

Chapter 7

Conclusions

O snail
Climb Mount Fuji
But slowly, slowly!
—Issa

This chapter provides concluding remarks about the main findings, discusses related research work that has not been included, and proposes future directions.

7.1 Summary of major contributions

The central question of this thesis is how people are able to adapt to endlessly novel situations and to learn efficiently in vast and complex environments. Inspired by influential theories of stimulus and category generalization, we adapted the principles of similarity-based generalization into a function learning framework, where functions represent inferred relationships between actions and outcomes. This provides a model of predictive generalization operating across an entire continuum of possibilities. Our framework uses kernel similarity to describe relationships in both metric and structured representations of knowledge, allowing us to sidestep influential critiques (Tversky, 1977) and to draw connections to neurological theories about how knowledge is represented in the brain (Behrens et al., 2018; Kaplan et al., 2017; Schuck et al., 2016; Stachenfeld et al., 2017; Theves et al., 2019).

Across spatial, conceptual, and structured domains, we have shown that our model of generalization provides compelling predictions about how people search, make judgments, and rate confidence about novel items. Not only does our model produce predictive generalizations about unobserved rewards that correspond to human choices and judgments, but it also tracks uncertainty in a manner consistent with how people judge confidence and direct their exploration towards uncertain regions of the environment. Our model comparisons are recoverable, the parameters are identifiable, and we are able to reliably simulate learning curves resembling human performance.

7.1.1 Domain general principles

The results of the 8 experiments presented in this thesis point towards domain general principles governing how people reason about novel situations, by leveraging the predictive power of related experience. A central component for our framework of generalization is the theory that people organize knowledge in a way that facilitates the representation of similarity. This is not a new idea, and owes a great deal to Shepard's (1987) theory of "psychological space" and Tolman's (1948) concept of a "cognitive map". However, the contribution of this thesis has been to provide a common framework for generalization, based on a similarity representation that bridges the gap between spatial and structured organizations of knowledge.

Our framework closely mirrors current advances in neuroscience, which propose grid cells in the entorhinal cortex as the physical instantiation of a predictive map of the environment (Behrens et al., 2018; Hafting et al., 2005). Grid cells use the same encoding for representing both spatial and conceptual information (Constantinescu et al., 2016; Theves et al., 2019), and are also sensitive to the transition dynamics of the environment (Momennejad et al., 2017; Stachenfeld et al., 2017). The Successor Representation (SR; Dayan, 1993) has emerged as

a prominent model of how grid cells encode state dynamics (Gershman, 2018b; Gustafson & Daw, 2011; Stachenfeld et al., 2017), which has equivalencies to the diffusion kernel we use to represent similarity in Chapter 6. The diffusion kernel is in turn equivalent to the RBF kernel (used throughout this thesis) when the environment is continuous and unrestricted, which is specifically the type of flat topology used by Shepard (1987) to construct his psychological space and produce the law of generalization. These connections offer rich opportunities for theory integration, with the captivating implication that generalization depends how we represent the world.

7.1.2 Undergeneralization

An outstanding question that appears repeatedly in this thesis, is why the tendency towards undergeneralization? In Chapter 3 we show that this is not necessarily a maladaptive bias (Figs. 3.4,B.3), but can in fact lead to better outcomes than even an exact match to the true correlation structure of the environment. However, in Chapter 4 we find even lower levels of generalization in children, where simulation results make it harder to justify the undergeneralization of children as purely beneficial (Fig. C.3). Rather, this could also be due to computational limitations, where computing generalizations over larger representational distances is more demanding. One way to simplify the computational complexity of our model would be to set the similarity of $k(\mathbf{x}, \mathbf{x}') = 0$ when the distance between $\mathbf{x}$ and $\mathbf{x}'$ is sufficiently large. Conceptually similar approaches are used in sparse Gaussian process implementations (Bauer, van der Wilk, & Rasmussen, 2016; Quiñonero-Candela & Rasmussen, 2005), while the nearest neighbors models in Chapter 6 can also be interpreted as implementing a simplified similarity metric by using only binary values $k(\mathbf{x}, \mathbf{x}') \in \{0, 1\}$, for neighbors and non-neighbors, respectively. Although it may be tempting to empirically test the exact sparsity of human similarity representations (spanning a continuum between our current GP implementation and the radical simplification of the nearest neighbors model), we advise caution due to concerns with model identifiability. Nevertheless, this is an important component in our future enterprise of building a process level theory of generalization.

7.2 Towards a process level theory

While the GP model of generalization provides a useful computational level description of human behavior (Marr, 1982; Marr & Poggio, 1976), how could such a model be plausibly implemented? On one side, we are faced with the challenges of explaining attentional mechanisms and working memory limitations, which work together to define the set of previous

experiences we can draw from. On the other side, we require a computationally plausible implementation of our GP model, where in our current implementation, an important question is how people measure and represent uncertainty.

7.2.1 Working memory and attentional mechanisms

In the experiments presented in this thesis, we side-stepped the issue of working memory limitations by presenting the full history of observations. This is not generally true in most human environments, where we depend on an internal (and capacity limited) representation of previous experience. Limitations on what we can represent plays a large role in RL problems (Collins & Frank, 2012), and has been proposed as a causal mechanism underlying learning impairments in schizophrenia (Collins, Brown, Gold, Waltz, & Frank, 2014). An account of working memory my also introduce new sources of noise during storage and retrieval, which have been related to selective overweighting of numerical information (Spitzer, Waschke, & Summerfield, 2017) and attributed as a source of random exploration (Findling, Skvortsova, Dromnelle, Palminteri, & Wyart, 2018). However, rate-distortion theory (Sims, 2018) is able to relate Shepard's law of generalization to capacity limits on information channels, where an exponential generalization function arises naturally from a boundedly rational agent attempting to reduce costly errors. Thus, in the search for new details to add to our model, we may find ourselves at its origin.

Integrating a theory of working memory into our GP model requires selecting an appropriate subset of previous observations in order to compute posterior predictions. The question of which experiences to include is closely connected to the problem of selective attention (Dayan et al., 2000). Indeed, models of reinforcement learning are becoming increasingly popular as tools for studying affective disorders (Addicott, Pearson, Sweitzer, Barack, & Platt, 2017; Ziegler, Pedersen, Mowinckel, & Biele, 2016), where the modulation of attention is thought to play a key role. Additionally, the majority of the experiments presented in this thesis use two dimensional environments[†], whereas naturalistic problems may have an arbitrarily large number of dimensions. Consider how many potentially relevant features are at play when deciding where to go for holidays or selecting a job applicant. From this panoply of feature dimensions, we must selectively allocate attention to a compact set of relevant features (Niv et al., 2015). The building of this compact and relevant representation is an important question in reinforcement learning (Radulescu et al., 2019), which might be closely related to the problem of learning structured representations (Gershman & Niv, 2010) we discussed in Chapter 6.

[†]It is an NP-hard problem to determine the dimensionality of the graph structures used in Experiments 7 and 8, although there exists an upper bound of twice the maximal degree plus one (Erdős & Simonovits, 1980).

This might also be a type of search problem, where one must selectively explore which feature dimensions are most predictive of reward.

7.2.2 Representing uncertainty

Throughout this thesis, we find that representations of uncertainty play an essential role in guiding exploration. This is one of the main advantages of the GP framework, allowing us to model directed exploration and predict confidence judgments about unobserved stimuli. The expected reward predictions of the GP are quite straightforward to compute as a weighted sum of similarities (Eq. 6.6). This is equivalent to an RBF network[†], which has been proposed as a model of how the brain processes perceptual information (Poggio & Girosi, 1989, 1990). However, there is still a gap between how a GP and a human processes uncertainty.

In recent work not included in this thesis, we looked at how cognitive limitations imposed by time pressure influence how people respond to uncertainty (Wu, Schulz, Gerbaulet, et al., 2019). We constructed a simple four-armed bandit task with independent reward distributions specifically chosen to disentangle differences in expected value from differences in uncertainty. We found that given unlimited time, participants were positively influenced by both high rewards and high uncertainty, consistent with results throughout this thesis. But given only 400 milliseconds to make each decision, we found that participants responded by changing their behavior to actively avoid uncertainty. These results are a useful probe into how learning and exploration strategies are influenced by computational limitations (either internally or externally imposed). We hypothesized that compared to random exploration, direction exploration may require more deliberation and control, which would be negatively affected by time pressure. These results suggest that grappling with uncertainty takes time and effort, yet in Chapter 4 we find that even children as young seven are capable of using uncertainty directed exploration. Current work in progress is extending the framework of Experiment 4 to even younger children, by removing numbers (to avoid limitations on how high they can count) and relying only instead on colors instead. Taken together, these are the first steps towards building a process-level and developmental theory of how people represent uncertainty.

[†]There is also a more exact equivalency between GPs and Bayesian neural networks (Neal, 1994, 2012), which accounts for both expected rewards and uncertainty estimates. While Bayesian neural networks are certainly useful in machine learning, there lacks a theory of how Bayesian neural networks might be related to human information processing.

7.3 Related work

In defining the scope of this thesis, it was necessary to omit other bodies of research conducted during my PhD in order to present a more coherent account of a focused research question. Nevertheless, these other lines of research warrant a brief overview, since they address related and equally important questions about the nature of human learning and cognition.

7.3.1 Search in risky environments

The search tasks presented in this thesis described only situations where outcomes had positive and immediate rewards. This is of course only a useful abstraction to focus on the phenomena of generalization. However, a justifiable critique would be to ask how our results carry over to the domain of losses (Palminteri, Khamassi, Joffily, & Coricelli, 2015; Seymour, Daw, Dayan, Singer, & Dolan, 2007) or for alternative reward distributions (e.g., binary gambles; Brandstätter, Gigerenzer, & Hertwig, 2006; Hertwig et al., 2004; Lejarraga, Pachur, Frey, & Hertwig, 2016).

Using a search task similar to Experiments 1 and 2, we conducted two experiments where instead of only positive rewards, included a condition with risky outcomes (E. Schulz, Wu, et al., 2018). Specifically, if participants revealed a reward below a specified threshold, the rest of the round would be forfeit, along with any future rewards that could have been earned. In this risky context, we still find that a GP model of generalization provided the best prediction of human behavior, but instead of an optimistic UCB sampling strategy, participants were best described by a *probability of being safe* (POS) strategy to avoid potentially risky outcomes. This strategy resembles the probability of improvement (POI) strategy (Eq. 2.24), but rather than trying to improve on the current best outcome, it selectively samples options by the likelihood they are above the safe threshold. Thus, people are able to adapt to risky contexts by changing their sampling strategy, but still use similar principles of generalization.

Additionally, not all choices have the same reward structure. In Analytis, Wu, and Gelastopoulos (2019), we studied a dynamic decision-making problem where effort is allocated between a safe task with relatively predictable linear rewards, or a make-or-break task characterized by dramatic success or failure outcomes. This type of heterogeneous choice problem is quite familiar to anyone who has had to balance writing a risky yet potentially life-changing grant proposal with submitting a paper to a "safe" journal with a more predictable outcome. While it is possible to define an optimal solution when the reward structure is known, solving it via backwards induction is computationally expensive. We show that simple strategies relying on myopic or heuristic principles perform competitively against the optimal strategy, yet sidestep much of the computational burden. These boundedly rational strategies also seem

to produce distinctive risk-averse or risk-seeking behaviors, hinting towards the possibility that these behavioral patterns arise not only from inherent preferences, but perhaps also due to a reliance on boundedly rational strategies.

7.3.2 Searching for information instead of rewards

Instead of searching for rewards, we sometimes search for answers to our questions (Crupi, Nelson, Meder, Cevolani, & Tentori, 2018; Nelson, 2005; Wu, Meder, et al., 2017). For example, consider a doctor selecting a medical test to facilitate the diagnosis of a patient with an unknown disease, or a scientist designing an experiment to test a hypothesis. This paradigm of information search as query selection often assumes either descriptive information about the underlying environmental probabilities (Skov & Sherman, 1986; Slowiaczek, Klayman, Sherman, & Skov, 1992; Wu, Meder, et al., 2017) or sufficient experience such that hardly any uncertainty remains (Meder & Nelson, 2012; Nelson, McKenzie, Cottrell, & Sejnowski, 2010). Yet, an important questions is how people can learn what is a good and what is a bad question through experience. Using a variant of a multi-armed bandit paradigm, where each option corresponds to a query and yields information instead of rewards, we show that a *take-the-difference* (TTD) heuristic using only the absolute difference between absolute frequencies of outcomes (Wu, Meder, et al., 2017) produces competitive performance compared to substantially more complex Bayesian sampling models (Wu, Meder, & Nelson, in prep). This highlights the importance of how different representations (e.g., natural frequencies vs. conditional probabilities; Gigerenzer & Hoffrage, 1995; McDowell & Jacobs, 2017) can facilitate different cognitive strategies.

7.3.3 The social dimension of search

In focusing on the individual mechanisms of learning and search, we have omitted a crucial dimension of human cognition. We do not only learn from our own experiences, but also from that of other people (Hills et al., 2015; March, 1991; Rendell et al., 2010). The metaphor of "learning as search" is equally applicable in a social domain, where multiple individuals simultaneously search for rewards and can interact with one another by exchanging information or competing for resources.

In a social environment, the exploration-exploitation dilemma is distributed across the collective, dependent on the flow of information between individuals. In order to resolve contradictory results about which types of communication networks produce better collective outcomes (Lazer & Friedman, 2007; Mason & Watts, 2012), we conducted an environmental analysis using multi-agent simulations (Barkoczi, Analytis, & Wu, 2016). More connected networks result in faster collective convergence and thus more exploitation, while less connected

networks facilitate higher individual variability and more exploration at the group level (Mason, Jones, & Goldstone, 2008). Thus, the contradictions in the literature (Lazer & Friedman, 2007; Mason & Watts, 2012) were due to two separate research traditions studying disjoint sets of environments, whereas different environments simply enforce different demands on rational behavior.

Additionally, instead of only assuming information is unilaterally shared along a deterministic social network, we also examined the rationality of when to freely share information (Bouhlel, Wu, Hanaki, & Goldstone, 2018; Tump, Wu, Bouhlel, & Goldstone, 2019). Inspired by the open source movement, we used both agent-based simulations and evolutionary algorithms to study the conditions in which it is beneficial to openly share information without any expectations of reciprocity. Similar to research on swarm intelligence in biology (Berdahl, Torney, Ioannou, Faria, & Couzin, 2013; Brown, Brown, & Shaffer, 1991), we found that shared information facilitates the formation of a spatially coordinated group that searches better than any lone individual. This provides the original donor with by-product benefits that can outweigh the costs of increased competition. This work provides guidance for developing environments that can foster cooperation, even in anonymous settings (e.g., online) where traditional reputation-based mechanisms of reciprocity are absent.

7.4 Conclusion

In summary, this thesis presents a theory of generalization across spatial, conceptual, and structured spaces. Our theory offers compelling evidence about how generalization is used to navigate uncertainty and efficiently direct search efforts. We use Gaussian process regression as a model of generalization, which leverages feature-based or structural similarity to make predictive generalizations about where search seems most promising. Predictions are in the form of a distribution over functions, each corresponding to a hypothesis about the relationship between actions and outcomes. These predictions can be decomposed into expectations of reward along with an estimate of the underlying uncertainty, where both play a key role in navigating the exploration-exploitation dilemma. Converging evidence from search behavior and judgments about expected reward and confidence provide broad support for our model, which is both recoverable and capable of simulating human-like learning curves. This work is closely related to neurological theories of how the brain encodes a predictive map of the environment, and offers broad implications for understanding the nature of adaptive behavior in the face of uncertainty.

Appendix A

Statistical tests

A.1 Comparisons

Both frequentist and Bayesian statistics are reported throughout this thesis. Frequentist tests are reported as Student's t-tests (specified as either paired or independent) for parametric comparisons, while the Mann-Whitney U test or Wilcoxon signed-rank test are used for non-parametric comparisons (for independent samples or paired samples, respectively). Each of these tests are accompanied by a Bayes factors (BF) to quantify the relative evidence the data provide in favor of the alternative hypothesis (H_A) over the null (H_0).

Parametric comparison are tested using the default two-sided Bayesian t-test for either independent or dependent samples, where both use a Jeffreys-Zellner-Siow prior with its scale set to $\sqrt{2}/2$, as suggested by Rouder, Speckman, Sun, Morey, and Iverson (2009). All statistical tests are non-directional as defined by a symmetric prior (unless otherwise indicated).

Non-parametric comparisons are tested using either the frequentist Mann-Whitney-U test for *independent samples*, or the Wilcoxon signed-rank test for *paired samples*. In both cases, the Bayesian test is based on performing posterior inference over the test statistics (Kendall's r_τ for the Mann-Whitney-U test and standarized effect size $r = \frac{Z}{\sqrt{N}}$ for the Wilcoxon signed-rank test) and assigning a prior using parametric yoking (van Doorn, Ly, Marsman, & Wagenmakers, 2017). This leads to a posterior distribution for Kendall's r_τ or the standarized effect size r, which yields an interpretable Bayes factor via the Savage-Dickey density ratio test. The null hypothesis posits that parameters do not differ between the two groups, while the alternative hypothesis posits an effect and assigns an effect size using a Cauchy distribution with the scale parameter set to $1/\sqrt{2}$.

A.2 Correlations

For testing linear correlations with Pearson's r, the Bayesian test is based on Jeffreys (1961) test for linear correlation and assumes a shifted, scaled beta prior distribution $B(\frac{1}{k}, \frac{1}{k})$ for r, where the scale parameter is set to $k = \frac{1}{3}$ (Ly, Verhagen, & Wagenmakers, 2016).

For testing rank correlations with Kendall's tau, the Bayesian test is based on parametric yoking to define a prior over the test statistic (van Doorn, Ly, Marsman, & Wagenmakers, 2018), and performing Bayesian inference to arrive at a posterior distribution for r_τ. The Savage-Dickey density ratio test[†] is used to produce an interpretable Bayes Factor.

A.3 ANOVA

We use a one-way analysis of variance (ANOVA) to compare the means of $p \geq 2$ samples based on the F distribution. In general terms, we can define ANOVA as a linear model:

$$\mathbf{y} = \mu\mathbf{1} + \sigma\mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{A.1}$$

where $\mathbf{y}$ is a vector of N observations, μ is the aggregate mean, $\mathbf{1}$ is a column vector of length N, σ is the scale factor, $\mathbf{X}$ is the $N \times p$ design matrix, $\boldsymbol{\theta}$ is a column vector of the standardized effect sizes, and $\boldsymbol{\varepsilon}$ is a column vector containing the i.i.d. errors where $\varepsilon_i \overset{\text{i.i.d}}{\sim} \mathcal{N}(0, \sigma^2)$.

We assume independent g-priors (Zellner & Siow, 1980) for each effect size $\theta_1 \sim \mathcal{N}(0, g_1\sigma^2), \cdots, \theta_p \sim \mathcal{N}(0, g_p\sigma^2)$, where each g-value is drawn from an inverse chi-square prior with a single degree of freedom $g_i \overset{\text{i.i.d}}{\sim}$ inverse-$\chi^2(1)$. For μ and σ^2 we assume a Jeffreys (1946) prior. Following Rouder, Morey, Speckman, and Province (2012), we compute the Bayes factor by integrating the likelihoods with respect to the prior on parameters, where Monte Carlo sampling was used to approximate the g-priors. The Bayes factor reported in the text can be interpreted as the log-odds of the model relative to an intercept-only null model.

A.4 Regression coefficients

We use mixed effects regression to quantify the relationship between a fixed effect and the observed data, by accounting for random effects due to individual variability. We can describe the mixed-effects model as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \tag{A.2}$$

[†]If you are reading this footnote and you are on my committee, you win a bottle of whiskey! In the unlikely event of multiple readers, the prize will be awarded to the first to point it out during the defense. However, the prior probability of anybody reading this is extremely low and set to $\beta(10^{120}, 1)$.

where $\mathbf{y}$ is the vector of observations, $\boldsymbol{\beta}$ is a vector of fixed effects, $\mathbf{u}$ is a vector of random effects, and where $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices relating $\mathbf{y}$ to $\boldsymbol{\beta}$ and $\mathbf{u}$, respectively. We report the standarized effect size $\beta \in [-1,1]$[†], the t-statistic and corresponding p-value based on Satterthwaite's (1946) approximation of the degrees of freedom, where parameters were estimated using restricted maximum likelihood estimation (REML). However, because these frequentist statistics in the mixed-effects framework are prone to being anti-conservative (Luke, 2017), we also report an interpretable Bayes factor.

In the Bayesian framework, we estimate the model posteriors using No-U-Turn sampling (Hoffman & Gelman, 2014) as a form of Hamiltonian Monte Carlo Markov Chain method. We compute the Bayes factor quantifying the log-odds of our model against a null intercept-only model, where we use bridge sampling (Gronau et al., 2017) as a method to approximate the marginal likelihood of both models.

A.5 Protected exceedance probability

Protected exceedance probability (*pxp*; Rigoux et al., 2014; Stephan et al., 2009) is defined in terms of a Bayesian model selection framework for group studies. Intuitively, it can be described as a random-effect analysis, where models are treated as random effects and are allowed to differ between subjects. Inspired by the Polya's urn model (e.g., Wei, 1979), we can imagine a population containing K different types of models (i.e., people best described by each model) much like an urn containing different colored marbles. If we assume that there is a fixed but unknown distribution of models in the population, what is the probability of each model being more frequent in the population than all other models in consideration?

This is modelled hierarchically, using variational Bayes to estimate the parameters of a Dirichlet distribution describing the posterior probabilities of each model $P(m_k|\mathbf{y})$ given the data $\mathbf{y}$. The exceedance probability is thus defined as the posterior probability that the frequency of a model r_{m_k} is larger than all other models $r_{m_{k'\neq k}}$ under consideration:

$$xp(m_k) = p(r_{m_k} > r_{m_{k'\neq k}}|\mathbf{y}) \tag{A.3}$$

Rigoux et al. (2014) extends this approach by correcting for chance, based on the Bayesian Omnibus Risk (*BOR*), which is the posterior probability that all model frequencies are equal:

$$pxp(m_k) = xp(m_k)(1-BOR) + \frac{BOR}{K} \tag{A.4}$$

[†]Note the difference between ther scalar standarized effect-size β and the vector of fixed effects $\boldsymbol{\beta}$ used in the general equation for the linear mixed effects model

This produces the *protected exceedance probability* (*pxp*) reported throughout this chapter, and is implemented using `https://github.com/sjgershm/mfit/blob/master/bms.m`.

Appendix B

Chapter 3 supplementary materials

B.1 Model recovery

We present model recovery results that assess whether or not our predictive model comparison procedure allows us to correctly identify the true underlying model. To assess this, we generated data based on each individual participant's parameter estimates. We generated data using the Option Learning and the Function Learning Model for Experiment 1 and the Option Learning* Model and the Function Learning* Model for Experiments 2 and 3. In all cases, we used the UCB sampling strategy in conjunction with the specified learning model. We then utilized the same cross-validation method as before in order to determine if we could successfully identify which model generated the underlying data. Figure B.1 shows the cross-validated predictive performance (half boxplot with each data point representing a single simulated participant) for the simulated data, along with the number of simulated participants best described (inset icon array).

B.1.1 Experiment 1

In the simulation for Experiment 1, our predictive model comparison procedure shows that the Option Learning Model is a better predictor for data generated from the same underlying model, whereas the Function Learning model is only marginally better at predicting data generated from the same underlying model. This suggests that our main model comparison results are robust to Type I errors, and provides evidence that the better predictive accuracy of the Function Learning model for participant data is unlikely due to overfitting.

When the Function Learning Model has generated the underlying data, the same Function Learning Model achieves a predictive accuracy of $R^2 = .4$ and describes 41 out of 81 simulated participants best, whereas the Option Learning model achieves a predictive accuracy of $R^2 = .39$

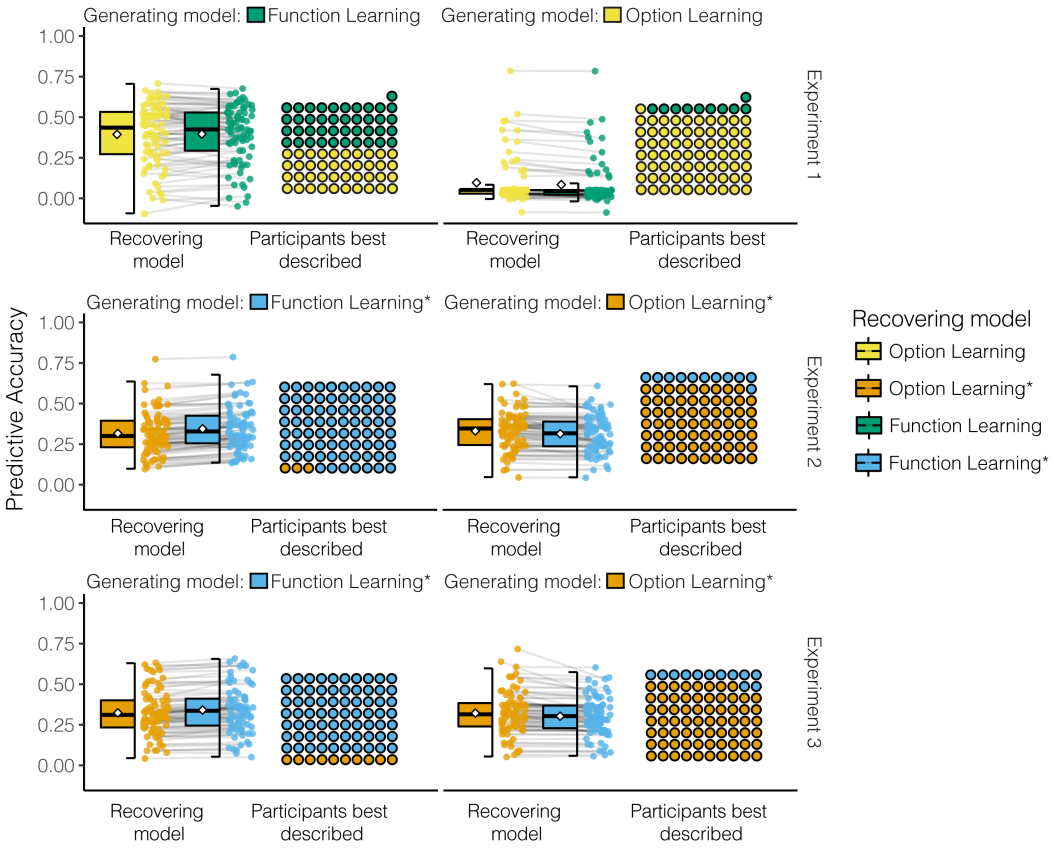

Fig. B.1 Model recovery (Experiments 1-3). Data was generated by the specified generating model (left and right columns) using individual participant parameter estimates. The recovery process used the same cross-validation method used in the model comparison. We report the predictive accuracy of each candidate recovery model (colours). Boxplots show the median (line), mean (diamond), interquartile range (box), and 1.5x IQR (whiskers). Each individual (simulated) participant is represented as a dot, with lines connecting each simulated participant. Icon arrays show the number of simulated participants best described. For both generating and recovery models, we used UCB sampling. Table B.3 reports the median values of the cross-validated parameter estimates used to specify each generating model.

and describes 40 participants best. Furthermore, the protected probability of exceedance for the Function Learning Model is pxp $= 0.51$. This makes our finding of the Function Learning Model as the best predictive model even stronger as, technically, the Option Learning Model could mimic parts of the Function Learning behavior.

When the Option Learning Model generates data using participant parameter estimates, the same Option Learning Model achieves an average predictive accuracy of $R^2 = .1$ and describes 71 out of 81 simulated participants best. On the same generated data, the Function Learning Model achieves an average predictive accuracy of $R^2 = .08$ and only describes 10 out of 81 simulated participants best. The protected probability of exceedance for the Option Learning Model is pxp $= 0.99$. If the counterfactual had occurred, namely that if data generated by

the Option Learning Model had been best predicted by the Function Learning Model, we would need to be sceptical about our modelling results on the basis that the wrong model could describe data better than the true generating model. However, here we see that the Function Learning Model does not make better predictions than the true model for data generated by the Option Learning Model.

B.1.2 Experiment 2

In the simulations for Experiment 2, we used the localized version of each type of learning model for both generation and recovery, since in both cases, localization improved model accuracy in predicting the human participants (Table B.3). Here, we find very clear recoverability in all cases, with the recovering model best predicting the vast majority of simulated participants when it is also the generating model (Fig. B.1).

When the Function Learning* Model generates the underlying data, the same Function Learning* Model achieves a predictive accuracy of $R^2 = .34$ and describes 77 out of 80 simulated participants best, whereas the Option Learning* Model describes only 3 out of 80 simulated participants best, with a average predictive accuracy of $R^2 = .32$. The protected probability of exceedance for the Function Learning* model is $\text{pxp} = 1$.

When the Option Learning* Model generates the data, the same Option Learning* Model achieves a predictive accuracy of $R^2 = .33$ and predicts 69 out of 80 simulated participants best, whereas the Function Learning* Model predicts only 11 simulated participants best, with an average predictive accuracy of $R^2 = .31$. The protected probability of exceedance for the Option Learning* model is $\text{pxp} = 1$. Again, we find evidence that the models are indeed discriminable, and that the Function Learning* Model does not overfit data generated by the wrong model.

B.1.3 Experiment 3

We again find in all cases the best recovery model is the same as the generating model. When the Function Learning* Model generates data, the matched recovery with the same Function Learning* Model best predicts 70 out of 80 participants, with an average predictive accuracy of $R^2 = .34$. The Option Learning* Model best predicts the remaining 10 participants, with an average predictive accuracy of $R^2 = .32$. The protected probability of exceedance for the Function Learning* model is $\text{pxp} = 1$.

When the Option Learning* Model generates the data, the same Option Learning* Model best predicts 68 out of 80 participants with an average predictive accuracy of $R^2 = .32$, whereas the Function Learning* Model only best predicts 12 out of 80 participants with an average

predictive accuracy of $R^2 = .3$. The protected probability of exceedance for the Option Learning* model is pxp $= 1$.

In all simulations, the model that generates the underlying data is also the best performing model, as assessed by predictive accuracy, the number of simulated participants predicted best, and the protected probability of exceedance. Thus, we can confidently say that our cross-validation procedure distinguishes between these model classes. Moreover, in the cases where the Function Learning or Function Learning* Model generated the underlying data, the predictive accuracy of the same model is not perfect (i.e., $R^2 = 1$), but rather close to the predictive accuracies we found for participant data (Table B.3).

B.1.4 High temperature recovery

We also assessed how much each model's recovery can be affected by the underlying randomness of the softmax choice function. For every recovery simulation, we selected the 10 simulations with the highest underlying softmax temperature parameter τ (ranges: $\tau^{10}_{Exp1} = [0.09, 0.42]$, $\tau^{10}_{Exp2} = [0.11, 0.25]$, $\tau^{10}_{Exp3} = [0.21, 9.7]$) and again calculated the probability of exceedance for the true underlying model. The results of this analysis led to a probability of exceedance for the Function Learning Model in Experiment 1 of pxp $= .81$, for the Function Learning* Model in Experiment 2 of pxp $= 0.99$, for the Function Learning* Model in Experiment 3 of pxp $= 0.93$, for the Option Learning Model in Experiment 1 of pxp $= 0.97$, for the Option Learning* Model in Experiment 2 of pxp $= 0.99$, and for the Option Learning Model in Experiment 3 of pxp $= 0.98$. Thus, the models seem to be well-recoverable even in scenarios with high levels of random noise in the generated responses.

B.2 Parameter recovery

Another important question is whether or not the reported parameter estimates of the two Function Learning models are reliable and robust. We address this question by assessing the recoverability of the three parameters of the Function Learning model, the length-scale λ, the exploration factor β, and the temperature parameter τ of the softmax choice rule. We use the results from the model recovery simulation described above, and correlate the empirically estimated parameters used to generate data (i.e., the estimates based on participants' data), with the parameter estimates of the recovering model (i.e., the MLE from the cross-validation procedure on the simulated data). We assess whether the recovered parameter estimates are similar to the parameters that were used to generated the underlying data. We present parameter recovery results for the Function Learning Model for Experiment 1 and the Function Learning*

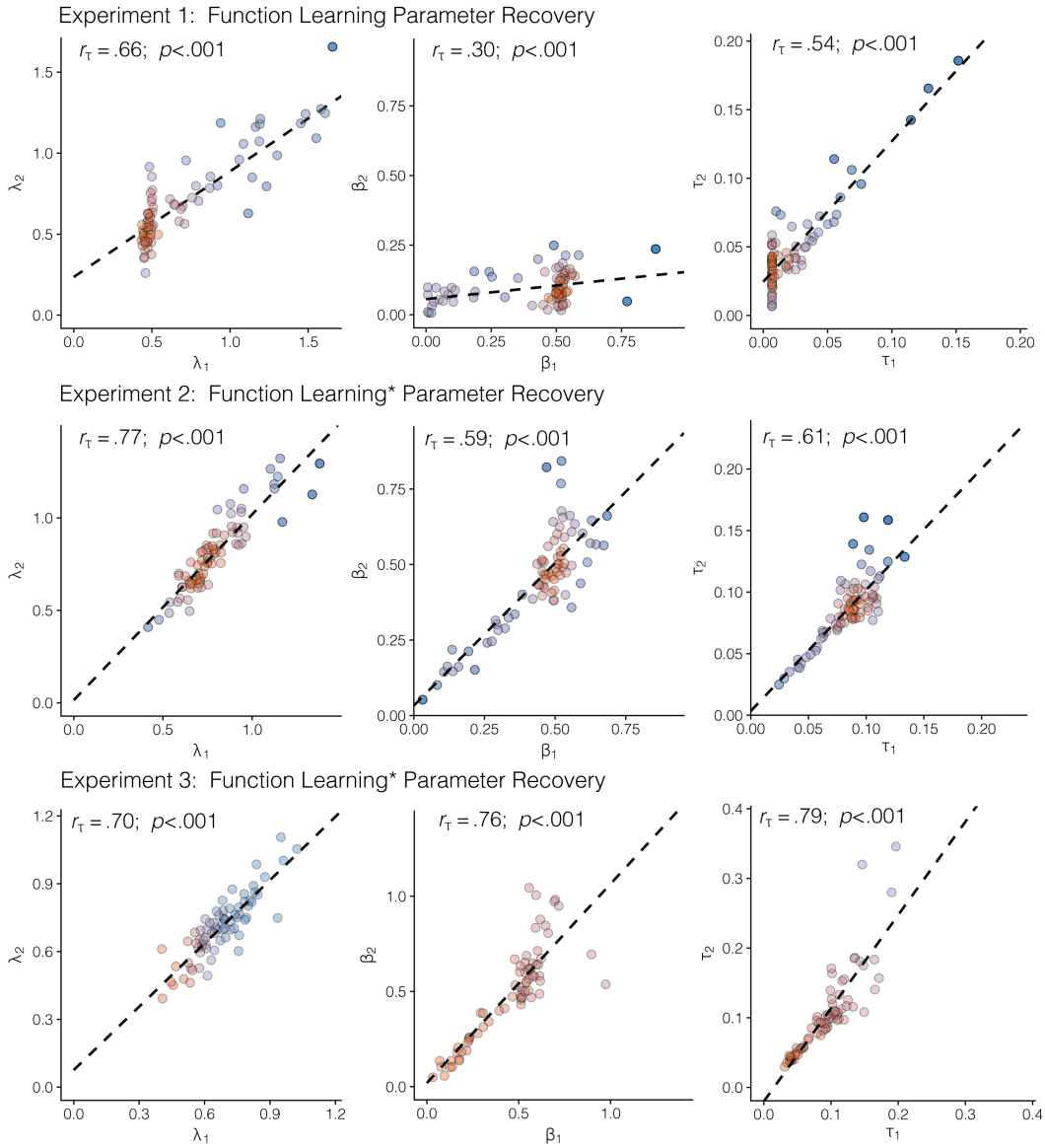

Fig. B.2 Parameter recovery (Experiments 1-3). The generating parameter estimate is on the x-axis and the recovered parameter estimate is on the y-axis. The generating parameter estimates are from the cross-validated participant parameter estimates, which were used to simulate data. Recovered parameter estimates are the result of the cross-validated model comparison on the simulated data. While the cross-validation procedure yielded k estimates per participant, one for each round ($k_{Exp1} = 16$; $k_{Exp2} = k_{Exp3} = 8$), we show the median estimate per (simulated) participant. The dashed line shows a linear regression on the data, with the rank correlation (Kendall's tau) and p-value shown above. For readability, colours represent the bivariate kernel density estimate, with red indicating higher density. The axis limits are chosen based on 1.5× the IQR for the larger of the two values (generating or recovered parameter estimates). Thus, some outliers are omitted from these plots (2.3% in Exp. 1, 1.7% in Exp. 2, and 5.2% in Exp. 3) but all datapoints are used to calculate the rank correlations.

Model for Experiments 2 and 3, in all cases using the UCB sampling strategy. We report the results in Supplementary Figure B.2, with the generating parameter estimate on the x-axis and the recovered parameter estimate on the y-axis. We report rank-correlation using Kendall's tau (r_τ), which should not be confused with the temperature parameter τ of the softmax function. Additionally, we calculate the Bayes Factor (BF_τ) to quantify the evidence for the presence of a positive correlation using non-informative, shifted, and scaled beta-priors as recommended by (Wagenmakers, Verhagen, & Ly, 2016).

For Experiment 1, the rank-correlation between the generating and the recovered length-scale λ is $r_\tau = .66$, $p < .001$, $BF_\tau > 100$, the correlation between the generating and the recovered exploration factor β is $r_\tau = .30$, $p < .001$, $BF_\tau > 100$, and the correlation between the generating and the recovered softmax temperature parameter τ is $r_\tau = .54$, $p < .001$, $BF_\tau > 100$. For Experiment 2, the correlation between the generating and the recovered λ is $r_\tau = .77$, $p < .001$, $BF_\tau > 100$, for β the correlation is $r_\tau = .59$, $p < .001$, $BF_\tau > 100$, and for τ the correlation is $r =_\tau .61$, $p < .001$, $BF_\tau > 100$. For Experiment 3, the correlation between the generating and the recovered λ is $r_\tau = .70$, $p < .001$, $BF_\tau > 100$, for β the correlation is $r_\tau = .76$, $p < .001$, $BF_\tau > 100$, and for τ the correlation is $r = .79$, $p < .001$, $BF_\tau > 100$.

These results show that the rank-correlation between the generating and the recovered parameters is very high for all experiments and for all parameters. Thus, we have strong evidence to support the claim that the reported parameter estimates of the Function Learning Model (Table B.3) are reliable, and therefore interpretable. Importantly, we find that estimates for β (exploration bonus) and τ (softmax temperature) are indeed separately identifiable, providing evidence for the existence of a *directed* exploration bonus Wilson et al. (2014), as a separate phenomena from noisy, undirected exploration (Daw et al., 2006) in our data.

B.3 Mismatched generalization

B.3.1 Generalized mismatch

A mismatch is defined as estimating a different level of spatial correlations (captured by the per participant λ-estimates) than the ground truth in the environment. In Chapter 3 (Fig. 3.4), we report a generalized Bayesian optimization simulation where we simulate every possible combination between $\lambda_0 = \{0.1, 0.2, \cdots, 1\}$ and $\lambda_1 = \{0.1, 0.2, \cdots, 1\}$, leading to 100 different combinations of student-teacher scenarios. For each of these combinations, we sample a continuous bivariate target function from a GP parameterized by λ_0 and then use the Function Learning-UCB Model parameterized by λ_1 to search for rewards. The exploration parameter β was set to 0.5 to resemble participant behaviour (Table B.3). The input space was continuous

between 0 and 1, i.e., any number between 0 and 1 could be chosen and GP-UCB was optimized (sometimes called the inner-optimization loop) per step using NLOPT (Johnson, 2014) for non-linear optimization. It should be noted that instead of using a softmax choice rule, the optimization method uses an argmax rule, since the former is not defined for continuous input spaces. Additionally, since the interpretation of λ is always relative to the input range, a length-scale of $\lambda = 1$ along the unit input range would be equivalent to $\lambda = 10$ in the $x, y = [0, 10]$ input range of Experiments 2 and 3. Thus, this simulation represents a broad set of potential mismatch alignments, while the use of continuous inputs extends the scope of the task to an infinite state space.

B.3.2 Experiments 1 and 2

In both Experiments 1 and 2, we found that participant λ-estimates were systematically lower than the true value ($\lambda_{Rough} = 1$ and $\lambda_{Smooth} = 2$), which can be interpreted as a tendency to undergeneralize compared to the spatial correlation between rewards. In order to test how this tendency to undergeneralize (i.e., underestimate λ) influences task performance, we conducted two additional sets of simulations using the exact experimental design for Experiments 1 and 2 (Fig. B.3a-b). These simulations used different combinations of λ values in a *teacher* kernel (x-axis) to generate environments and in a *student* kernel (y-axis), to simulate human search behaviour with the Function Learning Model.

Both teacher and student kernels were always RBF kernels, where the teacher kernel (used to generate environments) was parameterized with a length-scale λ_0 and the student kernel (used to simulate search behaviour) with a length-scale λ_1. For situations in which $\lambda_0 \neq \lambda_1$, the assumptions of the student can be seen as mismatched with the environment. The student *overgeneralizes* when $\lambda_1 > \lambda_0$ (Fig. B.3a-b above the dotted line), and *undergeneralizes* when $\lambda_1 > \lambda_0$ (Fig. B.3a-b below the dotted line), as was captured by our behavioural data. We simulated each possible combination of $\lambda_0 = \{0.1, 0.2, \cdots, 3\}$ and $\lambda_1 = \{0.1, 0.2, \cdots, 3\}$, leading to 900 different combinations of student-teacher scenarios. For each of these combinations, we sampled a target function from a GP parameterized by λ_0 and then used the Function Learning-UCB Model parameterized by λ_1 to search for rewards using the median parameter estimates for β and τ from the matching experiment (see Table B.3).

Figures B.3a-b show the results of the Experiment 1 and Experiment 2 simulations, where the colour of each tile shows the median reward obtained at the indicated trial number, for each of the 100 replications using the specified teacher-student scenario. The first simulation assessed mismatch in the univariate setting of Experiment 1 (Fig. B.3a), using the median participant estimates of both the softmax temperature parameter $\tau = 0.01$ and the exploration parameter $\beta = 0.50$ and simulating 100 replications for every combination between $\lambda_0 = \{0.1, 0.2, \cdots, 3\}$

and $\lambda_1 = \{0.1, 0.2, \cdots, 3\}$. This simulation showed that it can be beneficial to undergeneralize (Fig. B.3a, area below the dotted line), in particular during the first five trials. Repeating the same simulations for the bivariate setting of Experiment 2 (using the median participant estimates $\tau = 0.02$ and $\beta = 0.47$), we found that undergeneralization can also be beneficial in a more complex two-dimensional environment (Fig. B.3b), at least in the early phases of learning. In general, assumptions about the level of correlations in the environment (i.e., extent of generalization λ) only influence rewards in the short term, and can disappear over time once each option has been sufficiently sampled Srivastava et al. (2015).

B.3.3 Experiment 3

Given the robust tendency to undergeneralize in Experiments 1 and 2 (where there was a true underlying level of spatial correlation), we ran one last simulation to examine how adaptive participant λ estimates were in the real-world datasets used in Experiment 3, compared to other possible λ values. Figure B.3c shows the performance of different student λ values in the range $\{0.1, 0.2, \cdots, 3\}$ simulated over 10,000 replications sampled (with replacement) from the set of 20 natural environments. Red lines show performance in terms of average cumulative reward (Accumulation criterion) and blue lines show performance in terms of maximum reward (Maximization criterion). Vertical dashed lines indicate the interquartile range of participant λ estimates. As student λ values increase, performance by both metrics typically peaks within the range of human λ estimates, with performance largely staying constant or decreasing for larger levels of λ (with the exception of average reward at $t = 40$). Thus, we find that the extent of generalization observed in participants is generally adaptive to the real-world environments they encountered. It should also be noted that higher levels of generalization beyond what we observed in participant data have only marginal benefits, yet could potentially come with additional computational costs (depending on how it is implemented). Recall that a λ of 1 corresponds to assuming the correlation of rewards effectively decays to 0 for options with a distance greater than 3. If we assume a computational implementation where information about uncorrelated options is disregarded (e.g., in a sparse GP (Herbrich, Lawrence, & Seeger, 2003)), then the range of participant λ estimates could suggest a tendency towards lower complexity and memory requirements, while sacrificing only marginal benefits in terms of either average cumulative reward or maximum reward.

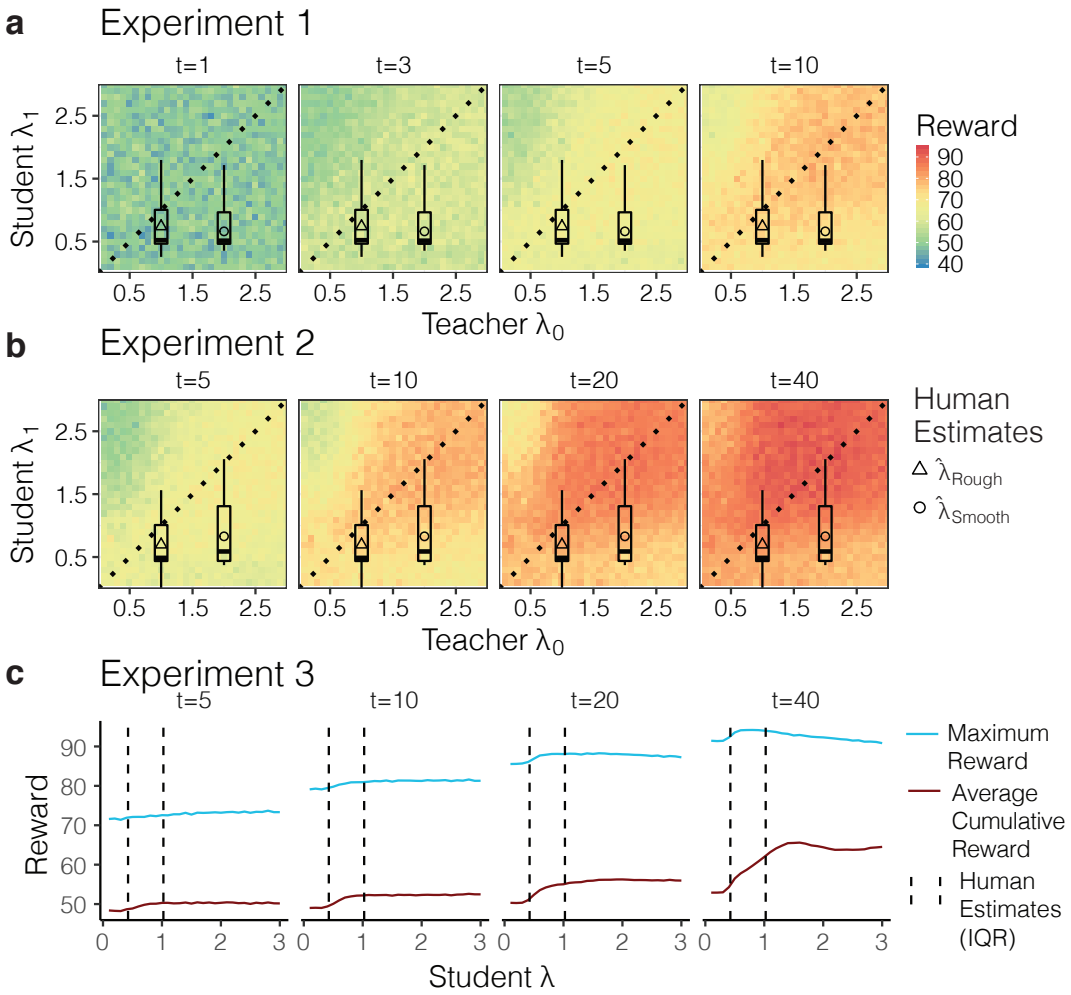

Fig. B.3 Mismatched length-scale (λ) simulation results. **a-b**) The teacher length-scale λ_0 is on the x-axis, the student length-scale λ_1 is on the y-axis, and each panel represents a different trial t. The teacher λ_0 values were used to generate environments, while the student λ_1 values were used to parameterize the Function Learning-UCB Model to simulate search performance. The dotted lines show where $\lambda_0 = \lambda_1$ and mark the difference between undergeneralization and overgeneralization, with points below the line indicating undergeneralization. Each tile of the heat-map indicates the median reward obtained for that particular λ_0-λ_1-combination, aggregated over 100 replications. Triangles and circles indicate mean participant λ estimates from Rough and Smooth conditions, with boxplots showing the interquartile range, the median (line), and 1.5x IQR (whiskers). **c**) Simulations with student λ values in the range $[0,3]$ over 10,000 samples (sampled with replacement) from the set of 20 different natural environments. Red lines show average cumulative reward and blue lines show the maximum reward. Vertical dashed lines show the interquartile range of participant λ estimates.

B.4 Natural environments

The environments used in Experiment 3 were compiled from various agricultural datasets (Table B.1), where payoffs correspond to normalized crop yield (by weight), and the rows and columns of the 11x11 grid correspond to the rows and columns of a field. Because agricultural data is naturally discretized into a grid, we did not need to interpolate or transform the data in

any way (so as not to introduce any additional assumptions), except for the normalization of payoffs in the range $[0, 100]$, where 0 corresponds to the lowest yield and 100 corresponds to the largest yield. Note that as in the other experiments, Gaussian noise $\sim \mathcal{N}(0,1)$ was added to each observed payoff in the experiment.

In selecting datasets, we used three inclusion criteria. Firstly, the datasets needed to be at least as large as our 11x11 grid. If the dataset was larger, we randomly sampled a 11x11 subsection from the data. Secondly, to avoid datasets where payoffs were highly skewed (e.g., with the majority of payoffs around 0 or around 100), we only included datasets where the median payoff was in the range $[25, 75]$. Lastly, we required that the spatial autocorrelation of each environment (computed using Moran's I) be positive:

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{B.1}$$

where N is the total number of samples (i.e., each of the 121 sections of land in a 11x11 grid), x_i is the normalized yield (i.e., payoff) for option i, $\bar{x}$ is the mean payoff over all samples, and W is the spatial weights matrix where $w_{ij} = 1$ if i and j are the same or neighbouring samples and $w_{ij} = 0$ otherwise. Moran's I ranges between $[-1, 1]$ where intuitively $I = -1$ would resemble a checkerboard pattern (with black and white tiles reflecting the highest and lowest values in the payoff spectrum), indicating maximum difference between neighbouring samples. On the other hand, $I \to 1$ would reflect a linear step function, with maximally high payoffs on one side of the environment and maximally low payoffs on the other side. We included all environments where $I > 0$, indicating that there exists some level of positive spatial correlation that could be used by participants to guide search.

Although the structure of rewards in real-world data can sometimes be distributed differently and in particular more discretely (for example, imagine a bitmap or other structural patterns such as a checkerboard or a crop circle), we believe that our environment inclusion criteria allow us to appropriately model generalization using our pool of models, while at the same time extending the scope to more complex and challenging natural structures.

B.5 Additional behavioural analyses

B.5.1 Learning over trials and rounds

We assessed whether participants improved more strongly over trials or over rounds (Fig. B.4). If they improved more over trials, this means that they are indeed finding better and better options, whereas if they are improving over rounds, this would also suggest some kind of meta-

Table B.1 Agricultural datasets used in Experiment 3

Dataset Name	Spatial Autocorrelation (Moran's I)	Crop	Source
batchelor.lemon.uniformity	0.053	Lemon	Batchelor and Reed (1918)
batchelor.navel1.uniformity	0.028	Navel Orange	Batchelor and Reed (1918)
batchelor.valencia.uniformity	0.098	Valencia Orange	Batchelor and Reed (1918)
draper.safflower.uniformity	0.075	Safflower	Draper (1959)
goulden.barley.uniformity	0.036	Barley	Goulden (1939)
iyer.wheat.uniformity	0.047	Wheat	Krishna Iyer (1942)
kalamkar.wheat.uniformity	0.004	Wheat (Yeoman II)	Kalamkar (1932)
khin.rice.uniformity	0.011	Rice	Khin (2016)
kristensen.barley.uniformity	0.146	Barley	Kristensen (1925)
montgomery.wheat.uniformity	0.243	Wheat (Winter)	Montgomery (1912)
moore.polebean.uniformity	0.119	Blue Lake Pole Beans	Moore and Darroch (1956)
moore.bushbean.uniformity	0.028	Bush Beans	Moore and Darroch (1956)
moore.sweetcorn.uniformity	0.039	Sweet Corn	Moore and Darroch (1956)
moore.carrots.uniformity	0.030	Carrots	Moore and Darroch (1956)
moore.springcauliflower.uniformity	0.013	Spring Cauliflower	Moore and Darroch (1956)
nonnecke.corn.uniformity	0.117	Sweet Corn	Nonnecke (1959)
odland.soybean.uniformity	0.105	Soybean	Odland and Garber (1928)
odland.soyhay.uniformity	0.069	Soyhay	Odland and Garber (1928)
polson.safflower.uniformity	0.059	Safflower	Polson (1964)
stephens.sorghum.uniformity	0.043	Sorghum	Stephens and Vinall (1928)

learning as they would get better at the task the more rounds they have performed previously. To test this, we fit a linear regression to every participant's outcome individually, either only with trials or only with rounds as the independent variable. Afterwards, we extract the mean standardized slopes for each participant including their standard errors. Notice that these estimates are based on a linear regression, whereas learning curves are probably non-linear. Thus, this method might underestimate the true underlying effect of learning over time.

Results (from one-sample t-tests with $\mu_0 = 0$) show that participants' scores improve significantly over trials for Experiment 1 ($t(80) = 5.57$, $p < .001$, $d = 0.6$, 95% CI $(0.2, 1.1)$, $BF > 100$), Experiment 2 ($t(79) = 2.78$, $p < .001$, $d = 0.31$, 95% CI $(-0.1, 0.8)$, $BF = 4.4$), and Experiment 3 ($t(79) = 5.91$, $p < .001$, $d = 0.7$, 95% CI $(0.2, 1.1)$, $BF > 100$). Over successive rounds, there was a negative influence on performance in Experiment 1 ($t(80) = -2.78$, $p = .007$, $d = -0.3$, 95% CI $(-0.7, 0.1)$, $BF = 4.3$), no difference in Experiment 2 ($t(79) = 0.21$, $p = .834$, $d = 0.02$, 95% CI $(-0.4, 0.5)$, $BF = 0.1$), and a minor positive influence in Experiment 3 ($t(79) = 2.16$, $p = .034$, $d = 0.2$, 95% CI $(-0.2, 0.7)$, $BF = 1.1$). Overall, participants robustly improved over trials in all experiments, with the largest effect sizes found in Experiments 1 and 3. There was no improvement over rounds in all of the experiments, suggesting that the four fully revealed example environments presented prior to the start of the task was sufficient for familiarizing participants with the task.

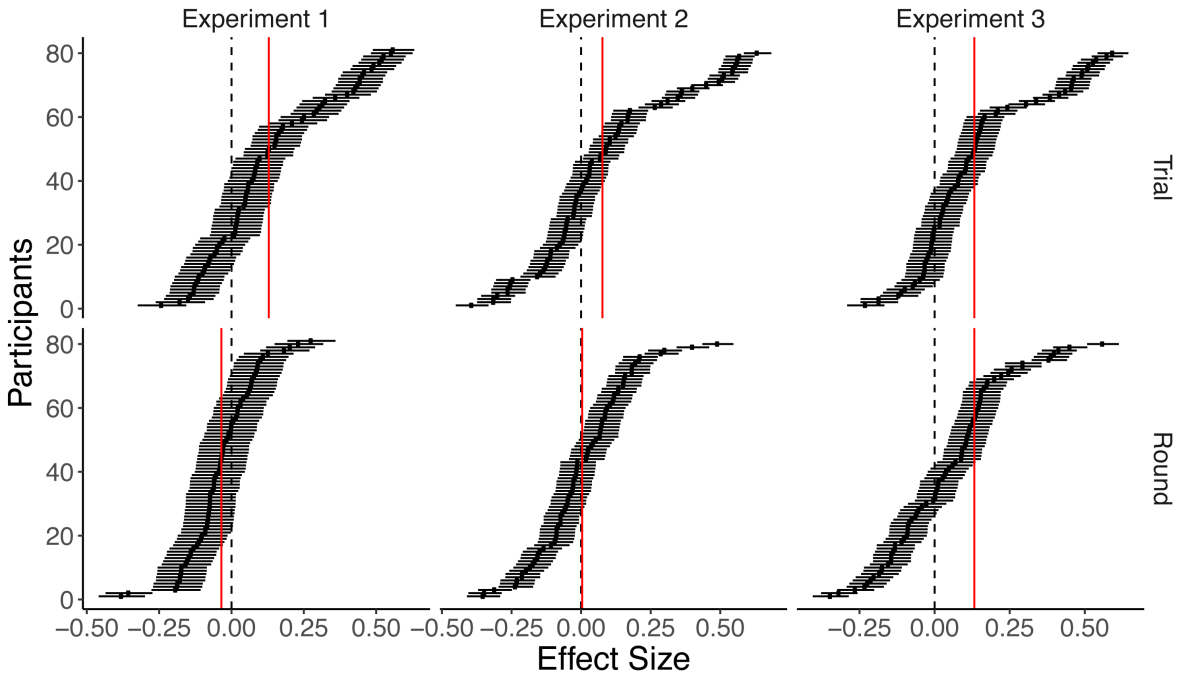

Fig. B.4 Learning over trials and rounds. Average correlational effect size of trial and round on score per participant as assessed by a standardized linear regression. Participants are ordered by effect size in decreasing order. Dashed lines indicate no effect. Red lines indicate average effect size.

B.5.2 Experimental conditions and model characteristics

To further assess how the experimental conditions influenced the model's behaviour, we performed Bayesian linear regressions of the experimental conditions onto the models' predictive accuracy and parameter estimates. To do so, we assumed a Gaussian prior on the coefficients, and an inverse Gamma prior on the conditional error variance, while inference was performed via Gibbs sampling. The results of these regressions are shown in Table B.2. Whereas the smoothness of the underlying environments (in Experiments 1 and 2) had no effect on the model's predictive accuracy and almost no effect on parameter estimates (apart from a small effect on directed exploration in Experiment 1), participants in the Accumulation payoff condition showed decreased levels of directed exploration (as captured by β) in Experiment 1 and Experiment 3, and decreased levels of random exploration in Experiment 3. Thus, our model seems to capture meaningful differences between the two reward conditions in these two experiments.

Table B.2 Bayesian linear regression of experimental conditions on model performance and parameter estimates (Experiments 1-3)

	Predictive Accuracy R^2	Generalization λ	Exploration Bonus β	Temperature τ
Experiment 1				
Intercept	**0.23 (0.18, 0.28)**	**0.71 (0.59, 0.84)**	**0.40 (0.33, 0.47)**	**0.02 (0.01, 0.02)**
Smooth	0.02 (-0.03, 0.09)	-0.07 (-0.22, 0.09)	**0.09 (0.01, 0.18)**	0.00 (-0.01, 0.01)
Accumulator	**0.12 (0.05, 0.18)**	0.03 (-0.13, 0.18)	**-0.10 (-0.19, -0.02)**	0.00 (-0.01, 0.01)
Experiment 2				
Intercept	**0.33 (0.28, 0.37)**	**0.76 (0.69, 0.82)**	**0.50 (0.47, 0.53)**	**0.09 (0.08, 0.10)**
Smooth	0.03 (-0.02, 0.08)	0.04 (-0.03, 0.06)	0.01 (-0.03, 0.04)	0.00 (-0.01, 0.01)
Accumulator	**0.07 (0.01, 0.12)**	-0.01 (-0.08, 0.06)	0.00 (-0.04, 0.02)	-0.01 (0.00, 0.01)
Experiment 3				
Intercept	**0.28 (0.24, 0.33)**	**0.64 (0.60, 0.69)**	**0.56 (0.49, 0.63)**	**0.11 (0.10, 0.12)**
Accumulator	**0.10 (0.03, 0.16)**	0.06 (-0.01, 0.12)	**-0.15 (-0.24, -0.05)**	**-0.03 (-0.04, -0.01)**

Note: We use the Function Learning model for Experiment 1 and the localized Function Learning* model for Experiment 2 and Experiment 3. Columns indicate dependent variable, whereas rows shows independent variables' regression coefficients including 95% posterior credible sets in brackets. Boldface indicates estimates whose credible sets do not overlap with 0.

B.6 Individual learning curves

To better understand why the aggregated participant learning curves sometimes decrease in average reward over time, whereas the simulated model curves tend not to (Fig. 3.3b), we present individual participant learning curves in Fig, B.5. Here, we separate the behavioural data by horizon (colour), payoff condition (rows), and environment (columns), where each line represents a single participant. We report performance in terms of both average reward (top section: Accumulation goal) and maximum reward (bottom section: Maximization goal).

The individual learning curves reveal two main causes for the decrease in reward over time when aggregating over conditions and participants. Firstly, looking at the learning curves for participants assigned to the Accumulation condition (Fig. B.5 top row), we see that roughly half of participants in the long search horizon (blue lines) show a decreasing trend at the midway point of the round. However, the other half of participants continue to gain increasingly higher rewards, more like the simulated learning curves of the Function Learning model in Figure 3b. This may be a by-product of the alternating search horizon manipulation, since the curves typically tend to decrease near the trial where a short horizon round would have ended, but also a tendency towards over-exploration that more closely resembles the Maximization goal.

Secondly, in aggregating over conditions and participants, the performance of the Accumulation and Maximization participants are averaged together. Whereas many Accumulation

payoff condition participants display more positively increasing average reward, these data points are washed out by the Maximization payoff condition participants who tend to have flatter average reward curves in pursuit of the global optimization goal.

Lastly, one additional insight from the individual learning curves comes from the flat-lined maximum reward lines (Fig. B.5, bottom section). Found more often in Accumulation participants, these flat lines represent participants who have reached a satisfactory payoff and cease additional exploration in order to exploit it. This is yet another behavioural signature of the payoff manipulations.

B.7 Extended model comparison

We report the full model comparison of 27 models, of which 12 (i.e., four learning models and three sampling strategies) are included in Chapter 3. We use different *Models of Learning* (i.e., Function Learning and Option Learning), which combined with a *Sampling Strategy* can make predictions about where a participant will search, given the history of previous observations. We also include comparisons to *Simple Heuristic Strategies* (Gigerenzer et al., 1999), which make predictions about search decisions without maintaining a representation of the world (i.e., without a learning model). Table B.3 shows the predictive accuracy, the number of participants best described, the protected probability of exceedance (Stephan et al., 2009) and the median parameter estimates of each model. Figure B.6 shows a more detailed assessment of predictive accuracy and model performance, with participants separated by payoff condition and environment type.

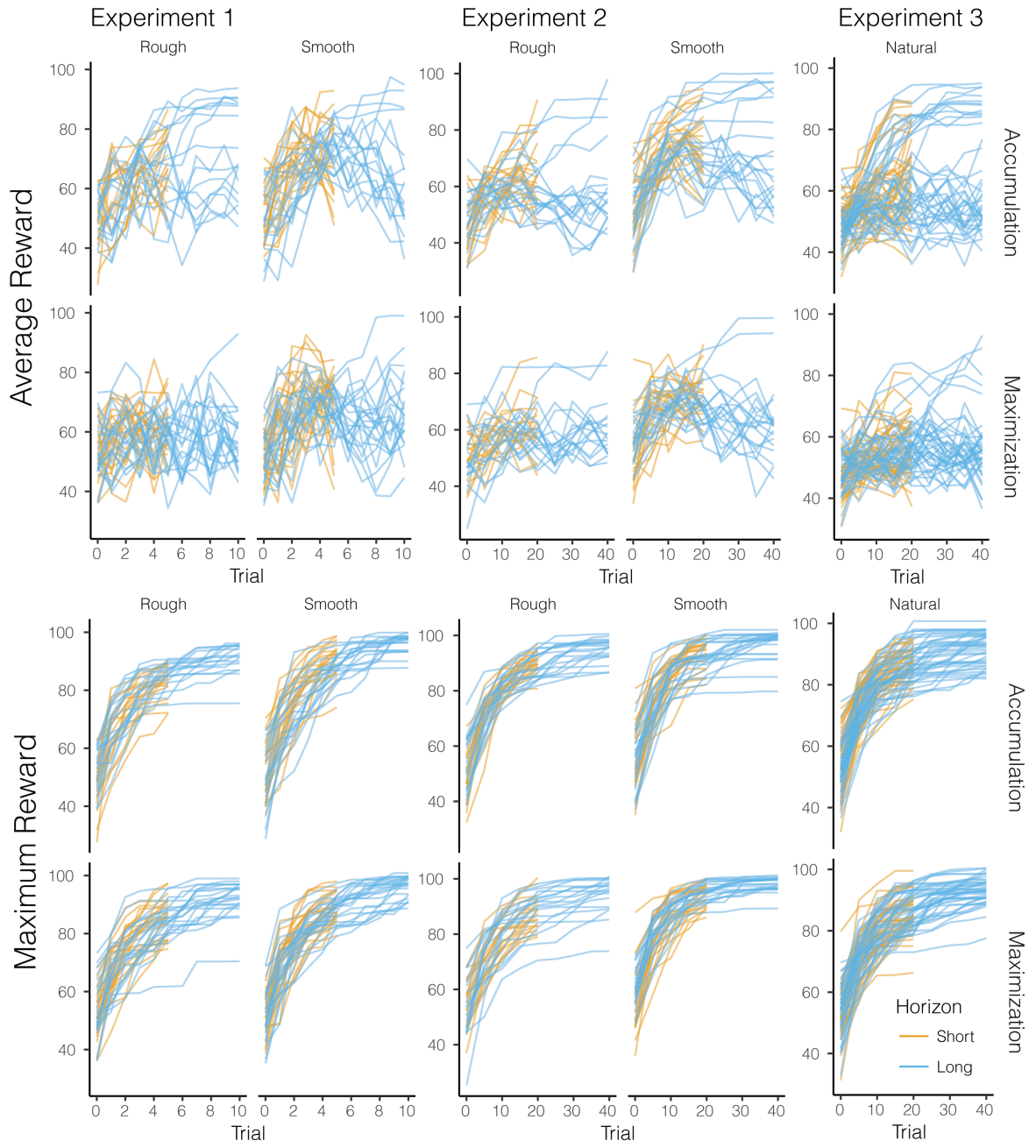

Fig. B.5 Individual participant learning curves. Each line represents a single participant, separated by search horizon (colour), by payoff condition (rows), and environment (columns). The top section shows performance in terms of average reward, while the bottom section shows performance in terms of maximum reward.

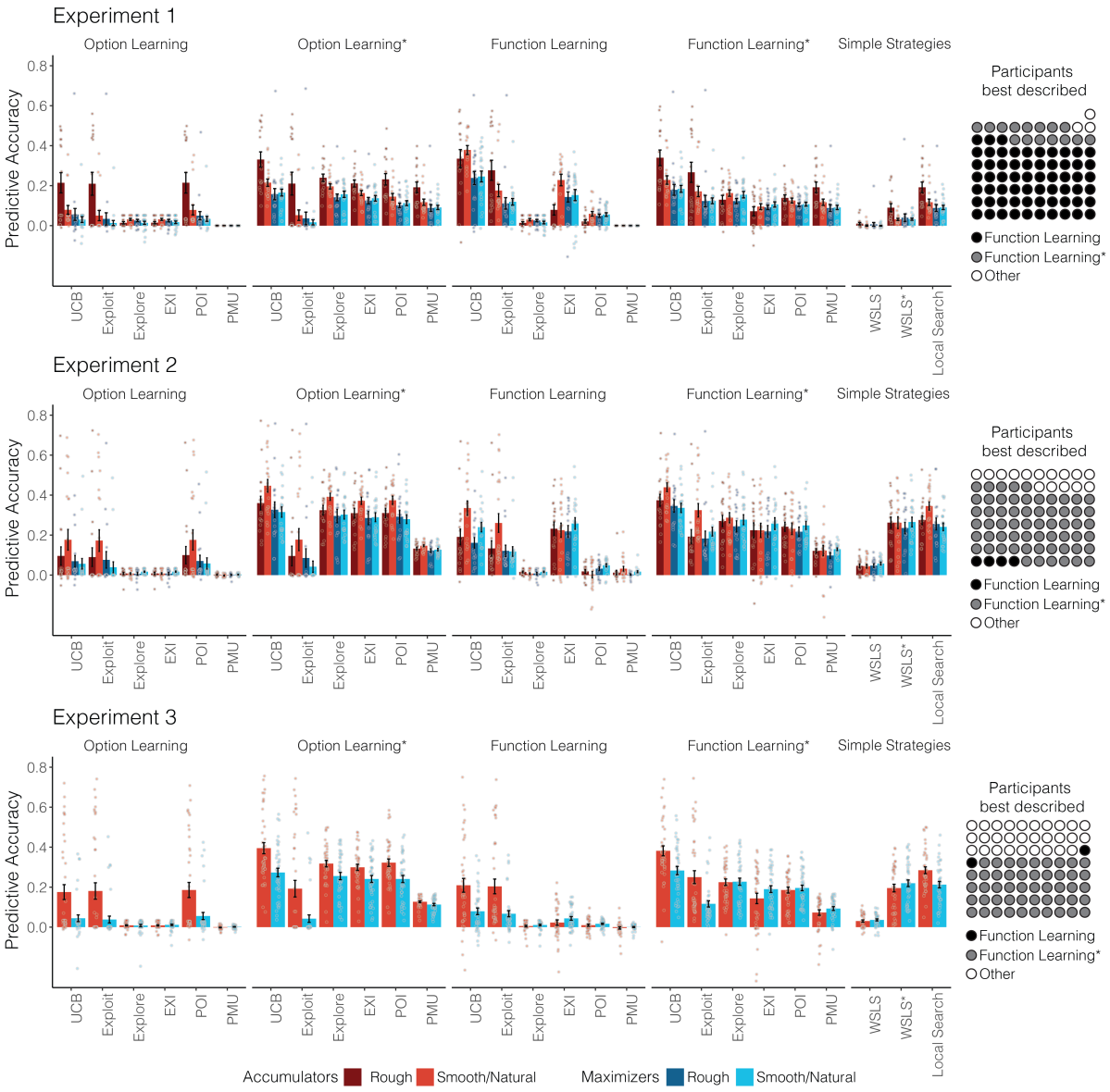

Fig. B.6 Extended model comparison of all 27 models for Experiments 1-3. The learning model is indicated above (or lack of in the case of simple heuristic strategies), and sampling strategy are along the x-axis. Bars indicate predictive accuracy (group mean) along with standard error, and are separated by payoff condition (colour) and environment type (darkness), with individual participants overlaid as dots. Icon arrays (right) show the number participants best described (out of the full 27 models) and are aggregated over payoff conditions, environment types, and sampling strategy. Table B.3 provides more detail about the number of participants best described by each model as well as the protected probability of exceedance.

Table B.3 Extended modelling results (Experiments 1-3)

Model	Experiment 1							Experiment 2							Experiment 3						
	Model Comparison			Parameter Estimates				Model Comparison			Parameter Estimates				Model Comparison			Parameter Estimates			
	R^2	Best Described	pxp	Length Scale λ	Exploration Bonus β	Error Variance $\sqrt{\theta_\varepsilon^2}$	Softmax Temperature τ	R^2	Best Described	pxp	Length Scale λ	Exploration Bonus β	Error Variance $\sqrt{\theta_\varepsilon^2}$	Softmax Temperature τ	R^2	Best Described	pxp	Length Scale λ	Exploration Bonus β	Error Variance $\sqrt{\theta_\varepsilon^2}$	Softmax Temperature τ
Option Learning																					
Upper Confidence Bound	0.09	0	0	–	3.51	0.94	0.03	0.1	0	0	–	0.97	1.96	0.02	0.11	0	0	–	0.85	2.08	0.03
Pure Exploitation	0.07	1	0	–	–	54.6	54.6	0.1	0	0	–	–	148.41	148.41	0.11	0	0	–	–	148.41	148.41
Pure Exploration	0.02	0	0	–	–	0.32	0.02	0.01	0	0	–	–	15.9	0.03	0.01	0	0	–	–	5.42	0.05
Expected Improvement	0.02	0	0	–	–	0.37	0.01	0.01	0	0	–	–	1.56	0.02	0.01	0	0	–	–	0.73	0.21
Probability of Improvement	0.09	0	0	–	–	0.01	0.15	0.1	0	0	–	–	0.01	0.11	0.12	0	0	–	–	0.02	0.11
Probability of Maximum Utility	0.00	0	0	–	–	0.69	0.69	0	0	0	–	–	0.54	0.01	0.00	0	0	–	–	0.65	0.01
Option Learning*																					
Upper Confidence Bound	0.21	1	0	–	44.7	0.01	28.07	0.36	12	0	–	44.08	0.07	15.79	0.33	21	0.05	–	42.25	0.01	16.25
Pure Exploitation	0.07	1	0	–	–	54.6	0.01	0.1	0	0	–	–	148.41	148.41	0.12	0	0	–	–	148.41	148.41
Pure Exploration	0.18	0	0	–	–	0.01	0.71	0.33	3	0	–	–	0.58	0.43	0.29	5	0	–	–	0.45	0.46
Expected Improvement	0.16	0	0	–	–	0.01	0.27	0.32	0	0	–	–	0.63	0.14	0.27	0	0	–	–	0.4	0.16
Probability of Improvement	0.14	0	0	–	–	0.01	0.19	0.32	0	0	–	–	0.01	0.09	0.28	0	0	–	–	0.01	0.1
Probability of Maximum Utility	0.12	0	0	–	–	0.67	0.46	0.13	0	0	–	–	0.36	0.01	0.00	0	0	–	–	0.54	0.01
Function Learning																					
Upper Confidence Bound	**0.29**	**48**	**1**	**0.5**	**0.51**	–	**0.01**	0.24	4	0	0.54	0.47	–	0.02	0.14	2	0	0.52	0.4	–	0.02
Pure Exploitation	0.16	6	0	1.94	–	–	0.15	0.16	0	0	1.55	–	–	0.11	0.14	0	0	1.16	–	–	0.13
Pure Exploration	0.02	0	0	0.11	–	–	0.03	0.01	0	0	0.17	–	–	0.55	0.01	0	0	0.17	–	–	0.55
Expected Improvement	0.15	9	0	0.56	–	–	0.01	0.23	0	0	0.67	–	–	0.05	0.03	0	0	0.49	–	–	0.01
Probability of Improvement	0.05	0	0	3.43	–	–	0.18	0.02	0	0	0.87	–	–	0.09	0.01	0	0	0.78	–	–	0.14
Probability of Maximum Utility	0.00	0	0	0.69	–	–	7.17	0.02	0	0	0.49	–	–	0.01	0.00	0	0	0.42	–	–	0.01
Function Learning*																					
Upper Confidence Bound	0.23	10	0	0.96	0.54	–	0.16	**0.38**	**60**	**1**	**0.76**	**0.49**	–	**0.09**	**0.33**	**47**	**0.95**	**0.67**	**0.52**	–	**0.1**
Pure Exploitation	0.16	1	0	7.13	–	–	0.12	0.23	0	0	14.4	–	–	0.06	0.18	0	0	10.87	–	–	0.06
Pure Exploration	0.14	3	0	0.08	–	–	0.32	0.27	0	0	0.17	–	–	.19	0.23	2	0	0.17	–	–	0.2
Expected Improvement	0.09	1	0	0.71	–	–	0.11	0.23	1	0	0.67	–	–	0.05	0.17	0	0	0.64	–	–	0.06
Probability of Improvement	0.12	0	0	7.14	–	–	0.2	0.24	0	0	0.84	–	–	0.09	0.19	0	0	0.72	–	–	0.1
Probability of Maximum Utility	0.12	0	0	0.67	–	–	0.46	0.12	0	0	0.46	–	–	0.01	0.08	0	0	0.27	–	–	0.01
Simple Heuristics																					
Win-Stay Lose-Sample	0.00	0	0	–	–	–	3.72	0.05	0	0	–	–	–	0.32	0.03	0	0	–	–	–	0.39
Win-Stay Lose-Sample*	0.05	0	0	–	–	–	0.73	0.26	0	0	–	–	–	0.22	0.21	1	0	–	–	–	0.24
Local Search	0.12	0	0	–	–	–	0.46	0.28	0	0	–	–	–	0.22	0.25	2	0	–	–	–	0.23

Note: R^2 indicates out-of-sample predictive accuracy, "Best Described" indicates the number of participants in each experiment that were best described by a model, and pxp is the protected probability of exceedance Rigoux et al. (2014); Stephan et al. (2009) using the model's out-of-sample log-evidence. Parameter estimates are the median over all participants. There were 81 participants in Experiment 1, 80 participants in Experiment 2, and 80 participants in Experiment 3. The best performing model for each experiment is highlighted in boldface. Asterisks (*) indicate a localized variant of a model.

Appendix C

Chapter 4 supplementary materials

C.1 Other forms of random exploration

A softmax function with a temperature parameter τ is only one way to define random exploration. Another approach towards assessing random exploration is so-called ε-greedy exploration. Given k number of arms (64 in Experiment 4), ε-greedy exploration chooses

$$p(\mathbf{x}) = \begin{cases} 1-\varepsilon, & \text{if } \arg\max \mathrm{UCB}(\mathbf{x}) \\ \varepsilon/(k-1), & \text{otherwise} \end{cases} \tag{C.1}$$

where ε is a free parameter. We test the ε-greedy method of exploration by using it instead of a softmax function in combination with the GP regression model and a UCB-sampling strategy. The results of this comparison show that the ε-greedy exploration model was systematically worse at predicting behavior than the softmax model reported in the main text (mean predictive accuracy: $R^2 = 0.21, t(159) = 6.67, p < .001, d = 0.53, BF > 100$). Additionally, the softmax model also had better predictive accuracy than the ε-greedy exploration model for adults (mean predictive accuracy: $R^2 = 0.26, t(49) = 9.29$ $p < .001$, $d = 1.31$, $BF > 100$), and for older children (mean predictive accuracy: $R^2 = 0.21, t(54) = 3.60, p < .001, d = 0.49, BF = 39$), but not for younger children (mean predictive accuracy: $R^2 = 0.17$, $t(54) = 0.33$, $p = .74$, $d = 0.04$, $BF = 0.2$).

Next, we looked for age-related differences in the parameter estimates of the ε-greedy model[†], specifically the directed exploration parameter β and the alternative random exploration parameter ε. As in the softmax-parameterized models, we find larger λ-estimates for adults than for older children (0.99 vs. 0.24, $U = 1975$, $r_\tau = 0.31$, $p < .001$, $BF > 100$), whereas

[†]Note that interpreting estimates of inferior computational models can be problematic and should only be done with caution.

the two children groups do not differ in their λ-estimates (0.31 vs. 0.24, $U = 1299$, $r_\tau = 0.10$, $p = .20$, $BF = 0.4$). Furthermore, we find more directed exploration (larger β parameters) for older children than for adults (17.30 vs. 5.38, $U = 555$, $r_\tau = 0.42$, $p < .001$, $BF > 100$), but no difference between the two groups of children (17.20 vs. 17.30, $U = 1684$,$r_\tau = 0.12$, $p = .3$, $BF = 0.3$). We also found a difference in ε-greedy exploration parameter between adults and older children (0.00012 vs. 0.00014, $U = 960$,$r_\tau = 0.21$, $p = .007$, $BF = 7$), but not between the two groups of children (0.00014 vs. 0.00016, $U = 1774$, $r_\tau = 0.12$, $p = .11$, $BF = 0.5$). Notice that the relative proportion of random exploration decisions according to the ε-parameter estimates is so small, that over the 200 choices in our task, this accounts for a difference of approximately 1 in every 250 choices. Thus, there is almost no practical difference in participants' ε-parameters. The overall age-related effect in the ε-greedy analysis was also larger for directed exploration than for ε-greedy exploration ($r_\tau = 0.40$ vs. $r_\tau = 0.25$).

Thus, there are two reasons to believe that children are driven more strongly by directed than by random exploration. Firstly, the GP-UCB model combined with a softmax formulation of random exploration predicted participants better than an ε-greedy model, and finds no age-related difference in terms of random exploration described by the temperature parameter τ. Secondly, parameter estimates of the ε-greedy model find only small and practically meaningless age-related difference in ε-exploration, but again a large age-related differences in the directed exploration parameter β.

C.2 Parameter estimates

All parameter estimates were included in the statistical analyses, although we exclude outliers larger than 5 in Figures 4.1e, C.2, and in Table C.1. For GP-UCB estimates, 1.1% of λ-estimates, for 3.4% of β-estimates, and 2.7% of τ-estimates were removed this way. The presence of these outliers motivated us to use non-parametric tests (without removing outliers) to compare the different parameters across age groups, in order to achieve more robustness. However, the significance of our test results does not change even if we use parametric t-tests, but remove outliers before performing these tests. After removing values higher than 5, we still find that adults show higher λ-estimates than older children ($t(103) = 3.88$, $p < .001$, $d = 0.76$, $BF > 100$), who do not differ in their λ-estimates from younger children ($t(108) = 1.00$, $p = .32$, $d = 0.19$, $BF = 0.3$). Moreover, the β-estimates are higher for older children than for adults ($t(103) = 4.41$, $p < .001$, $d = 0.87$, $BF > 100$), but do not differ between the two groups of children $t(108) = 1.41$, $p = .160$, $d = 0.27$, $BF = 0.5$). Crucially, the random exploration parameters τ do neither differ between older children and adults ($t(103) = .55$, $p = .58$,

$d = 0.11$, $BF = 0.2$), nor do they differ between the two groups of children ($t(108) = 1.74$, $p = .08$, $d = 0.33$, $BF = 0.8$).

Our criterion for outlier removal is less strict than other criteria, such as removing values higher than 1.5×IQR. In our data, this would remove 1.8% of λ-estimates, 5.1% of β-estimates, and 11.2% of τ-estimates. For completeness, we also reanalyzed the data after performing Tukey's procedure of outlier removal, where we find the same results. Adults still show a higher λ-estimates than older children ($t(102) = 3.87$, $p < .001$, $d = 0.76$, $BF > 100$), who do not differ from younger children ($t(108) = 1.04$, $p = .30$, $d = 0.20$, $BF = 0.3$). The β-estimates are again higher for older children than for adults ($t(100) = 4.69$, $p < .001$, $d = 0.93$, $BF > 100$), and do not differ between the two groups of children ($t(103) = 1.06$, $p = .29$, $d = 0.21$, $BF = 0.3$). Finally, estimates for τ-parameter do not differ between adults and older children ($t(103) = 1.36$, $p = .18$, $d = 0.26$, $BF = 0.5$), nor between the two groups of children ($t(108) = 0.96$, $p = .34$, $d = 0.18$, $BF = 0.3$). Taken together, our results hold no matter which method of outlier removal is applied.

C.3 Model recovery

We present model recovery results that assess whether or not our predictive model comparison procedure allows us to correctly identify the true underlying model. To assess this, we generated data based on each individual participant's mean parameter estimates (excluding outliers larger than 5 as before). More specifically, for each participant and round, we use the cross-validated parameter estimates to specify a given model, and then generate new data in the attempt to mimic participant data. We generate data using the BMT and the GP regression model. In all cases, we use the UCB sampling strategy in conjunction with the specified learning model. We then utilize the same cross-validation method as before in order to determine if we can successfully identify which model has generated the underlying data. Figure C.1 shows the cross-validated predictive performance for the simulated data.

Our predictive model comparison procedure shows that the GP model is a better predictor for data generated from the same underlying model, whereas the BMT model is only marginally (if at all) better at predicting data generated from the same underlying model. This suggests that our main model comparison results are robust to Type II errors, and provides evidence that the better predictive accuracy of the GP model on participant data is unlikely due to differences in model mimicry.

When the BMT model generates data using participant parameter estimates, the same Mean Tracker model performs better than the GP model ($t(159) = 1.42$, $p = .16$, $d = 0.11$, $BF = 4.1$) and predicts 86 out of 160 simulated participants best. Notice, however, that both models

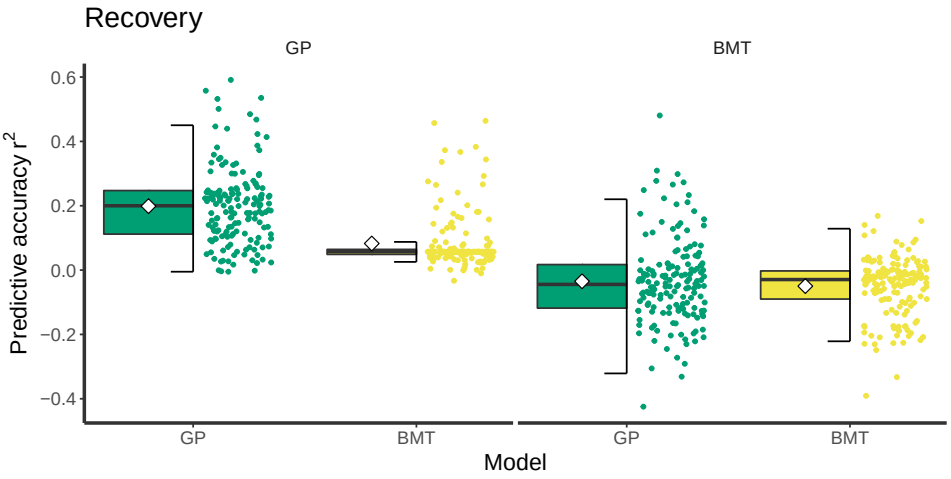

Fig. C.1 Tukey box plots of model recovery results including individual data points (points) and overall means (diamonds). **Left:** Model performance based on data generated by the GP-UCB model specified with participant parameter estimates. The matching GP-UCB model makes better predictions than a mismatched BMT-UCB model. **Right:** Model performance based on data generated by the BMT-UCB model specified with participant parameter estimates. Both GP and BMT models predict this data poorly, and perform similar to a random model.

perform poorly in this case, with both models achieving an average pseudo-r-squared of around $R^2 = -0.04$. This also shows that the BMT is not a good generative model of human-like behavior in our task.

When the GP model has generated the underlying data, the same model performs significantly better than the BMT model ($t(159) = 18.6$, $p < .001$, $d = 1.47$, $BF > 100$) and predicts 153 of the 160 simulated participants best. In general, both models perform better when the GP has generated the data with the GP achieving a predictive accuracy of $R^2 = .20$ and the BMT of achieving $R^2 = .12$. This makes our finding of the GP model as the best predictive model even stronger as—technically—the BMT model can mimic parts of its behavior. Moreover, the resulting values of predictive accuracy are similar to the ones we found using participant data. This indicates that our empirical values of predictive accuracy were as good as possible under the assumption that a GP model generated the data.

C.4 Parameter recovery

Another important question is whether the reported parameter estimates of the GP-UCB model are reliable and recoverable. We address this question by assessing the recoverability of the three underlying parameters, the length-scale λ, the directed exploration factor β, and the random exploration (temperature) parameter τ of the softmax choice rule. We use the results from the model recovery simulation described above, and correlate the empirically estimated parameters used to generate data (i.e., the estimates based on participants' data), with the

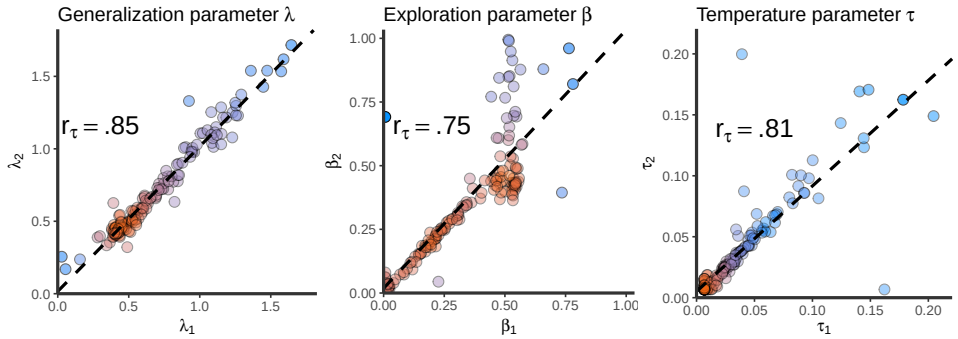

Fig. C.2 Parameter recovery results. The generating parameter estimate is on the x-axis and the recovered parameter estimate is on the y-axis. The generating parameter estimates are from the cross-validated participant parameter estimates, which were used to simulate data (see Model recovery). Recovered parameter estimates are the result of the cross-validated model comparison on the simulated data. While the cross-validation procedure yielded 8-estimates per participant, one for each round, we show the mean estimate per (simulated) participant. The dashed line shows a linear regression on the data, while Kendall's rank correlation r_τ is shown in the plot. For readability, colors represent the bivariate kernel density estimate, with red indicating higher density.

parameter estimates of the recovering model (i.e., the MLE from the cross-validation procedure on the simulated data). We assess whether the recovered parameter estimates are similar to the parameters that were used to generate the underlying data. We present parameter recovery results for the Gaussian Process regression model using the UCB sampling strategy. We report the results in Figure C.2, with the generating parameter estimate on the x-axis and the recovered parameter estimate on the y-axis.

The correlation between the generating and the recovered length-scale λ is $r_\tau = .85$, $p < .001$, the correlation between the generating and the recovered exploration factor β is $r_\tau = 0.75$, $p < .001$, and the correlation between the generating and the recovered softmax temperature parameter τ is $r_\tau = 0.81$, $p < .001$.

These results show that the correlation between the generating and the recovered parameters is very high for all parameters. Thus, we have strong evidence to support the claim that the reported parameter estimates of the GP-UCB model are recoverable, reliable, and therefore interpretable. Importantly, we find that estimates for β (exploration bonus) and τ (softmax temperature) are indeed recoverable, providing evidence for the existence of a *directed* exploration bonus, as a separate phenomena from *random* exploration in our behavioral data.

Next, we analyze whether or not the same differences between the parameter estimates that we found for the experimental data can also be found for the simulated data. Thus, we compare the recovered parameter estimates from the data generated by the estimated parameters for the different age groups. This comparison shows that the recovered data exhibits the same characteristics as the empirical data. The recovered λ-estimates for simulated data from adults was again larger than the recovered lambda estimates for older children ($U = 2021$, $r_\tau = 0.33$,

$p < .001$, $BF > 100$), whereas there was no difference between the recovered parameters for the two simulated groups of children ($U = 1800$, $p = .08$, $r_\tau = 0.14$, $BF = 1$). As in the empirical data, the recovered estimates also showed a difference between age groups in their directed exploration behavior such that the recovered β was higher for simulated older children than for simulated adults ($U = 730$, $p < .001$, $r_\tau = 0.33$, $BF > 100$), whereas there was no difference between the two simulated groups of children ($U = 1730$, $p = .19$, $r_\tau = 0.10$, $BF = 0.6$). There was no difference between the different recovered τ-parameters (max-$BF = 0.5$). Thus, our model can also reproduce similar group differences between generalization and directed exploration as found in the empirical data.

C.5 Counter-factual parameter recovery

Another explanation of the finding that children differ from adult participants in their directed exploration parameter β but not in their random exploration parameter τ could be that the softmax temperature parameter τ can sometimes track more of the random behavioral difference between participants than the directed exploration parameter β. If the random exploration parameter τ indeed tends to absorb more of the random variance in the data than the directed exploration parameter β, then perhaps one is always more likely to find differences in β rather than differences in τ. To assess this claim, we simulate data using our GP-UCB model as before but swap participants' parameter estimates of β with their estimates of τ and vice versa. Ideally, this simulation can reveal whether it is possible for our method to find differences in τ but not β in a counter-factual parameter recovery where the age groups differ in their random but not their directed exploration behavior. Thus, we generate data from the swapped GP-UCB model and then use our model fitting procedure to assess the GP-UCB-model's parameters from this generated data. The results of this simulation reveal that simulated adults do not differ from simulated older children in their estimated directed exploration parameters β ($U = 1170$, $r_\tau = 0.11$, $p = .19$, $BF = 0.7$). Furthermore, the two simulated children groups also do not differ in terms of their directed exploration parameter β ($U = 1696$,$r_\tau = 0.09$, $p = .27$, $BF = 0.4$). However, the random exploration parameter τ is estimated to be somewhat higher for simulated older children than for simulated adults ($U = 1771$, $r_\tau = 0.21$, $p = .01$, $BF = 2.5$) and shows no difference between the two simulated children groups ($U = 1631$, $p = .48$, $r_\tau = 0.06$, $BF = 0.4$). This means that the GP-UCB model can pick up on differences in random exploration as well and therefore that our findings are unlikely due to a false positive.

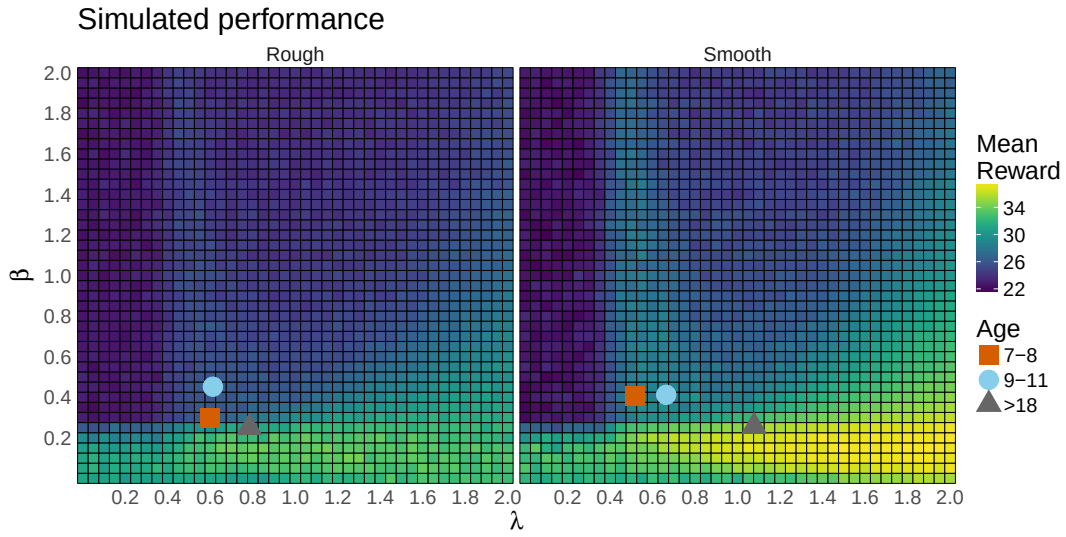

Fig. C.3 Simulated performance for different parameter values of λ (generalization) and β (directed exploration bonus) in the rough (left) and smooth (right) environment. Each tile shows the mean performance over 100 replications on each environment type. The mean participant parameter estimates (separated by age group) are overlaid.

C.6 Comparison to optimal parameter estimates

We compare the performance of participants' parameter estimates to the simulated performance of alternative parameter combinations in the GP-UCB model (Fig. C.3). For this, we simulate the GP-UCB model on both the smooth and the rough environments for the same number of trials as participants experienced, and track performance for each run. Since there were no meaningful differences for the random exploration parameter τ, we set $\tau = 0.03$ (i.e., the median over all participants) for all simulations. We vary the parameter values of the directed exploration bonus β and the generalization parameter λ to all permutations of values stemming from $[0.05, 0.1, \cdots, 2]$, leading to 1,600 differently parameterized models in total. We then run each model for 100 replications on both the smooth and the rough environments individually, always calculating mean performance over all runs. Figure C.3 shows the performance of different λ-β-combinations with participants' parameter estimates overlaid.

We extract the best parameters of this simulation by using all parameters that are not significantly different in their performance from the overall best-performing parameters using an α-level of 0.05. The best-performing parameters for the rough condition have a median generalization parameter of $\lambda = 0.95$ (range: 0.65-1.30) and a median exploration parameter of $\beta = 0.15$ (range: 0.1-0.15). The best-performing parameters for the smooth condition have a median generalization parameter of $\lambda = 1.78$ (range: 1.4-2) and a median exploration parameter of $\beta = 0.15$ (range: 0.05-0.2). Unsurprisingly, adults' parameter estimates are closer to best-performing parameters than children's parameter estimates. These simulations also

replicate earlier findings (Wu, Schulz, Speekenbrink, et al., 2018) showing that lower values of λ (i.e., undergeneralization) can lead to better performance than values of λ that are higher and closer to the true underlying λ that generated the environments.

C.7 Further behavioral analyses

C.7.1 Learning over trials and rounds

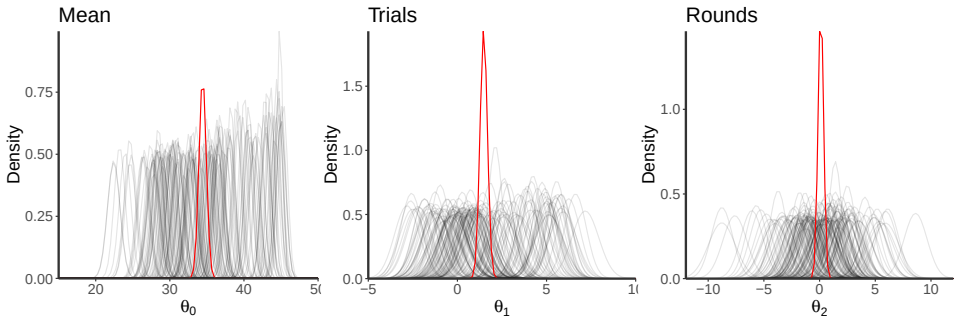

Fig. C.4 Posterior regression coefficients of mean performance and the effects of both trials and rounds onto participants' rewards. Individual lines densities correspond to participant-wise estimates, whereas red lines show hierarchical estimates.

We analyze participants learning over trials and rounds using a hierarchical Bayesian regression approach. Formally, we assume the regression weight parameters θ_x for $x \in \{0,1,2\}$ are hierarchically distributed as

$$\theta_x \sim \mathcal{N}(\mu_x, \sigma_x^2); \tag{C.2}$$

we further assume a weakly informative prior of the means and standard deviation over the regression equation, defined as:

$$\mu_x \sim \mathcal{N}(0, 100) \tag{C.3}$$

$$\sigma_x^2 \sim \text{Half-Cauchy}(0, 100) \tag{C.4}$$

We fit one hierarchical model of means and standard deviations over all participants using Hamiltonian Markov chain Monte Carlo sampling as implemented in the `PyMC`-environment. This yields hierarchical estimates for each regression coefficient overall as well as individual estimates for each participant. Doing so for a model containing an intercept (mean performance), a standardized coefficient of the effect of trials on rewards, as well as a standardized coefficient of the effect of rounds on rewards results into the posterior distributions shown in Figure C.4.

The overall posterior mean of participants' rewards is estimated to be 34.4 with a 95%-credible set of $[33.5, 35.4]$. The standardized effect of trials onto rewards is estimated to be

1.48 with a 95%-credible set of $[1.1, 1.9]$, indicating a strong effect of learning over trials. The standardized effect of rounds onto participants' rewards is estimated to be 0.11 with a 95%-credible set of $[-0.4, 0.6]$, indicating no effect. Taken together, these results indicate that participants performed well above the chance-level of 25 overall and improved their score greatly over trials. They did not, however, learn or adapt their strategies across rounds.

C.7.2 Reaction times

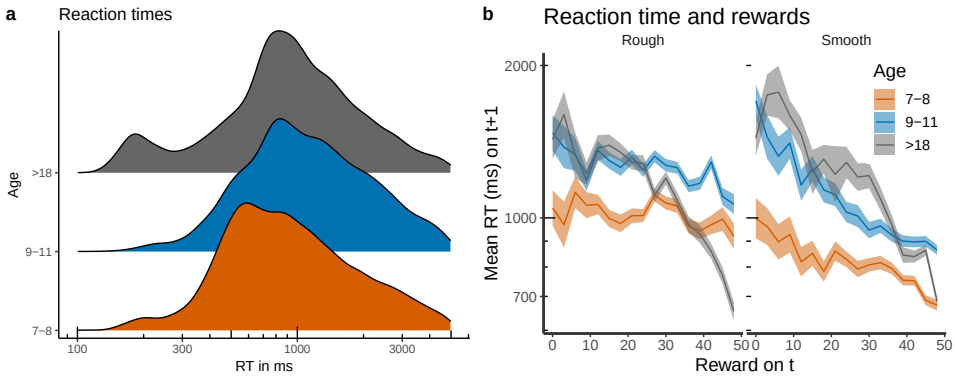

Fig. C.5 Reaction times. **a**) Reaction time (RT) in miliseconds (ms) by age group and condition, with the x-axis on log-scale. **b**) Reaction times as a function of previous reward.

We analyze participants log-reaction times as function of previous reward, age group, and condition. Reaction times are filtered to be smaller than 5000ms and larger than 100ms (3.05% removed in total). We find that reaction times are slower for the rough as compared to the smooth condition (see Fig. C.5a; $t(158) = 4.44$, $p < .001$, $d = 0.70$, $BF > 10$). Moreover, adults are somewhat faster than children aged 9-11 ($t(103) = -2.60$, $p = .01$, $d = 0.51$, $BF = 4$). The difference in reaction times between children aged 9-11 and children aged 7-8 is only small ($t(108) = 2.27$, $p = .03$, $d = 0.43$, $BF = 2$).

We also analyze how the value of the previous reward influences reaction times, i.e., whether participants slow down after a bad outcome and/or speed up after a good outcome (see Fig. C.5b). The correlation between the previous reward and reaction times is negative overall with $r = -0.18$, $t(159) = -15$, $p < .001$, $BF > 100$, indicating that larger rewards lead to faster reaction times, whereas participants might slow down after having experienced low rewards. This effect is even stronger for the smooth as compared to the rough condition ($t(158) = -4.43$, $p < .001$, $d = 0.70$, $BF > 100$). Moreover, this effect is also stronger for adults than for older children ($t(103) = -5.51$, $p < .001$, $d = 1.08$, $BF > 100$) and does not differ between the two groups of children ($t(108) = 1.93$, $p = .06$, $d = 0.37$, $BF = 1$).

Table C.1 Extended modeling results (Experiment 4).

Model	R^2	# Best	Log-loss	*pxp*	Generalization λ	Exploration β	Error Var. $\sqrt{\theta_\varepsilon^2}$	Softmax τ
Overall:								
BMT-PureExploit	0.02	10	103.9	0	–	–	2.98	0.01
BMT-PureExplore	0.01	3	102.7	0	–	–	0.58	0.08
BMT-UCB	0.05	5	99.3	0	–	0.41	1.60	0.16
GP-PureExploit	0.09	34	94.4	0	1.46	–	–	0.16
GP-PureExplore	0.02	2	102.1	0	0.18	–	–	0.56
GP-UCB	0.28	106	74.8	1	0.59	0.45	–	0.03
Age 7-8:								
BMT-PureExploit	0.00	3	103.9	0	–	–	0.53	0.01
BMT-PureExplore	0.00	3	100.8	0	–	–	1.22	0.08
BMT-UCB	0.03	3	100.5	0	–	0.47	1.20	0.11
GP-PureExploit	0.07	14	96.9	0	1.37	–	–	0.18
GP-PureExplore	0.01	0	102.7	0	0.17	–	–	0.41
GP-UCB	0.14	32	89.2	1	0.46	0.48	–	0.03
Age 9-11:								
BMT-PureExploit	0.01	0	103.9	0	–	–	2.86	0.01
BMT-PureExplore	0.01	0	101.7	0	–	–	0.65	0.08
BMT-UCB	0.05	0	99.3	0	–	0.33	1.34	0.07
GP-PureExploit	0.08	9	95.1	0	1.52	–	–	0.17
GP-PureExplore	0.01	2	101.5	0	0.17	–	–	0.38
GP-UCB	0.26	44	76.6	1	0.74	0.53	–	0.02
Adults:								
BMT-PureExploit	0.29	7	72.8	0	–	–	3.38	0.03
BMT-PureExplore	0.00	0	103.4	0	–	–	0.13	0.87
BMT-UCB	0.29	2	74.2	0	–	0.18	2.94	0.03
GP-PureExploit	0.33	11	69.1	0	1.38	–	–	0.07
GP-PureExplore	0.01	0	103.9	0	0.10	–	–	0.23
GP-UCB	0.40	30	62.7	1	0.85	0.24	–	0.03

Note: Columns indicate (from left to right) the average predictive accuracy of each model (R^2), how many participants each model best predicted (# Best), the average log-loss over all 8 rounds of predictions (Log-loss), the protected probability of exceedance (*pxp*), the GP's generalization parameter (λ), the directed exploration parameter (β) for all models involving a UCB sampling strategy, the BMT's error variance ($\sqrt{\theta_\varepsilon^2}$), and the random exploration (softmax temperature) parameter (τ). Parameter estimates are based on median over per-participant mean parameter estimates (excluding estimates larger than 5 as outliers).

References

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

Acuna, D., & Schrater, P. R. (2009). Structure learning in human sequential decision-making. In *Advances in Neural Information Processing Systems* (pp. 1–8).

Addicott, M., Pearson, J., Sweitzer, M., Barack, D., & Platt, M. (2017). A primer on foraging and the explore/exploit trade-off for psychiatry research. *Neuropsychopharmacology*, *42*, 1931–1939.

Analytis, P. P., Wu, C. M., & Gelastopoulos, A. (2019). Make-or-break: chasing risky goals or settling for safe rewards? *Cognitive Science*, *43*, e12743. doi: 10.1111/cogs.12743

Attneave, F. (1950). Dimensions of similarity. *The American Journal of Psychology*, *63*, 516–556.

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

Barkoczi, D., Analytis, P. P., & Wu, C. M. (2016). Collective search on rugged landscapes: a crossenvironmental analysis. In A. Papafragou, D. Grodner, D. Mirman, & J. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 918–923). Austin, TX: Cognitive Science Society.

Batchelor, L., & Reed, H. (1918). Relation of the variability of yields of fruit trees to the accuracy of field trials. *Journal of Agricultural Research*, *12*, 461–468.

Bauer, M., van der Wilk, M., & Rasmussen, C. E. (2016). Understanding probabilistic sparse Gaussian process approximations. In *Advances in Neural Information Processing Systems* (pp. 1533–1541).

Beals, R., Krantz, D. H., & Tversky, A. (1968). Foundations of multidimensional scaling. *Psychological Review*, *75*, 127.

Behrens, T. E., Muller, T. H., Whittington, J. C., Mark, S., Baram, A. B., Stachenfeld, K. L., & Kurth-Nelson, Z. (2018). What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron*, *100*, 490–509.

Bellman, R. (1952). On the theory of dynamic programming. *Proceedings of the National Academy of Sciences*, *38*, 716–719.

Bellman, R. (1957). *Dynamic Programming*. Princeton University Press: Princeton, NJ.

Berdahl, A., Torney, C. J., Ioannou, C. C., Faria, J. J., & Couzin, I. D. (2013). Emergent sensing of complex environments by mobile animal groups. *Science*, *339*, 574–576.

Berns, G. S., McClure, S. M., Pagnoni, G., & Montague, P. R. (2001). Predictability modulates human brain response to reward. *Journal of Neuroscience*, *21*, 2793–2798.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Berlin, Heidelberg: Springer-Verlag.

Blanco, N. J., Love, B. C., Ramscar, M., Otto, A. R., Smayda, K., & Maddox, W. T. (2016). Exploratory decision-making as a function of lifelong experience, not cognitive decline.

Journal of Experimental Psychology: General, *145*, 284–297.

Blough, D. S. (1969). Generalisation gradient shape and summation in steady-state tests. *Journal of the Experimental Analysis of Behavior*, *12*, 91–104.

Blough, D. S. (1975). Steady state data and a quantitative model of operant generalization and discrimination. *Journal of Experimental Psychology: Animal Behavior Processes*, *1*, 3.

Bonawitz, E. B., Denison, S., Gopnik, A., & Griffiths, T. L. (2014). Win-stay, lose-sample: A simple sequential algorithm for approximating bayesian inference. *Cognitive Psychology*, *74*, 35–65.

Bonawitz, E. B., van Schijndel, T. J., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*, 215–234.

Borji, A., & Itti, L. (2013). Bayesian optimization explains human active search. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 26* (pp. 55–63). Curran Associates, Inc.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144–152).

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT 2010* (pp. 177–186). Springer.

Bouhlel, I., Wu, C. M., Hanaki, N., & Goldstone, R. L. (2018). Sharing is not erring: Pseudo-reciprocity in collective search. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 156–161). Austin, TX: Cognitive Science Society.

Boyan, J. A., & Moore, A. W. (1995). Generalization in reinforcement learning: Safely approximating the value function. *Advances in Neural Information Processing Systems*, 369–376.

Bramley, N. R., Dayan, P., Griffiths, T. L., & Lagnado, D. A. (2017). Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological Review*, *124*(3), 301-338.

Brandstätter, E., Gigerenzer, G., & Hertwig, R. (2006). The priority heuristic: making choices without trade-offs. *Psychological Review*, *113*, 409.

Brehmer, B. (1974). Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance*, *11*(1), 1–27.

Bromberg-Martin, E. S., Matsumoto, M., Hong, S., & Hikosaka, O. (2010). A pallidus-habenula-dopamine pathway signals inferred stimulus values. *Journal of Neurophysiology*, *104*.

Brown, C. R., Brown, M. B., & Shaffer, M. L. (1991). Food-sharing signals among socially foraging cliff swallows. *Animal Behaviour*, *42*(4), 551–564.

Bubeck, S., & Cesa-Bianchi, N. (2012). Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, *5*, 1–122.

Busemeyer, J. R., Byun, E., DeLosh, E. L., & McDaniel, M. A. (1997). Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In K. Lamberts & D. Shanks (Eds.), *Concepts and Categories* (p. 405-437). Cambridge: MIT Press.

Bush, R. R., & Mosteller, F. (1951a). A mathematical model for simple learning. *Psychological Review*, *58*, 313.

Bush, R. R., & Mosteller, F. (1951b). A model for stimulus generalization and discrimination. *Psychological Review*, *58*, 413.

Carroll, J. D. (1963). Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series*, *1963*, i–144.

Cauffman, E., Shulman, E. P., Steinberg, L., Claus, E., Banich, M. T., Graham, S., & Woolard, J. (2010). Age differences in affective decision making as indexed by performance on the iowa gambling task. *Developmental Psychology*, *46*, 193–207.

Cavanagh, P., Hunt, A. R., Afraz, A., & Rolfs, M. (2010). Visual stability based on remapping of attention pointers. *Trends in Cognitive Sciences*, *14*, 147–153.

Chater, N., & Vitányi, P. M. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, *47*, 346–369.

Cheng, K. (2000). Shepard's universal law supported by honeybees in spatial generalization. *Psychological Science*, *11*, 403–408.

Christakou, A., Gershman, S. J., Niv, Y., Simmons, A., Brammer, M., & Rubia, K. (2013). Neural and psychological maturation of decision-making in adolescence and young adulthood. *Journal of Cognitive Neuroscience*, *25*, 1807–1823.

Christakou, A., Murphy, C. M., Chantiluke, K., Cubillo, A. I., Smith, A. B., Giampietro, V., … Rubia, K. (2013). Disorder-specific functional abnormalities during sustained attention in youth with attention deficit hyperactivity disorder (adhd) and with autism. *Molecular Psychiatry*, *18*, 236–244.

Coenen, A., Nelson, J. D., & Gureckis, T. M. (2018). Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, 1–41.

Collins, A. G., Brown, J. K., Gold, J. M., Waltz, J. A., & Frank, M. J. (2014). Working memory contributions to reinforcement learning impairments in schizophrenia. *Journal of Neuroscience*, *34*, 13747–13756.

Collins, A. G., & Frank, M. J. (2012). How much of reinforcement learning is working memory, not reinforcement learning? A behavioral, computational, and neurogenetic analysis. *European Journal of Neuroscience*, *35*, 1024–1035.

Collins, A. G., & Frank, M. J. (2018). Within-and across-trial dynamics of human eeg reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, *115*, 2502–2507.

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

Crupi, V., Nelson, J. D., Meder, B., Cevolani, G., & Tentori, K. (2018). Generalized information theory meets human cognition: Introducing a unified framework to model uncertainty and information search. *Cognitive Science*, *42*, 1410–1456.

Crupi, V., & Tentori, K. (2014). State of the field: Measuring information and confirmation. *Studies in History and Philosophy of Science Part A*, *47*, 81–90.

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

D'Ardenne, K., McClure, S. M., Nystrom, L. E., & Cohen, J. D. (2008). Bold responses reflecting dopaminergic signals in the human ventral tegmental area. *Science*, *319*, 1264–1267.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1–25.

Davidson, D. (1996). The effects of decision characteristics on children's selective search of predecisional information. *Acta Psychologica*, *92*, 263–281.

Daw, N. D., Gershman, S. J., Seymour, B., Dayan, P., & Dolan, R. J. (2011). Model-based influences on humans' choices and striatal prediction errors. *Neuron*, *69*, 1204–1215.

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

Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, *5*(4), 613–624.

Dayan, P., Kakade, S., & Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, *3*, 1218–1223.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology*, *18*, 185–196.

Decker, J. H., Otto, A. R., Daw, N. D., & Hartley, C. A. (2016). From creatures of habit to goal-directed learners: tracking the developmental emergence of model-based reinforcement learning. *Psychological science*, *27*, 848–858.

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

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

Draper, A. D. (1959). *Optimum plot size and shape for safflower yield tests* (Unpublished doctoral dissertation). The University of Arizona.

Dresler, M., Shirer, W. R., Konrad, B. N., Müller, N. C., Wagner, I. C., Fernández, G., ... Greicius, M. D. (2017). Mnemonic training reshapes brain networks to support superior memory. *Neuron*, *93*, 1227–1235.

Ekman, G. (1954). Dimensions of color vision. *The Journal of Psychology*, *38*, 467–474.

Erdős, P., & Simonovits, M. (1980). On the chromatic number of geometric graphs. *Ars Combin*, *9*, 229–246.

Erickson, M. A., & Kruschke, J. K. (2002). Rule-based extrapolation in perceptual categorization. *Psychonomic Bulletin & Review*, *9*, 160–168.

Findling, C., Skvortsova, V., Dromnelle, R., Palminteri, S., & Wyart, V. (2018). Computational noise in reward-guided learning drives behavioral variability in volatile environments. *BioRxiv*, 439885.

Frank, M. J., Doll, B. B., Oas-Terpstra, J., & Moreno, F. (2009). Prefrontal and striatal dopaminergic genes predict individual differences in exploration and exploitation. *Nature Neuroscience*, *12*, 1062–1068.

Frank, S. A. (2018). Measurement invariance explains the universal law of generalization for psychological perception. *Proceedings of the National Academy of Sciences*, *115*, 9803–9806.

Fruchterman, T. M., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and experience*, *21*, 1129–1164.

Gaissmaier, W., & Schooler, L. J. (2008). The smart potential behind probability matching. *Cognition*, *109*, 416–422.

Galesic, M., Goode, A. W., Wallsten, T. S., & Norman, K. L. (2018). Using Tversky's contrast model to investigate how features of similarity affect judgments of likelihood. *Judgment & Decision Making*, *13*.

Gardner, M. P., Schoenbaum, G., & Gershman, S. J. (2018). Rethinking dopamine as generalized prediction error. *Proceedings of the Royal Society B*, *285*, 20181645.

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge

in the human hippocampal–entorhinal cortex. *eLife*, *6*, e17086.
Geist, M., Piot, B., & Pietquin, O. (2017). Is the bellman residual a bad proxy? In *Advances in Neural Information Processing Systems* (pp. 3205–3214).
Gershman, S. J. (2015). A unifying probabilistic view of associative learning. *PLoS Computational Biology*, *11*, e1004567.
Gershman, S. J. (2017). Dopamine, inference, and uncertainty. *Neural Computation*, *29*, 3311–3326.
Gershman, S. J. (2018a). Deconstructing the human algorithms for exploration. *Cognition*, *173*, 34–42.
Gershman, S. J. (2018b). The successor representation: Its computational logic and neural substrates. *Journal of Neuroscience*, *38*, 7193–7200.
Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*, 101–128.
Gershman, S. J., Malmaud, J., & Tenenbaum, J. B. (2017). Structured representations of utility in combinatorial domains. *Decision*, *4*, 67.
Gershman, S. J., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current opinion in neurobiology*, *20*, 251–256.
Gershman, S. J., & Niv, Y. (2015). Novelty and inductive generalization in human reinforcement learning. *Topics in Cognitive Science*, *7*, 391–415.
Gershman, S. J., Pesaran, B., & Daw, N. D. (2009). Human reinforcement learning subdivides structured action spaces by learning effector-specific values. *Journal of Neuroscience*, *29*, 13524–13531.
Ghirlanda, S., & Enquist, M. (2003). A century of generalization. *Animal Behaviour*, *66*, 15–36.
Gigerenzer, G., & Hoffrage, U. (1995). How to improve bayesian reasoning without instruction: frequency formats. *Psychological Review*, *102*, 684–704.
Gigerenzer, G., & Todd, P. M. (2012). Ecological rationality: the normative study of heuristics. In *Ecological rationality: Intelligence in the world* (pp. 487–497). Oxford University Press.
Gigerenzer, G., Todd, P. M., & the ABC Research Group. (1999). *Simple heuristics that make us smart*. New York: Oxford University Press.
Gittins, J. C. (1979). Bandit processes and dynamic allocation indices. *Journal of the Royal Statistical Society: Series B (Methodological)*, *41*, 148–164.
Gittins, J. C., & Jones, D. (1974). A dynamic allocation index for the sequential design of experiments. In J. Gani (Ed.), *Progress in statistics* (p. 241-266). Amsterdam: North-Holland.
Gittins, J. C., & Jones, D. M. (1979). A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, *66*, 561–565.
Glimcher, P. W. (2011). Understanding dopamine and reinforcement learning: the dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*, 15647–15654.
Goldstone, R. L., & Son, J. Y. (2012). Similarity. In K. J. Holyoak & R. G. Morrison (Eds.), *The oxford handbook of thinking and reasoning*. Oxford University Press.
Gopnik, A., Griffiths, T. L., & Lucas, C. G. (2015). When younger learners can be better (or at least more open-minded) than older ones. *Current Directions in Psychological Science*, *24*, 87–92.
Gopnik, A., O'Grady, S., Lucas, C. G., Griffiths, T. L., Wente, A., Bridgers, S., ... Dahl, R. E.

(2017). Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *Proceedings of the National Academy of Sciences*, *114*, 7892–7899.

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

Goulden, C. H. (1939). *Methods of statistical analysis*. John Wiley and Sons, Inc.

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

Griffiths, T. L., & Tenenbaum, J. B. (2009). Theory-based causal induction. *Psychological Review*, *116*, 661.

Gronau, Q. F., Singmann, H., & Wagenmakers, E.-J. (2017). Bridgesampling: An R package for estimating normalizing constants. *arXiv preprint arXiv:1710.08162*.

Guez, A., Silver, D., & Dayan, P. (2013). Scalable and efficient Bayes-adaptive reinforcement learning based on monte-carlo tree search. *Journal of Artificial Intelligence Research*, *48*, 841–883.

Gustafson, N. J., & Daw, N. D. (2011). Grid cells, place cells, and geodesic generalization for spatial reinforcement learning. *PLoS Computational Biology*, *7*, e1002235.

Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology*, *51*, 79.

Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801.

Hagen, J. W., & Hale, G. A. (1973). The development of attention in children. *ETS Research Report Series*, *1973*.

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*, *87*, 1–32.

Hanson, H. M. (1959). Effects of discrimination training on stimulus generalization. *Journal of experimental psychology*, *58*, 321.

Hartley, C. A., & Somerville, L. H. (2015). The neuroscience of adolescent decision-making. *Current Opinion in Behavioral Sciences*, *5*, 108–115.

Herbrich, R., Lawrence, N. D., & Seeger, M. (2003). Fast sparse gaussian process methods: The informative vector machine. In *Advances in Neural Information Processing Systems 15* (pp. 625–632).

Hertwig, R., Barron, G., Weber, E. U., & Erev, I. (2004). Decisions from experience and the effect of rare events in risky choice. *Psychological Science*, *15*, 534–539.

Hertwig, R., Hoffrage, U., & Martignon, L. (1999). Quick estimation. In G. Gigerenzer, P. M. Todd, & the ABC Research Group (Eds.), *Simple heuristics that make us smart* (pp. 209–234). New York: Oxford University Press.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science*, *30*, 3–41.

Hills, T. T., & Hertwig, R. (2010). Information search in decisions from experience: Do our patterns of sampling foreshadow our decisions? *Psychological Science*, *21*, 1787–1792.

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

Hills, T. T., Kalff, C., & Wiener, J. M. (2013). Adaptive lévy processes and area-restricted search in human foraging. *PLoS One*, *8*, e60488.

Hills, T. T., Todd, P. M., & Goldstone, R. L. (2008). Search in external and internal spaces: Evidence for generalized cognitive search processes. *Psychological Science*, *19*, 802–

808.

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., & the Cognitive Search Research Group. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*, 46–54.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, *15*, 1593–1623.

Hull, C. (1943). *Principles of behavior*. Appleton-Century, New York.

Humphries, N. E., Weimerskirch, H., Queiroz, N., Southall, E. J., & Sims, D. W. (2012). Foraging success of biological lévy flights recorded in situ. *Proceedings of the National Academy of Sciences*, *109*, 7169–7174.

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

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008a). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, *15*, 256–271.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008b). Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, *52*, 297–303.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences*, *13*, 381–388.

James, W. (1890). *The Principles of Psychology*. Dover, New York.

Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences*, *186*, 453–461.

Jeffreys, H. (1961). *The Theory of Probability*. Oxford, UK: Oxford University Press.

Johnson, S. G. (2014). *The nlopt nonlinear-optimization package.* Retrieved from `http://ab-initio.mit.edu/nlopt`

Josef, A. K., Richter, D., Samanez-Larkin, G. R., Wagner, G. G., Hertwig, R., & Mata, R. (2016). Stability and change in risk-taking propensity across the adult life span. *Journal of Personality and Social Psychology*, *111*, 430.

Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, *3*, 430–454.

Kakade, S., & Dayan, P. (2002). Dopamine: generalization and bonuses. *Neural Networks*, *15*, 549–559.

Kalamkar, R. (1932). A study in sampling technique with wheat. *The Journal of Agricultural Science*, *22*, 783–796.

Kalish, M. L., Griffiths, T. L., & Lewandowsky, S. (2007). Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychonomic Bulletin & Review*, *14*, 288–294.

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

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, *82*, 35–45.

Kaplan, R., Schuck, N. W., & Doeller, C. F. (2017). The role of mental maps in decision-making. *Trends in Neurosciences*, *40*, 256–259.

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

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the*

National Academy of Sciences, *105*, 10687–10692.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*, 20.

Keresztes, A., Bender, A., Bodammer, N., Lindenberger, U., Shing, Y. L., & Werkle-Bergner, M. (2017). Hippocampal maturity promotes memory distinctiveness in childhood and adolescence. *Proceedings of the National Academy of Sciences*, *114*, 9212–9217.

Khin, S. (2016). *Investigation into the relative costs of rice experiments based on the efficiency of designs* (Unpublished doctoral dissertation). University of the West Indies.

Kierkegaard, S. (2004/1849). *The sickness unto death: a Christian psychological exposition of edification and awakening by anti-Climacus*. Penguin UK.

Klahr, D. (1982). Nonmonotone assessment of monotone development: An information processing analysis. In S. Strauss & R. Stavy (Eds.), *U-shaped behavioral growth* (pp. 63–86). New York: Academic Press.

Kober, J., Bagnell, J. A., & Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, *32*, 1238–1274.

Koh, K., & Meyer, D. E. (1991). Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*, 811.

Kolling, N., Behrens, T. E., Mars, R. B., & Rushworth, M. F. (2012). Neural mechanisms of foraging. *Science*, *336*, 95–98.

Kondor, R. I., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning* (pp. 315–322).

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

Krishna Iyer, P. V. (1942). Studies with wheat uniformity trial data. I. Size and shape of experimental plots and the relative efficiency of different layouts. *The Indian Journal of Agricultural Science*, *12*, 240–262.

Kristensen, R. (1925). Anlaeg og opgoerelse af markforsoeg. *Tidsskrift for landbrugets planteavl*, *31*, 464–494.

Kruschke, J. K. (1992). Alcove: an exemplar-based connectionist model of category learning. *Psychological Review*, *99*, 22.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*, 3–36.

Kruschke, J. K. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, *36*, 210–226.

Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*, 1–27.

Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, *29*, 115–129.

Kwantes, P. J., & Neal, A. (2006). Why people underestimate y when extrapolating in linear functions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1019.

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

Lak, A., Stauffer, W. R., & Schultz, W. (2014). Dopamine prediction error responses integrate subjective value from different reward dimensions. *Proceedings of the National Academy of Sciences*, *111*, 2343–2348.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*, 1332–1338.

Lake, B. M., Salakhutdinov, R. R., & Tenenbaum, J. (2013). One-shot learning by inverting a compositional causal process. In *Advances in Neural Information Processing Systems* (pp. 2526–2534).

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, 1–101.

Lakoff, G., & Johnson, M. (2008). *Metaphors We Live By*. University of Chicago press.

Landau, B., & Jackendoff, R. (1993). Whence and whither in spatial language and spatial cognition? *Behavioral and Brain Sciences*, *16*, 255–265.

Landauer, T. K., & Dumais, S. T. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.

Lazer, D., & Friedman, A. (2007). The network structure of exploration and exploitation. *Administrative Science Quarterly*, *52*, 667–694.

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

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

Lejarraga, T., Pachur, T., Frey, R., & Hertwig, R. (2016). Decisions from experience: From monetary to medical gambles. *Journal of Behavioral Decision Making*, *29*, 67–77.

Leuker, C., Pachur, T., Hertwig, R., & Pleskac, T. J. (2018). Exploiting risk–reward structures in decision making under uncertainty. *Cognition*, *175*, 186–200.

Lichtenberg, J. M., & Simsek, Ö. (2016). Simple regression models. In *Proceedings of the NIPS 2016 Workshop on Imperfect Decision Makers: Admitting Real-World Rationality, Barcelona, Spain, December 9, 2016.* (pp. 13–25).

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

Ljungberg, T., Apicella, P., & Schultz, W. (1992). Responses of monkey dopamine neurons during learning of behavioral reactions. *Journal of Neurophysiology*, *67*, 145–163.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: a network model of category learning. *Psychological Review*, *111*, 309.

Lucas, C. G., Bridgers, S., Griffiths, T. L., & Gopnik, A. (2014). When children are better (or at least more open-minded) learners than adults: Developmental differences in learning the forms of causal relationships. *Cognition*, *131*, 284–299.

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

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods*, *49*, 1494–1502.

Lusena, C., Goldsmith, J., & Mundhenk, M. (2001). Nonapproximability results for partially observable markov decision processes. *Journal of Artificial Intelligence Research*, *14*, 83–103.

Ly, A., Verhagen, J., & Wagenmakers, E.-J. (2016). Harold jeffreys's default bayes factor hypothesis tests: Explanation, extension, and application in psychology. *Journal of Mathematical Psychology*, *72*, 19–32.

Machado, M. C., Rosenbaum, C., Guo, X., Liu, M., Tesauro, G., & Campbell, M. (2018). Eigenoption discovery through the deep successor representation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Mandelbrot, B. B. (1982). *The Fractal Geometry of Nature* (Vol. 1). WH freeman New York.

March, J. G. (1991). Exploration and exploitation in organizational learning. *Organization Science*, *2*, 71–87.

Markowitz, H. (1952). Portfolio selection. *The Journal of Finance*, *7*, 77–91.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc.

Marr, D., & Poggio, T. (1976). *From understanding computation to understanding neural circuitry* (Tech. Rep. No. AIM-357). Cambridge, MA: Massachusetts Institute of Technology, Institute of Artificial Intelligence.

Mason, W. A., Jones, A., & Goldstone, R. L. (2008). Propagation of innovations in networked groups. *Journal of Experimental Psychology: General*, *137*, 422.

Mason, W. A., & Watts, D. J. (2012). Collaborative learning in networks. *Proceedings of the National Academy of Sciences*, *109*, 764–769.

Mata, R., Wilke, A., & Czienskowski, U. (2013). Foraging across the life span: is there a reduction in exploration with aging? *Frontiers in Neuroscience*, *7*, 53.

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

McDowell, M., & Jacobs, P. (2017). Meta-analysis of the effect of natural frequencies on Bayesian reasoning. *Psychological Bulletin*, *143*, 1273–1312.

Meder, B., & Nelson, J. D. (2012). Information search with situation-specific reward functions. *Judgment and Decision Making*, *7*, 119–148.

Meder, B., Nelson, J. D., Jones, M., & Ruggeri, A. (2018). *Stepwise versus globally optimal search in children and adults.* (Manuscript submitted for publication)

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254.

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

Menegas, W., Babayan, B. M., Uchida, N., & Watabe-Uchida, M. (2017). Opposite initialization to novel cues in dopamine signaling in ventral and posterior striatum in mice. *eLife*, *6*, e21886.

Mercer, J. (1909). Functions of positive and negative type, and their connection the theory of integral equations. *Philosophical Transactions of the Royal Society A*, *209*, 415–446.

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

Meyer, J. (2014). The theory of risk and risk aversion. In *Handbook of the Economics of Risk and Uncertainty* (Vol. 1, pp. 99–133). Elsevier.

Michels, J., Saxena, A., & Ng, A. Y. (2005). High speed obstacle avoidance using monocular vision and reinforcement learning. In *Proceedings of the 22nd International Conference on Machine Learning* (pp. 593–600).

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some english consonants. *The Journal of the Acoustical Society of America*, *27*, 338–352.

Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, *49*, 8–30.

Mitchell, T. M. (1982). Generalization as search. *Artificial Intelligence*, *18*, 203–226.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533.

Močkus, J. (1975). On bayesian methods for seeking the extremum. In *Optimization Techniques*

IFIP Technical Conference (pp. 400–404).
Močkus, J. (2012). *Bayesian approach to global optimization: Theory and applications*. Springer Science & Business Media.
Momennejad, I., & Howard, M. W. (2018). Predicting the future with multi-scale Successor Representations. *bioRxiv*, 449470.
Momennejad, I., Russek, E. M., Cheong, J. H., Botvinick, M. M., Daw, N., & Gershman, S. J. (2017). The successor representation in human reinforcement learning. *Nature Human Behaviour*, *1*, 680.
Montague, P. R., Dayan, P., Nowlan, S. J., Pouget, A., & Sejnowski, T. (1993). Using aperiodic reinforcement for directed self-organization during development. In *Advances in Neural Information Processing Systems* (pp. 969–976).
Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.
Montgomery, E. (1912). Variation in yield and methods of arranging plats to secure comparative results. In *Twenty-Fifth Annual Report of the Agricultural Experiment Station of Nebraska* (pp. 164–180).
Moore, J. F., & Darroch, J. (1956). *Field plot technique with blue lake pole beans, bush beans, carrots, sweet corn, spring and fall cauliflower*. Washington Agricultural Experiment Stations, Institute of Agricultural Sciences, State College of Washington.
Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences*, *97*, 11170–11175.
Myung, I. J., Kim, C., & Pitt, M. A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory & Cognition*, *28*, 832–840.
Najemnik, J., & Geisler, W. S. (2005). Optimal eye movement strategies in visual search. *Nature*, *434*, 387.
Neal, R. M. (1994). *Bayesian learning for neural networks* (Unpublished doctoral dissertation). Dept. of Computer Science, University of Toronto.
Neal, R. M. (2012). *Bayesian learning for neural networks*. Springer Science & Business Media.
Neimark, E. D., & Shuford, E. H. (1959). Comparison of predictions and estimates in a probability learning situation. *Journal of Experimental Psychology*, *57*, 294.
Nelson, J. D. (2005). Finding useful questions: On bayesian diagnosticity, probability, impact, and information gain. *Psychological Review*, *112*, 979–999.
Nelson, J. D., & Cottrell, G. W. (2007). A probabilistic model of eye movements in concept formation. *Neurocomputing*, *70*, 2256–2272.
Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F., & Meder, B. (2014). Children's sequential information search is sensitive to environmental probabilities. *Cognition*, *130*, 74–80.
Nelson, J. D., McKenzie, C. R., Cottrell, G. W., & Sejnowski, T. J. (2010). Experience matters: Information acquisition optimizes probability gain. *Psychological Science*, *21*, 960–969.
Newton, I. (1687). *PhilosophiæNaturalis Principia Mathematica*. Royal Society, London.
Niv, Y., Daniel, R., Geana, A., Gershman, S. J., Leong, Y. C., Radulescu, A., & Wilson, R. C. (2015). Reinforcement learning in multidimensional environments relies on attention mechanisms. *The Journal of Neuroscience*, *35*, 8145–8157.
Nonnecke, I. (1959). The precision of field experiments with vegetable crops as influenced by plot and block size and shape: I. Sweet corn. *Canadian Journal of Plant Science*, *39*,

443-457.

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Perception & Psychophysics*, *38*, 415–432.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39.

Nosofsky, R. M. (1988a). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: learning, memory, and cognition*, *14*, 700–708.

Nosofsky, R. M. (1988b). Similarity, frequency, and category representations. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*, 54–65.

Odland, T., & Garber, R. (1928). Size of plat and number of replications in field experiments with soybeans. *Journal of the American Society of Agronomy*.

O'Doherty, J. P., Dayan, P., Friston, K., Critchley, H., & Dolan, R. J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron*, *38*, 329–337.

Ohl, S., & Rolfs, M. (2018). Saccadic selection of stabilized items in visuospatial working memory. *Consciousness and Cognition*, *64*, 32–44.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: Clarendon Press.

OpenAI, Andrychowicz, M., Baker, B., Chociej, M., Józefowicz, R., McGrew, B., ... Zaremba, W. (2018). Learning dexterous in-hand manipulation. *arXiv preprint arXiv:1808.00177*.

Oudeyer, P.-Y., Kaplan, F., Hafner, V. V., & Whyte, A. (2005). The playground experiment: Task-independent development of a curious robot. In *Proceedings of the AAAI Spring Symposium on Developmental Robotics* (pp. 42–47).

Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*, 97–98.

Palminteri, S., Khamassi, M., Joffily, M., & Coricelli, G. (2015). Contextual modulation of value signals in reward and punishment learning. *Nature Communications*, *6*, 8096.

Palminteri, S., Kilford, E. J., Coricelli, G., & Blakemore, S.-J. (2016). The computational development of reinforcement learning during adolescence. *PLoS Computational Biology*, *12*, e1004953.

Palminteri, S., Lefebvre, G., Kilford, E. J., & Blakemore, S.-J. (2017). Confirmation bias in human reinforcement learning: Evidence from counterfactual feedback processing. *PLoS Computational Biology*, *13*, e1005684.

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

Pavlov, I. P. (1927). *Conditional reflexes: an investigation of the physiological activity of the cerebral cortex.* Oxford University Press.

Penrose, R. (1955). A generalized inverse for matrices. In *Mathematical proceedings of the cambridge philosophical society* (Vol. 51, pp. 406–413).

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R. J., & Frith, C. D. (2006). Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*, 1042–1045.

Piaget, J. (1964). Part I: Cognitive development in children: Piaget development and learning. *Journal of Research in Science Teaching*, *2*, 176–186.

Pirolli, P., & Card, S. (1999). Information foraging. *Psychological Review*, *106*, 643.

Pitt, M. A., & Myung, I. J. (2002). When a good fit can be bad. *Trends in Cognitive Sciences*, *6*, 421–425.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computa-

tional models of cognition. *Psychological Review*, *109*, 472–491.
Poggio, T., & Bizzi, E. (2004). Generalization in vision and motor control. *Nature*, *431*, 768–774.
Poggio, T., & Girosi, F. (1989). *A theory of networks for approximation and learning* (Tech. Rep.). Cambridge, MA: Artificial Intelligence Laboratory, Massachusetts Institute of Technology.
Poggio, T., & Girosi, F. (1990). Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, *247*, 978–982.
Polat, U., Mizobe, K., Pettet, M. W., Kasamatsu, T., & Norcia, A. M. (1998). Collinear stimuli regulate visual responses depending on cell's contrast threshold. *Nature*, *391*, 580–584.
Polson, D. E. (1964). *Estimation of optimum size, shape, and replicate number of safflower plots for yield trials* (Unpublished doctoral dissertation). Utah State University.
Puterman, M. L. (1994). *Markov decision processes: discrete stochastic dynamic programming*. New York, NY: Wiley.
Quiñonero-Candela, J., & Rasmussen, C. E. (2005). A unifying view of sparse approximate gaussian process regression. *Journal of Machine Learning Research*, *6*, 1939–1959.
Radulescu, A., Niv, Y., & Ballard, I. (2019). Holistic reinforcement learning: The role of structure and attention. *Trends in Cognitive Sciences*, *23*, 278–292.
Raichlen, D. A., Wood, B. M., Gordon, A. D., Mabulla, A. Z., Marlowe, F. W., & Pontzer, H. (2014). Evidence of lévy walk foraging patterns in human hunter–gatherers. *Proceedings of the National Academy of Sciences*, *111*, 728–733.
Rasmussen, C. E., & Kuss, M. (2004). Gaussian Processes in Reinforcement Learning. In *Advances in Neural Information Processing Systems 16* (pp. 751–758). MIT Press: Cambridge, MA.
Rasmussen, C. E., & Williams, C. (2006). *Gaussian Processes for Machine Learning*. MIT Press: Cambridge, MA.
Reece, S., & Roberts, S. (2010). An introduction to Gaussian processes for the Kalman filter expert. In *13th Conference on Information Fusion (FUSION)* (pp. 1–9).
Rendell, L., Boyd, R., Cownden, D., Enquist, M., Eriksson, K., Feldman, M. W., ... Laland, K. N. (2010). Why copy others? insights from the social learning strategies tournament. *Science*, *328*, 208–213.
Rescorla, R. A., & Wagner, A. R. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64–99.
Reverdy, P. B., Srivastava, V., & Leonard, N. E. (2014). Modeling human decision making in generalized gaussian multiarmed bandits. *Proceedings of the IEEE*, *102*, 544–571.
Reynolds, A. M., Smith, A. D., Menzel, R., Greggers, U., Reynolds, D. R., & Riley, J. R. (2007). Displaced honey bees perform optimal scale-free search flights. *Ecology*, *88*(8), 1955–1961.
Riedmiller, M., Hafner, R., Lampe, T., Neunert, M., Degrave, J., Van de Wiele, T., ... Springenberg, J. T. (2018). Learning by playing-solving sparse reward tasks from scratch. *arXiv preprint arXiv:1802.10567*.
Rigoux, L., Stephan, K. E., Friston, K. J., & Daunizeau, J. (2014). Bayesian model selection for group studies—revisited. *Neuroimage*, *84*, 971–985.
Robbins, H. (1952). Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, *58*, 527–535.
Rolfs, M., Lawrence, B. M., & Carrasco, M. (2013). Reach preparation enhances visual performance and appearance. *Philosophical Transactions of the Royal Society B: Biological*

Sciences, *368*, 20130057.

Romero, P. A., Krause, A., & Arnold, F. H. (2013). Navigating the protein fitness landscape with Gaussian Processes. *Proceedings of the National Academy of Sciences*, *110*, 193–201.

Rothe, A., Lake, B. M., & Gureckis, T. M. (2018). Do people ask good questions? *Computational Brain & Behavior*, *1*, 69–89.

Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default bayes factors for anova designs. *Journal of Mathematical Psychology*, *56*, 356–374.

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*, 225–237.

Ruggeri, A., & Lombrozo, T. (2015). Children adapt their questions to achieve efficient search. *Cognition*, *143*, 203–216.

Sallab, A. E., Abdou, M., Perot, E., & Yogamani, S. (2017). Deep reinforcement learning framework for autonomous driving. *Electronic Imaging*, *2017*, 70–76.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin*, *2*, 110–114.

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

Schölkopf, B., Smola, A., & Müller, K.-R. (1997). Kernel principal component analysis. In *International Conference on Artificial Neural Networks* (pp. 583–588).

Schölkopf, B., & Smola, A. J. (2001). *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press.

Schuck, N. W., Cai, M. B., Wilson, R. C., & Niv, Y. (2016). Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, *91*, 1402–1412.

Schultz, W. (2016). Dopamine reward prediction-error signalling: a two-component response. *Nature Reviews Neuroscience*, *17*, 183–195.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593–1599.

Schulz, E., Franklin, N. T., & Gershman, S. J. (2018). Finding structure in multi-armed bandits. *bioRxiv*, 432534.

Schulz, E., Konstantinidis, E., & Speekenbrink, M. (2017). Putting bandits into context: How function learning supports decision making. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 927–943.

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.

Schulz, E., Tenenbaum, J., Duvenaud, D. K., Speekenbrink, M., & Gershman, S. J. (2016). Probing the compositionality of intuitive functions. In *Advances In Neural Information Processing Systems* (pp. 3729–3737).

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

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

Schulz, E., Wu, C. M., Huys, Q. J., Krause, A., & Speekenbrink, M. (2018). Generalization and search in risky environments. *Cognitive Science*, *42*, 2592–2620.

Schulz, E., Wu, C. M., Ruggeri, A., & Meder, B. (in press). Searching for rewards like a child means less generalization and more directed exploration. *Psychological Science*.

Schulz, L. E. (2015). Infants explore the unexpected. *Science*, *348*, 42–43.
Seymour, B., Daw, N., Dayan, P., Singer, T., & Dolan, R. (2007). Differential encoding of losses and gains in the human striatum. *Journal of Neuroscience*, *27*, 4826–4831.
Shannon, C. E. (1988). Programming a computer for playing chess. In *Computer chess compendium* (pp. 2–13). Springer.
Shepard, R. N. (1958). Stimulus and response generalization: Deduction of the generalization gradient from a trace model. *Psychological Review*, *65*, 242.
Shepard, R. N. (1962a). The analysis of proximities: multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*, 125–140.
Shepard, R. N. (1962b). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, *27*, 219–246.
Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–489.
Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology*, *41*, 1–20.
Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*, 652–656.
Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis*. Appleton-Century, New York.
Skov, R. B., & Sherman, S. J. (1986). Information-gathering processes: Diagnosticity, hypothesis-confirmatory strategies, and perceived hypothesis confirmation. *Journal of Experimental Social Psychology*, *22*, 93–121.
Slowiaczek, L. M., Klayman, J., Sherman, S. J., & Skov, R. B. (1992). Information selection and use in hypothesis testing: What is a good question, and what is a good answer? *Memory & Cognition*, *20*, 392–405.
Sollich, P. (1999). Learning curves for Gaussian processes. In *Advances in Neural Information Processing Systems* (pp. 344–350).
Solway, A., & Botvinick, M. M. (2015). Evidence integration in model-based tree search. *Proceedings of the National Academy of Sciences*, *112*, 11708–11713.
Somerville, L. H., Sasse, S. F., Garrad, M. C., Drysdale, A. T., Abi Akar, N., Insel, C., & Wilson, R. C. (2017). Charting the expansion of strategic exploratory behavior during adolescence. *Journal of Experimental Psychology: General*, *146*, 155–164.
Speekenbrink, M., & Konstantinidis, E. (2015). Uncertainty and exploration in a restless bandit problem. *Topics in Cognitive Science*, *7*, 351–367.
Spitzer, B., Waschke, L., & Summerfield, C. (2017). Selective overweighting of larger magnitudes during noisy numerical comparison. *Nature Human Behaviour*, *1*, 0145.
Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. (2010). Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. *Proceedings of the 27th International Conference on Machine Learning (ICML 2010)*, 1015–1022.
Srivastava, V., Reverdy, P., & Leonard, N. E. (2015). Correlated multiarmed bandit problem: Bayesian algorithms and regret analysis. *arXiv preprint arXiv:1507.01160*.
Stachenfeld, K. L., Botvinick, M., & Gershman, S. J. (2014). Design principles of the hippocampal cognitive map. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 2528–2536).
Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a

predictive map. *Nature Neuroscience*, *20*, 1643–1653.
Steinwart, I., Hush, D., & Scovel, C. (2006). An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels. *IEEE Transactions on Information Theory*, *52*, 4635–4643.
Stephan, K. E., Penny, W. D., Daunizeau, J., Moran, R. J., & Friston, K. J. (2009). Bayesian model selection for group studies. *Neuroimage*, *46*, 1004–1017.
Stephens, J. C., & Vinall, H. (1928). Experimental methods and the probable error in field experiments with sorghum. *Journal of Agricultural Research*, *37*, 629–646.
Steyvers, M., Lee, M. D., & Wagenmakers, E.-J. (2009). A Bayesian analysis of human decision-making on bandit problems. *Journal of Mathematical Psychology*, *53*, 168–179.
Stojic, H., Analytis, P. P., & Speekenbrink, M. (2015). Human behavior in contextual multi-armed bandit problems. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (pp. 2290–2295).
Sui, Y., Gotovos, A., Burdick, J., & Krause, A. (2015). Safe exploration for optimization with Gaussian processes. In *International Conference on Machine Learning* (pp. 997–1005).
Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*, 9–44.
Sutton, R. S. (1996). Generalization in reinforcement learning: Successful examples using sparse coarse coding. In *Advances in Neural Information Processing Systems 8* (pp. 1038–1044).
Sutton, R. S., & Barto, A. G. (1990). *Time-derivative models of pavlovian reinforcement.* The MIT Press.
Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (Second ed.). MIT press.
Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In *Advances in Neural Information Processing Systems* (pp. 59–68).
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–640.
Tesauro, G. (1992). Practical issues in temporal difference learning. *Machine Learning*, *8*, 257–277.
Theves, S., Fernandez, G., & Doeller, C. F. (2019). The hippocampus encodes distances in multidimensional feature space. *Current Biology*, *29*, 1226–1231.
Thomas, D. R., & Bistey, G. (1964). Stimulus generalization as a function of the number and range of generalization test stimuli. *Journal of Experimental Psychology*, *68*, 599–604.
Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, *25*, 285–294.
Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*, i–109.
Thorndike, E. L. (1911). *Animal intelligence: Experimental studies*. Hafner, Darien, CT.
Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. *Current Directions in Psychological Science*.
Todd, P. M., Hills, T. T., & Robbins, T. W. (2012). *Cognitive search: Evolution, algorithms, and the brain*. MIT press.
Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review*, *55*, 189.
Torgerson, W. S. (1952). Multidimensional scaling: I. Theory and method. *Psychometrika*, *17*, 401–419.
Tump, A. N., Wu, C. M., Bouhlel, I., & Goldstone, R. L. (2019). The evolutionary dynamics of

cooperation in collective search. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (p. 883—889). Montreal, QB: Cognitive Science Society.

Turing, A. (1950). Computing intelligence and machinery. *Mind*, *59*, 433–460.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological Review*, *89*, 123.

Tymula, A., Belmaker, L. A. R., Roy, A. K., Ruderman, L., Manson, K., Glimcher, P. W., & Levy, I. (2012). Adolescents' risk-taking behavior is driven by tolerance to ambiguity. *Proceedings of the National Academy of Sciences*, *109*, 17135-17140.

Ullman, T. D., Goodman, N. D., & Tenenbaum, J. B. (2012). Theory learning as stochastic search in the language of thought. *Cognitive Development*, *27*, 455–480.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2017). Bayesian latent-normal inference for the rank sum test, the signed rank test, and Spearman's ρ. *arXiv preprint arXiv:1712.06941*.

van Doorn, J., Ly, A., Marsman, M., & Wagenmakers, E.-J. (2018). Bayesian inference for kendall's rank correlation coefficient. *The American Statistician*, *72*, 303–308.

Vinyals, O., Babuschkin, I., Chung, J., Mathieu, M., Jaderberg, M., Czarnecki, W. M., ... Silver, D. (2019). *AlphaStar: Mastering the Real-Time Strategy Game StarCraft II*. `https://deepmind.com/blog/alphastar-mastering-real-time-strategy-game-starcraft-ii/`.

Viswanathan, G. M., Buldyrev, S. V., Havlin, S., Da Luz, M., Raposo, E., & Stanley, H. E. (1999). Optimizing the success of random searches. *Nature*, *401*, 911.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*, 599–637.

Wagenmakers, E.-J., Verhagen, J., & Ly, A. (2016). How to quantify the evidence for the absence of a correlation. *Behavior Research Methods*, *48*, 413–426.

Wald, A. (1947). *Sequential analysis*. Wiley, New York.

Wei, L. (1979). The generalized Polya's urn design for sequential medical trials. *The Annals of Statistics*, 291–296.

Wesman, A. G., & Bennett, G. K. (1959). Multiple regression vs. simple addition of scores in prediction of college grades. *Educational and Psychological Measurement*, *19*, 243–246.

White, J. M. (2013). *The role of delayed consequences in human decision-making.* (Unpublished doctoral dissertation). Princeton University.

Wilke, A., Minich, S., Panis, M., Langen, T. A., Skufca, J. D., & Todd, P. M. (2015). A game of hide and seek: Expectations of clumpy resources influence hiding and searching patterns. *PloS One*, *10*, e0130976.

Williams, C. K. (1998). Prediction with gaussian processes: From linear regression to linear prediction and beyond. In *Learning in graphical models* (pp. 599–621). Springer.

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*, 2074–2081.

Wright, K. (2017). agridat: Agricultural datasets [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=agridat` (R package version 1.13)

Wu, C. M., Meder, B., Filimon, F., & Nelson, J. D. (2017). Asking better questions: How presentation formats influence information search. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *8*, 1274–1297.

Wu, C. M., Meder, B., & Nelson, J. D. (in prep). Navigating uncertainty through information

search.

Wu, C. M., Schulz, E., Garvert, M. M., Meder, B., & Schuck, N. W. (2018). Connecting conceptual and spatial search via a model of generalization. In T. T. Rogers, M. Rau, X. Zhu, & C. W. Kalish (Eds.), *Proceedings of the 40th Annual Conference of the Cognitive Science Society* (pp. 1183–1188). Austin, TX: Cognitive Science Society.

Wu, C. M., Schulz, E., Gerbaulet, K., Pleskac, T. J., & Speekenbrink, M. (2019). Under pressure: The influence of time limits on human exploration. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (p. 1219—1225). Montreal, QB: Cognitive Science Society.

Wu, C. M., Schulz, E., & Gershman, S. J. (2019). Generalization as diffusion: human function learning on graphs. In A. Goel, C. Seifert, & C. Freksa (Eds.), *Proceedings of the 41st Annual Conference of the Cognitive Science Society* (p. 3122—3128). Montreal, QB: Cognitive Science Society.

Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2017). Mapping the unknown: The spatially correlated multi-armed bandit. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. J. Davelaar (Eds.), *Proceedings of the 39th Annual Conference of the Cognitive Science Society* (pp. 1357–1362). Austin, TX: Cognitive Science Society.

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*, 915–924.

Yates, F. A. (2013). *Art of Memory*. Routledge.

Zellner, A., & Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, D. V. Lindley, & A. F. M. Smith (Eds.), *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)* (pp. 585–603). University of Valencia.

Zhang, W., & Zhang, N. L. (2005). Restricted value iteration: Theory and algorithms. *Journal of Artificial Intelligence Research*, *23*, 123–165.

Ziegler, S., Pedersen, M. L., Mowinckel, A. M., & Biele, G. (2016). Modelling ADHD: A review of ADHD theories through their predictions for computational models of decision-making and reinforcement learning. *Neuroscience & Biobehavioral Reviews*, *71*, 633–656.

Declaration

I hereby declare that:

- I completed the doctoral thesis independently. Except where otherwise stated, I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.
- I have not applied for a doctoral degree elsewhere and do not have a corresponding doctoral degree.
- I have not submitted the doctoral thesis, or parts of it, to another academic institution and the thesis has not been accepted or rejected.
- I have acknowledged the Doctoral Degree Regulations which underlie the procedure of the Faculty of Life Sciences of Humboldt-Universität zu Berlin, as amended on 5th March 2015.
- No collaboration with commercial doctoral degree supervisors took place.
- The principles of Humboldt-Universität zu Berlin for ensuring good academic practice have been complied with.

Charley M. Wu
August 2019