Structure and similarity form the foundation of generalization

Charley M. Wu[1,2,3] (charley.wu@uni-tuebingen.de), Björn Meder[2,4], & Eric Schulz[3]
[1]University of Tübingen
[2]Max Planck Institute for Human Development
[3]Max Planck Institute for Biological Cybernetics
[4]Health and Medical University Potsdam

Abstract

The ability to generalize from limited experiences to novel situations is crucial for intelligent behavior, and provides key insights into the structure of our mental representations and learning mechanisms. Recent work in diverse areas have proposed computational models describing how the generalization of value can support efficient learning and flexible behavior in novel situations. Although the computational methods are new, the conceptual ideas and theoretical underpinnings are very familiar to psychology. In this review, we relate psychological theories of concept learning and function learning to modern methods for value generalization in reinforcement learning settings. Classic debates about the key mechanisms of generalization have largely coalesced around two ingredients: extracting regularities of the environment in the form of generic "rules", and using "similarity" to compare new situations to previously encountered instances. Yet today, the most prolific models are considered hybrids, where the use of Bayesian principles offer both rule- and similarity-based interpretations. While a duality of interpretations suggests an exchangeability of representations, the computations used by hybrid models typically operate over either one or the other—over hypothesized rules or over similarity representations—each conferring distinct advantages. Rules unlock compositionality and rapid transfer, while similarity is easy to compute and can flexibly capture any learned relationships in the environment. Here, we ask whether the unification of rules and similarity is truly complete, and provide perspectives on how to forge new hybrids to capture the full power of human generalization.

Introduction

Generalization is the ability to leverage past experience to inform novel situations. Since life is constantly brimming with new experiences and novel situations, it is no surprise that generalization is considered to be at the heart of intelligence[1,2], and widely studied in psychology[2–5], neuroscience[6–8], and machine learning[9–11]. Indeed, cognition without generalization would hardly resemble intelligence at all: each new situation or minor deviation from past experience would leave us floundering about, unable to leverage relevant experiences to form an adaptive response or make sense of the world.

Over the years, there have been numerous debates about the key mechanisms underlying human generalization, spanning multiple domains. Research in concept learning has studied how people generalize learned category labels when asked to classify new instances, for instance, identifying the breed of a dog or deciding whether a hotdog is a sandwich. Meanwhile, research in function learning has studied how people generalize by learning the relationship between inputs and outputs allowing for interpolation or extrapolation beyond observed data, such as predicting how much study time is needed to pass a test or anticipating how much you will enjoy a new menu item at your favorite restaurant. In both domains, theories about the underlying mechanisms of generalization have largely coalesced around two ingredients: extracting regularities of the environment in the form of generic "rules" to apply in novel settings and using "similarity" to compare new situations to previously encountered instances with the expectation that similar outcomes will result from similar situations.

Today, debates over which ingredient is more central have largely been settled in favor of hybrid models, which have both rule- and similarity-based interpretations and are often based on Bayesian principles (see REF[12] for a review). While a duality of interpretations suggests an exchangeability[13] between rule- and similarity-based representations, the computations used by hybrid models typically operate over either one or the other—over hypothesized rules or over representations of similarity[14]—each conferring distinct advantages. Rules unlock compositionality and rapid transfer, while similarity is easy to compute and can flexibly capture any learned relationships in the environment. Here, we ask whether the unification of rules and similarity is truly complete, and examine whether further integration of using rules for learning structure and using similarity for generalizing outcomes is necessary to achieve the full scope of human generalization.

Understanding the mechanisms of human generalization is all the more relevant in light of current advances in reinforcement learning[15] (RL). In RL problems, agents (biological or artificial) learn through interactions with the environment, receiving feedback in the form of rewards. Actions can thus be selected on the basis of value (i.e., reward expectations). Yet in complex tasks, it simply is not feasible to learn the value of all possible actions. Consider the simple problem of choosing what to cook for dinner. Given the immense

combinatorial space of possible meals you could make, it would be impossible to independently learn the value of each option through exhaustive exploration. Thus, a crucial feature of human learning efficiency and behavioral flexibility is being able to generalize value from limited experiences to a wide range of novel settings [2,7,16–18]. While the computational methods used to study human value generalization in RL settings are certainly new, the theoretical underpinnings and core questions are very familiar to psychology, and can be traced back to foundational research in concept and function learning.

In this review, we revisit the distinction between similarity- and rule-based mechanisms of generalization, bridging the domains of concept learning, function learning, and RL. Although flexibility can be seen as a strength, similarity-based approaches have historically been criticized for being too flexible to the point of being arbitrary[19,20].Yet, through the lens of modern theories, we show there exists a rational rather than arbitrary basis for how similarity representations are actively constructed to capture task-relevant aspects of the environment. And while similarity-based generalization can flexibly capture task-relevant and rich relational structures, rule-based methods may provide the more promising avenue for describing how the hidden structure in the environment can be inferred in the first place. The benefit of rules is that they can be compositionally combined, scaffolding upon previously learned schemas to propose novel yet likely structures, thus supporting inference from sparse data. We conclude by proposing new directions for further integrating similarity and rules, combining their relative advantages to unlock faster and more efficient generalization over hidden structure in the environment.

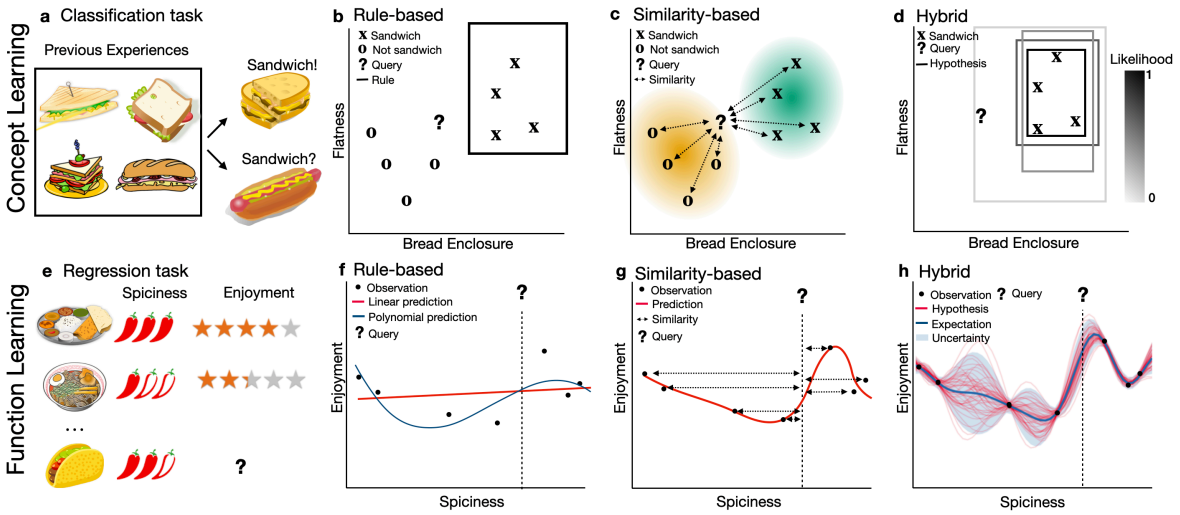

Fig. 1 | **From concept learning to function learning. a)** Concept learning is often studied based on classifying discrete stimuli (e.g., sandwich vs. not sandwich). **b)** Rule-based methods describe explicit category boundaries (rectangle), while **c)** similarity-based methods generalize category labels based on similarity (arrows) to previous exemplars or learned prototypes. **d)** Bayesian concept learning[5] provides a hybrid approach by defining a distribution over rules (rectangles), which produce patterns of generalization consistent with similarity-based accounts[3,21]. The likelihood of each hypothesis is indicated by the shading of lines, with narrower hypotheses being more likely. **e)** Function learning is the problem of learning a relational mapping between inputs (e.g., spiciness) and outputs (e.g., enjoyment), akin to a regression analysis. **f)** Rule-based methods describe specific parametric families of functions (e.g., linear or polynomial), while **g)** similarity-based methods commonly use Artificial Neural Networks (ANNs) to approximate nonlinear functions, where the influence of each data point is inversely proportional to their distance (arrows). **h)** Gaussian Process regression provides a hybrid approach using kernel similarity (Box 1) to compute a distribution over hypothesized functions (red lines), which can be summarized in terms of an expectation (blue line) and uncertainty (blue ribbon). Food images are from OpenClipArt under CC0 1.0.

Generalizing concepts, functions, and value

We first review theories of generalization in concept learning and function learning, which broadly map onto the distinction between classification and regression problems, respectively. Psychological research in both domains has largely focused on supervised learning tasks where participants generalize from labeled training data to predict the outcomes of novel stimuli (i.e., category labels or function outputs).

However, generalization is also central to reward learning, where computational approaches to RL have benefited greatly from advances in value function approximation[22,23]. In RL, a value function maps actions or states to

expectations of reward (i.e., value). For instance, learning which parts of a foreign city have the best nightlife or learning what to order at a new restaurant. Since you likely have not experienced all possible options, learning to approximate a value function allows you to extrapolate and interpolate from previous experiences, thus providing a mechanism for inferring the value of novel options.

In this section, we examine the common computational principles of generalization across domains and explore how theories of value generalization in RL tasks benefit from integration with classic theories of concept learning and function learning. Ultimately, the goal of this theory integration is to develop new models of human learning and decision-making that can scale up to increasingly more real-world and compositional problems.

Concept learning

Researchers have long studied generalization in the domain of concept learning using classification problems with discrete stimuli[24–27]. For instance, learning the category "sandwich" from examples of paninis and subs, and then generalizing confidently when shown a grilled cheese for the first time, but perhaps hesitating when shown a hotdog (Fig. 1a). An important debate in this field concerns how representations about categories are stored and the mechanisms used for generalizing about novel stimuli[28,29].

Rule-based concept learning

One influential class of theories proposes rule-based methods, which describe the explicit boundaries of category membership[30] (line in Fig. 1b). For instance, one might describe the necessary and sufficient features[24] of a sandwich as "food flattened between two pieces of bread", and thus classify any novel food that satisfies this rule as a sandwich. The specificity of rules facilitates rapid generalization, while their compositionality (i.e., the ability to combine multiple rules) makes them infinitely productive[13]. However, for the same reasons, they can be inflexible ("what about open-faced sandwiches?") and difficult to learn, since infinite productivity also implies an infinite hypothesis space of candidate rules to consider. Even with mechanisms for learning exceptions to rules[31] for added flexibility, rule-based methods only seem to offer partial explanations of human category learning[32], and perform best when paired together with other learning mechanisms[28,33,34]. Nevertheless, the basic

mechanisms of rule-based generalization (i.e., proposing explicit hypotheses) play an important role in modern theories of structure learning[35–37], as we discuss later.

Similarity-based concept learning

Another important class of theories uses similarity-based methods for predicting the category of novel stimuli (Fig. 1c). Through comparison to either previously encountered exemplars[25,26], or to a learned prototype[27,38] aggregated over multiple experiences, the most similar category can provide a predictive basis for generalization. A general principle is that stimuli with similar features are more likely to belong to the same category or concept[3]. Distance in feature space (arrows in Fig. 1c) thus provides a simple quantification of similarity. Yet the notion of similarity has been famously criticized for being too flexible, with endless and arbitrary ways to define similarity for any pair of stimuli[19,20,39]. However, modern theories provide new approaches for describing the psychological mechanisms people use to construct context-relevant similarity representations, forming a rational rather than arbitrary basis for computing similarity. And as we discuss later, advances in similarity-based approaches to generalization are now able to capture rich relational structure and represent the temporal dynamics of the environment.

Hybrid concept learning

Today, the most prolific theories of concept learning are considered hybrids and have a duality of both rule- and similarity-based interpretations. One influential example is the Bayesian concept learning framework[5] (Fig. 1d; Box 1), which uses a distribution over hypothesized category boundaries (rule-based interpretation) to infer the category membership of novel stimuli. Each of the boxes in Figure 1d represent a candidate hypothesis about the explicit boundaries of the category "sandwich", where a preference for simpler hypotheses makes narrower boundaries more likely[5,32,40].

Although the computations are over rule-based representations, Bayesian concept learning can replicate behavioral patterns of several influential similarity-based theories, such as the smooth generalization gradients characterized by Shepard's "Law of Generalization" (Fig. 3a) or Tversky's set-theoretic model of similarity[21]. And while other hybrid models advocated for a "separate-but-equal" approach[28], incorporating rules and similarity as separate

mechanisms, Bayesian concept learning represents a "unified" approach, suggesting rules and similarity are two sides of the same coin[13,32,41,42].

Box 1 | **Bayesian concept learning and Gaussian Process function learning**

Both Bayesian concept learning[5] and Gaussian Process function learning[43] exhibit a duality of rule- and similarity-based interpretations. This duality is achieved through Bayesian principles (see REF[12] for a review) to describe distributions over hypothesized category boundaries or over hypothesized functions.

Bayesian concept learning

The goal in concept learning is to predict whether a novel stimulus $\mathbf{x}_*$ falls within the same category C as previously observed examples $\mathbf{x}_i \in \mathcal{X}$. Bayesian concept learning[5] defines this probabilistically, by aggregating over all hypotheses h (i.e., category boundaries) consistent with $\mathbf{x}_*$ belonging to C:

$$p(\mathbf{x}_* \in C|\mathcal{X}) = \sum_{h:\mathbf{x}_* \in C} p(h|\mathcal{X}) \quad . \tag{1}$$

This represents a sum of posterior probabilities $p(h|\mathcal{X})$ for each hypothesis that encapsulates $\mathbf{x}_*$ (i.e., all category boundaries in Fig. 1d containing the query). The posterior probability for each hypothesis is defined by Bayes' rule:

$$p(h|\mathcal{X}) \propto p(\mathcal{X}|h)p(h). \tag{2}$$

Bayes' rule describes the integration of prior beliefs $p(h)$ and the likelihood of the data $p(\mathcal{X}|h)$ (conditioned on the hypothesis under consideration), in order to arrive at a posterior estimate (Eq. 2). In Bayesian concept learning, the prior is assumed to be uniform and the likelihood favors narrower hypotheses that are still consistent with the data (see REF[5] for details), as illustrated by the shading of hypotheses in Figure 1d. By aggregating over multiple rule-like hypotheses, Bayesian concept learning reproduces behavioral patterns characteristic of similarity-based theories, such as Shepard's Law of Generalization[3] and Tversky's set-theoretic model of similarity[21]. In practice, Eqs. 1-2 are typically intractable for most interesting problems and approximated via sampling[5].

Gaussian Process function learning

The goal in function learning is to learn a hypothesisized function h (e.g., a red line from Fig. 1h) mapping some input space $\mathcal{X}$ to real-valued outputs y (e.g., rewards). A prominent model of function learning is Gaussian Process regression[43], thus named because observations are assumed to be jointly Gaussian (see REF[44] for a tutorial). This allows us to describe a prior distribution over hypothesized functions, defined by a mean $m(\mathbf{x})$ and a covariance function $k(\mathbf{x}, \mathbf{x}')$:

$$p(h) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right). \tag{3}$$

While the mean function $m(\mathbf{x})$ is typically set to 0 without loss of generality[43], the covariance function $k(\mathbf{x}, \mathbf{x}')$ is defined by the choice of kernel function and provides an explicit similarity metric between any pair of inputs $\mathbf{x}$ and $\mathbf{x}'$. A common choice is the radial basis function (RBF) kernel:

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{-||\mathbf{x} - \mathbf{x}'||^2}{2\lambda^2}\right), \quad (4)$$

describing similarity as an exponentially decaying function of squared Euclidean distance (Fig. 3c). This enforces the inductive bias often present in similarity-based theories (Fig. 1g) that closer inputs are assumed to produce similar outcomes. The lengthscale parameter λ governs the rate that similarity decays over larger distances, with larger values corresponding to broader generalization and smoother functions (Fig. 3c). There are, however, a host of other kernels describing all manner of similarity relationships, including over graph structures[45–47] (Fig. 3d). Additionally, kernels are compositional[48,49], meaning multiple kernels can be combined to describe new composite relationships.

Gaussian Processes also use Eq. 2 to make posterior predictions, but here the Gaussian assumptions provide an analytically tractable solution[43]. These posterior predictions (conditioned on data $\mathcal{D} = (\mathcal{X}, \mathbf{y})$ of encountered inputs $\mathbf{x}_i \in \mathcal{X}$ and outputs $y_i \in \mathbf{y}$) are also Gaussian distributed, meaning the predicted outcome for any target input $\mathbf{x}_*$ is characterized by a posterior mean $m(\mathbf{x}_*|\mathcal{D})$ (expected outcome; blue line in Fig. 1h) and posterior variance $v(\mathbf{x}_*|\mathcal{D})$ (uncertainty; blue ribbon in Fig. 1h).

This core Bayesian framework—based on describing a distribution of hypotheses and adapting them to new data—has since proliferated computational theories across a wide range of phenomena, such as causal learning[50,51], word learning[52], structure induction[35], and learning compositional programs[1,36,37]. A distinct advantage of operating over rule-based representations is the ability to reason compositionally, by syntactically manipulating and combining multiple rules. Yet given an expressive hypothesis space, exact Bayesian inference is usually intractable, with most approaches relying on sample-based[5,35,53,54] or variational approximations[55,56]. Thus, it remains an open question how humans achieve the power and productivity of rule learning, but with limited cognitive resources[57,58].

Function learning

Generalization has also been studied in the domain of function learning, in the context of inferring the relationship between inputs and outputs[59–63]. For example, learning how spiciness (input) relates to one's enjoyment of a meal (output; Fig. 1e), or how the amount of time spent studying (input) predicts test

scores (output). Pioneering research by Carroll[64] used function learning to show that human generalization goes beyond merely predicting previously observed outcomes, in contrast to predicting only known category labels in the domain of concept learning. Rather, people can extrapolate beyond their previous experiences, generalizing not only to new inputs but also predicting new outputs (e.g., an off-the-charts food experience), Here too exists a parallel debate between rule- and similarity-based theories, which has culminated in hybrid formalizations[60,62,63].

Rule-based function learning

Many early theories of function learning are considered rule-based, by assuming people used a specific parametric model (e.g., a linear or polynomial function), and then learned by optimizing the parameters to best explain the data[64,65] (Fig. 1f). In function learning, rules thus correspond to a hypothesized relationship between variables, much like the law of gravity describes a polynomial relationship between mass and distance, or fitting a linear regression assumes a linear relationship. While rule-based methods could capture the systematicity of human extrapolation patterns[66], they lacked the flexibility of humans, who can learn to interpolate almost any function with enough training[67].

Similarity-based function learning

To account for the flexibility of human generalization, similarity-based models (Fig. 1g) of function learning were developed, often using artificial neural networks (ANNs) to encode the generic principle that similar inputs produce similar outputs[63,68]. The influence of previous observations thus decreases as a function of distance (arrows in Fig. 1g) to a given input, with nearby observations exerting a larger influence. But while these models could theoretically learn any function and matched human interpolation patterns, they lacked the inductive biases observed when humans extrapolate beyond observed data[49,69], for instance, the tendency to extrapolate functions linearly[70,71].

Hybrid function learning

To combine the rule-like systematicity of human extrapolation with the similarity-like flexibility of interpolation, hybrid function learning models were developed. One notable example is Gaussian Process regression[43,44] (Fig. 1h), which account for many empirical patterns of human function learning[60,72] while using similar Bayesian computations as hybrid models of concept learning models (Box 1). At the core of the Gaussian Process framework is a kernel

function, which provides a similarity metric with desirable mathematical properties[73]. Put simply, a kernel can take any two data points and provide a numerical quantification of their similarity. Take for instance the common Radial basis function (RBF) kernel, which quantifies similarity based on distance in feature space (Fig. 3c). Other common kernels can encode periodic relationships (periodic kernel), strictly linear functions (linear kernel), or even relational similarity on a graph-structure (e.g., a diffusion kernel[45,46]; Fig. 3d).

Based on a choice of kernel, Gaussian Processes describe a distribution over functions, where each function constitutes a candidate hypothesis (red lines in Fig. 1h). Given a set of observed data points, Gaussian Processes use Bayesian inference to compute a posterior distribution over functions, adapting the complexity to the available data rather than assuming a fixed level of complexity (as in rule-based accounts). This posterior distribution is also Gaussian, meaning that for any input, the predictions of the Gaussian Process can be characterized by a mean (i.e., expected outcome; blue line in Fig. 1h) and variance (i.e., uncertainty; blue ribbon in Fig. 1h).

Gaussian Process regression has been described as a hybrid model because it has both similarity- and rule-based interpretations[42,60]. The similarity-based interpretation is straightforward, since the kernel explicitly encodes similarity between data points, facilitating computations that operate directly over similarity representations. However, the framework also lends support to two rule-based interpretations. The first and most common interpretation is based on a mathematical property known as Mercer's theorem[74], which describes how any kernel can be decomposed into a sequence of basis functions[42,60], each corresponding to some abstract rule. Just as any color can be decomposed into red, green, and blue components, the basis functions that collectively constitute a kernel form the rule-like building blocks that allow Gaussian processes to express a potentially unlimited range of functions. Thus, inversely analogous to Bayesian concept learning, Gaussian Processes operate on similarity-based computations but provide equivalent rule-based interpretations. There is also a second rule-based interpretation, with more direct applications based on the compositionality of Gaussian Process kernels[48,49]. Any number of kernels can be combined via addition or multiplication operations to produce new kernels. Since each kernel can be seen as providing rule-like biases about the hypothesized form of a function (e.g., a

linear kernel for linear relationships, or a periodic kernel for periodic functions), compositional kernels thus allow for new compositional biases (e.g., a linear periodic relationship describing our alarming climate trends), similar to how rules can be combined to create new composite rules. As we describe later, since kernels can also be defined on graph structures (Fig. 3d), composing multiple kernels allows for aggregating over multiple hypotheses about the hidden structure of the environment (e.g., the floorplan of a foreign airport, or the layout of a city).

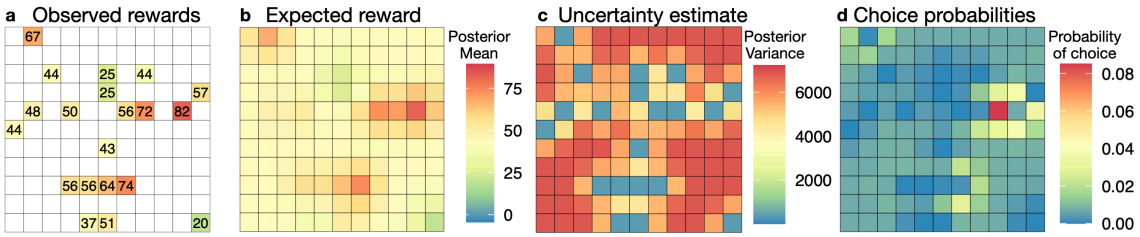

Fig. 2| **Value generalization and uncertainty-directed search. a)** Depiction of the spatially-correlated bandit task from REF[2]. Participants were incentivized to maximize reward by clicking tiles on a grid. Each click yielded normally distributed rewards, with nearby tiles having similar reward expectations based on an initially unknown spatial reward structure. Participants were given a limited search horizon that only allowed exploration of a small fraction of the environment. Thus, generalization and uncertainty-directed exploration greatly influenced performance (Box 2). **b-c**) Conditioned on the observations in panel a, the Gaussian Process model of value generalization makes predictions about expected reward and uncertainty (posterior mean and posterior variance, respectively). **d**) Reward expectations and uncertainty estimates are combined under UCB sampling, and transformed into probabilistic predictions of where the participant will search next using a softmax function (see Box 2 for details). All plots use participant parameter estimates from REF[2].

Value function learning in RL

Compared to concept learning, research on human function learning has received relatively less attention and produced fewer experiments. Yet there has been a revival of interest in understanding how humans learn functions, given the crucial contributions of value function approximation[23,75,76] in solving RL problems. In complex games such as Go[22]—but also in many real-world contexts[2,77]—the space of possible outcomes is too vast to be experienced exhaustively. Thus, many modern RL algorithms implement a form of function learning by learning a value function, mapping the space of possible actions to expectations of reward[15,22,23,75,78]. This estimated value function can then be used to generalize a limited number of experienced outcomes to a vast and

potentially infinite space of possibilities. Given the significance of value generalization for RL and AI, there is a great impetus to better understand how humans generalize value in complex tasks with sparse data.

Recently, several studies have studied human value generalization in a variety of RL contexts[2,7,18,45,79–81], often using the same Bayesian hybrid models (i.e., Gaussian Processes) as in classic function learning paradigms. In these RL tasks, participants can leverage features of the environment, such as spatial location[2] or abstract features[80] to make predictive generalizations about the value of novel actions or stimuli (Fig. 2). Instead of exploring randomly, predictive generalizations efficiently guide exploration towards promising choices (Box 2), with Gaussian Processes providing the best account of human generalization in a variety of spatial[2,82,83], conceptual[80,81,84], risky[79], and graph-structured[45] RL tasks.

Given similar historical developments in concept and function learning, there is much to be gained from integrating theories of generalization across domains. For instance, the Bayesian concept learning framework[5] has made substantial contributions towards tackling problems with increasingly structured hypothesis spaces. Similar methods that rely on rule-based computations have been developed for inferring graph-structured representations implicit in data[35,51] or learning complex behavioral programs built upon compositional primitives[1,36,37]. These successes suggest important opportunities in the domain of function learning and value generalization, with early work showing that function learning on discrete graph-structures[45] and abstract conceptual stimuli[80] operate using similar principles as human function learning on spatial features.

In the following sections, we re-examine the mechanisms of similarity- and rule-based approaches to generalization and discuss how modern theories greatly expand the scope of what “similarity” and “rules” can describe. While previous hybrid models have both similarity- and rule-based interpretations, the representational and computational commitments fall predominantly on one side or the other, conferring respective advantages and disadvantages. We then provide perspectives on how to forge new hybrids, combining similarity- and rule-based mechanisms in new ways to expand our understanding of how humans generalize in increasingly complex and structured tasks.

Box 2 | **Leveraging uncertainty for active learning**

A fundamental challenge for efficient learning is to determine what information should be acquired given current beliefs (i.e., active learning[85,86]). Bayesian models address this challenge by making predictions accompanied by a quantification of uncertainty. These uncertainty estimates support efficient exploration[2,87] and faster learning of concepts[88–90], functions[91,92], and rewards [2,80–83], by indicating where exploration would be most useful.

Figure 2 illustrates how uncertainty and generalization can be leveraged to efficiently explore large problem spaces[2,80,82,83]. In a spatially correlated bandit task (Fig. 2a), participants are given the goal of maximizing rewards by iteratively selecting options (tiles), each yielding stochastic rewards. Typically, the number of options greatly exceeds the search horizon, such that the agent can only experience a small fraction of the total options, thus facing a stark exploration–exploitation dilemma[93]: should they explore uncertain options to learn about novel opportunities or exploit known high-value options to maximize immediate rewards? However, as in many real-world problems[2,77,94], the environment is spatially-structured, such that nearby tiles yield similar rewards. Thus, generalization and uncertainty estimates can guide exploration towards both promising and informative regions of the search space.

Applying Gaussian Process regression as a model of value generalization yields posterior predictions about reward expectations (mean; Fig. 2b) and the associated uncertainty (variance; Fig. 2c). *Upper Confidence Bound* (UCB) sampling[2,87,95] offers an effective solution to the explore-exploit dilemma[96] by quantifying the value of any target option $\mathbf{x}_*$ using a weighted sum of reward expectations and uncertainty estimates:

$$\mathrm{UCB}(\mathbf{x}_*) = m(\mathbf{x}_*|\mathcal{D}) + \beta\sqrt{v(\mathbf{x}_*|\mathcal{D})}, \quad (5)$$

The coefficient β is the exploration bonus, governing how much exploring uncertain options (i.e., uncertainty-directed exploration) is valued relative to exploiting immediate rewards. In other words, UCB sampling optimistically inflates reward expectations by the attached uncertainty, providing an informational bonus for uncertain options. Finally, UCB values can be converted into probabilistic predictions about which option a participant will search next (Fig. 2d) using a softmax function:

$$p(\mathbf{x}_*) \propto \exp(\mathrm{UCB}(\mathbf{x}_*)/\tau), \quad (6)$$

where options are selected proportional to their UCB value. Here, τ is the temperature parameter, governing the level of randomness in these predictions, providing a source of random (undirected) exploration. Both uncertainty-directed (β) and random exploration (τ) play an important and dissociable role in human exploration[2,87,95,97], with different neural signatures[98] and distinct developmental trajectories in exploration and generalization behavior[82,83,99].

Similarity for generalizing outcomes

Common across domains is the notion that stimuli with similar features or occurring in similar contexts are more likely to belong to the same category or yield comparable outputs or rewards. Thus, it comes at no surprise that similarity-based mechanisms are ubiquitous in many psychological theories[3–5,10,17,21,100], with the quantification of similarity providing a powerful tool for generalizing from prior experiences to novel situations.

Similarity as distance in psychological space

Early theories of similarity-based generalization introduced the notion of a psychological space[101,102], where stimuli are embedded as geometric coordinates and a measure of distance (e.g., Euclidean distance) serves to represent the (dis-)similarity between stimuli. The most influential example is Shepard's "Law of Generalization"[3], which used confusability (i.e., the probability of responding to stimulus $\mathbf{x}$ when shown stimulus $\mathbf{x}'$) to construct a psychological space using Multidimensional Scaling[103,104]. In Shepard's psychological space, stimuli producing similar responses are embedded in similar locations, such that the same unit of distance in any direction corresponds to the same level of generalization. Stimuli located further apart in psychological space are thus less likely to yield the same response, becoming exponentially less likely as their distance increases (Fig. 3a). At the core of Shepard's theory is the assumption that representations about categories or natural kinds correspond to a "consequential region" in psychological space (Fig. 3a inset). Generalization thus arises due to uncertainty about the extent of these regions. As the distance between stimuli $\mathbf{x}$ and $\mathbf{x}'$ increases, they are less likely to belong to the same region and therefore less likely to produce similar outcomes, thus producing a smooth gradient of generalization.

Similarity in Episodic RL

More recently, related mechanisms of similarity-based generalization have been incorporated into RL models, spurred by advancements in value-function approximation[15,22]. Consider the Episodic RL framework[17,100], which stores episodic memories about previously encountered stimuli and their associated rewards (Fig, 3b inset). To predict the value of some novel stimuli, one first computes similarity to each previously encountered "episode", weighting these similarities by the corresponding reward values, and then summing it up

(Fig. 3b). Thus more similar episodes exert more influence in how their rewards generalize to the novel stimuli.

This use of a similarity-weighted sum for generalization is reminiscent of classic exemplar-based theories of concept learning[25], while also having direct equivalencies to computational methods used in function learning. For instance, when using an Radial Basis Function (RBF) kernel (Fig. 3c) as the similarity metric, Episodic RL is equivalent to an RBF network, which has featured prominently in machine learning approaches to value function approximation[15,105] and as a theory of human generalization in the visual and motor systems[8]. Moreover, an RBF network is in turn equivalent to the posterior mean of a Gaussian Process[44,105], thus establishing a direct relationship between Episodic RL and Gaussian Process function learning. However, a crucial difference is that the Gaussian Process—being a Bayesian model—also makes predictions with uncertainty, which play an essential role in describing human choice behavior[2,95,106], with distinct benefits for efficient learning and exploration (Box 2).

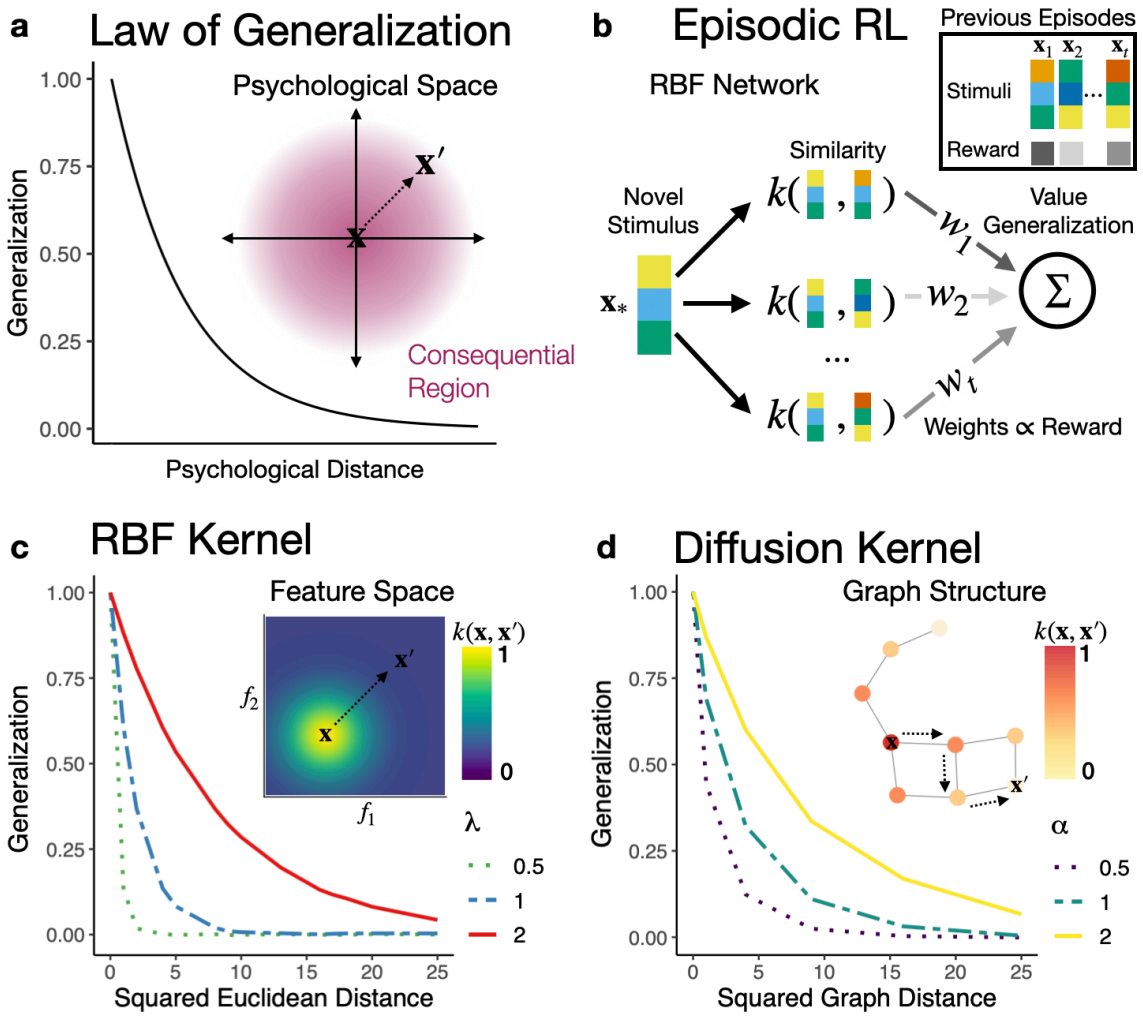

Fig. 3 | **Similarity. a)** Shepard's Law of Generalization[3] describes generalization as a function of distance (dashed arrow) between stimuli in psychological space (inset). The smooth, exponential gradient of generalization arises due to uncertainty about the extent of a consequential region, with more distant stimuli less likely to belong to the same region. **b)** The Episodic RL framework[17,100] describes how the value of some novel stimulus $\mathbf{x}_*$ can be predicted by comparing its similarity to all previously encountered stimuli (i.e., episodes). The degree of similarity is then weighted by the rewards for each previous episode, and summed to predict the value of the new stimulus. When using an RBF kernel as a measure of similarity, episodic RL is equivalent to both an RBF network[15,105] and the posterior mean of a Gaussian Process (but without uncertainty estimates). **c)** An RBF kernel provides a similarity metric as a function of the squared Euclidean distance between stimuli in feature space (inset: dashed arrow), producing similar generalization gradients as Shepard's model (quantified using Pearson correlation between expected outputs). The lengthscale parameter λ governs the rate which generalization decays as a function of distance. **d)** In structured environments, a diffusion kernel[45,46] offers an analogous similarity metric based on the connectivity structure of a graph, where the diffusion parameter α governs the rate that previous observations "diffuse" over the graph. In the limiting case of an infinitely fine lattice graph, the RBF and diffusion kernels are equivalent[46].

Similarity for capturing relational structure

Historically, similarity-based generalization has relied on comparing various stimuli based on their features, such as predicting the quality of a hotel based on location, amenities, or cost. However, recent approaches have derived similarity-based representations for capturing relational structure, for instance, comparing the similarity of individuals based on the connections of their social network[107,108] or comparing locations based on the transition dynamics of the environment[16,109]. You might be more likely to encounter similar people at two distant city hubs connected by a high-speed rail, compared to relatively local but isolated communities connected by winding country roads.

In environments that reflect such graph-structured relationships, a diffusion kernel[45,46] (Fig. 3d) offers an analogous similarity metric (among other graph kernels[47,110]), usable in a Gaussian Process framework. The diffusion kernel is directly computable from the (weighted) adjacency matrix of a graph. Generalization (of value) from observed to unobserved nodes in a graph is thus modelled as a heat diffusion process over the connectivity structure of the graph. Intuitively, just as heat from one room diffuses to another more quickly if they share a doorway vs. a sealed wall, so too is generalization influenced by the connectivity of the environment. The diffusion kernel thus extends similarity-based generalization to domains with rich relational structure, where outcomes can be predicted based on the connectivity structure of a graph representation rather than comparing features in some psychological space.

A rational basis for similarity

Altogether, similarity—whether defined over features or relational structure—provides a powerful tool for generalization. Yet we cannot simply go out into the world to measure how similar things are to one another. Consider how naturalistic stimuli have a host of different features and relationships, offering a potentially unlimited number of ways by which similarity can be computed[19]. Should an apple be compared to an orange on the basis of color, shape, taste, or country of origin? Thus, one must specify with respect to which features (or via which relationships) the stimuli are being compared[20]. This is often dependent on the underlying context: when at a fruit orchard, color might provide a useful comparison on the basis of ripeness, whereas at a customs office, country of origin is more relevant for determining the amount of tax to levy.

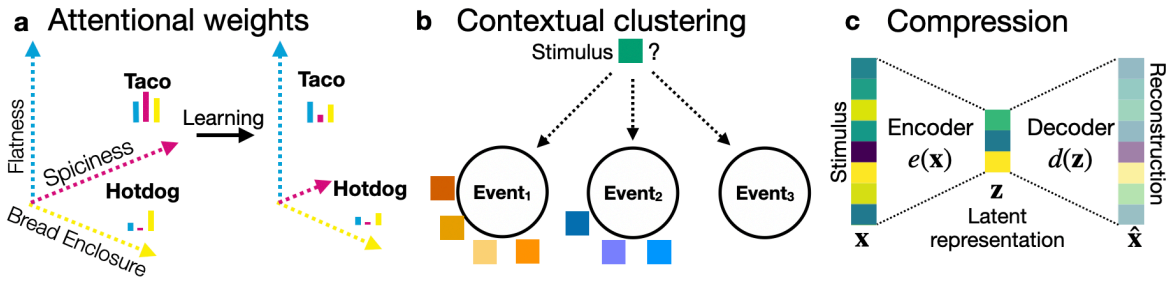

Fig. 4| **Attention, context, and compression. a)** Attentional weights define the relevance of each feature for generalization, and can be learned via associative mechanisms[111]. **b)** Contextual clustering can be used to infer latent event representations, to contextualize which features are relevant in the current context (e.g., different weights in different events). A common method is to use a Chinese Restaurant Process (CRP)[112] to group similar stimuli into common clusters, not unlike seating banquet guests at compatible tables. **c)** Compression provides a mechanism for constructing lower dimensional feature representations, rather than simply weighing and filtering features. One method is to use a Variational Autoencoder (VAE)[113], which simultaneously learns an encoder (mapping stimulus $\mathbf{x}$ into a low-dimensional latent variable $\mathbf{z}$) and a decoder (mapping $\mathbf{z}$ to some reconstruction of the stimulus $\hat{\mathbf{x}}$). Generalization can thus be performed on the latent variable $\mathbf{z}$.

The endless ways in which different stimuli can be compared has led to the criticism that similarity is too flexible to be useful as a concept in psychology[114]. The context-dependent nature of human similarity judgments is a key reason why we are often observed to violate logical axioms, such as the law of triangle inequality[21]. To adapt a famous example[21], people often judge Jamaica to be similar to Cuba (because of geography) and Cuba to be similar to Russia (because of political history). However, when directly comparing Jamaica and Russia to one another, they are judged to be quite dissimilar. Thus, by adding a comparison to Cuba, the similarity between Jamaica and Russia paradoxically increases.

Thus, even though similarity is ubiquitous in psychology[20,39], it is not always clear which metric or representation to use[4,21,115] for a given context. And while experimenters can exogenously define the relevant features of a task, this is not always possible in more naturalistic environments. In the following sections, we review recent psychological theories that provide potential solutions to this historic critique of similarity being too flexible, by anchoring similarity representations in a rational rather than arbitrary basis. These theories

have the potential to expand the scope to which we can apply similarity-based theories of generalizations.

Attentional weights

To account for contextual effects, many early models of concept learning included attentional weights[25,34] used to distort psychological space to prioritize similarity comparisons along relevant feature dimensions (Fig. 4a). While these early models simply assumed that these weights were given, recent theories of selective attention[116–120] have provided insight into the cognitive processes and control mechanisms that describe how people learn relevant feature weights through experience. These theories of selective attention largely coincide with rational models of attention[116,117,121], where over the course of learning, attention shifts towards more reward-relevant features[122,123], gradually filtering out irrelevant features[111,117]. This perspective is consistent with the modulation of attention incurring cognitive costs[124,125], where the costs of attending to more (or less) features needs to be traded off against the benefits of increased performance[126,127]. One example of selective attention is the feature-based RL model developed by Niv and colleagues[111], where the familiar computations of prediction-error learning[15,128] are used to describe how people learn which features to attend to, by updating attentional weights based on the difference between predictions and observed outcomes.

Contextual clustering

Yet, different features are relevant for different contexts. A computer, a cat, and a passport might seem quite dissimilar in everyday contexts, but in the event your house catches fire, they might be highly similar in the context of what you should save[129]. Thus, inferring context from our continuous stream of sensory data can provide guidance about which features are currently relevant, allowing for different attentional weights to be learned specific to each context (Fig. 4b). Entering the threshold of a library might make you more sensitive to speaking volume, whereas recognizing you are at a fancy dinner may make you more conscious of the shape and order of your cutlery.

While contextualized attention was already incorporated into classic concept learning models[25], more recent models describe the process behind how people infer latent context[112,130] rather than simply assuming context is given. This process of inferring context is often operationalized as a clustering problem, where stimuli we experience are grouped together into common

clusters, or "event representations"[131–133]. One widespread approach is to use a Chinese Restaurant Process[133–135] (CRP), which resembles how one might design a seating plan for banquet guests, by seating diners at compatible tables. Similar stimuli or experiences are thus grouped together, while dissimilar experiences are segmented and separated. Thus, as we transition from one inferred context to another, we can adaptively switch between different attentional weights, capturing the contextual-basis for how to represent similarity.

Compression

Lastly, distinct from merely filtering and weighing features, another important mechanism is compression[136–140], offering more radical transformations of natural stimuli features into new, low-dimensional representations. Similarity-based generalization could thus operate on these compressed representations rather than directly on the natural or even attention-weighted features of the stimuli. Given limitations on representational bandwidth[141,142] or working memory[143], building a reduced dimensional representation of stimuli (i.e., compressing irrelevant feature dimensions[140]) is not only relevant for predicting rewards, but also for protecting against transmission errors[144] when exchanging information between cognitive systems[145]. Indeed, rational solutions to information transmission under capacity limits[141,146](similar to the principles behind image or video compression), have been shown to recover the same exponential generalization gradients as Shepard's Law of Generalization[138] (Fig. 3a).

While many mechanisms of representational compression[136,147–149] have been proposed, one promising candidate is the Variational Autoencoder (VAE) framework[113,137,150,151](Fig. 4c). VAEs simultaneously learn an encoder and decoder (often as ANNs) mapping the features of input stimuli $\mathbf{x}$ onto a low-dimensional latent variable $\mathbf{z}$. The encoder is an inference model (inferring latent variable $\mathbf{z}$ from stimuli $\mathbf{x}$) and the decoder is the (approximately) inverse generative model, which reconstructs an approximation of the stimuli $\hat{\mathbf{x}}$ from $\mathbf{z}$. The compressed representation $\mathbf{z}$ attempts to capture the relevant characteristics of the stimuli, where reconstruction error (i.e., difference between the original stimuli $\mathbf{x}$ and the reconstruction $\hat{\mathbf{x}}$) provides the training signal.

Expanding the scope of similarity

In sum, similarity is no longer too flexible of a concept to be useful[19,114]. By incorporating modern accounts of how similarity representations are actively constructed—through selective attention, contextual clustering, and compression—similarity representations can be anchored on rational principles rather than being seen as arbitrary. These rational principles have the potential to expand the scope of similarity-based models, potentially allowing us to describe human generalization in more naturalistic, context-dependent settings. Yet are these mechanisms enough to describe the full depth of human generalization?

Rules for learning structure

Human generalization is anything but shallow. Learners are able to leverage rich relational structures[149,152–155], informed by domain theories and background knowledge[147,156,157]. Consider a chef figuring out how to substitute a missing ingredient in a recipe or an archaeologist deciding where to dig for ancient settlements. Generalization in both domains is informed by a rich tapestry of hidden structure—about the interactions between different foods and cooking techniques or about the dynamics of human migration patterns interwoven with geomorphological changes in the landscape. Understanding human generalization in richer and more naturalistic settings requires integrating theories about how people learn and incorporate structure into their generalization.

So far we have described how similarity-based methods are capable of capturing generalization over feature-based or relational structures[45,158]. We have also described how rational processes of attention, context, and compression, can be used to "warp" similarity representations to capture task-relevant aspects for generalization. Yet while similarity-based models demonstrate a high degree of flexibility in being able to capture a wide range of structures, they may not be the best suited for learning structure in the first place. Given the complex and productive hypothesis space of potential structures, tractable inference might require the ability to generate new hypotheses in an efficient and perhaps compositional manner. Just as a tent cannot stand without tentpoles, human generalization may require a more rigid

framework of rule-like structure learning on which to drape the canvas of similarity.

Rule-based mechanisms are a promising candidate for describing human structure learning and have a storied history in research on generalization[30,159,160] (see REF[29] for a review). Whether taught pedagogically (e.g., “i before e except after c”) or learned through experience (e.g., “don’t talk loudly in the library”), rules represent explicit hypotheses about regularities of the environment extracted from data[161]. This ability to reason about and use rules thus unlocks an unrivaled capacity of human intelligence, since rules allow for compositional and syntactic manipulation. Indeed, the power of logic and mathematics can be thought of as nothing more than the manipulation of syntactic rules[162].

In concept learning, rules can represent hypotheses about the boundaries between categories, while in function learning, rules can represent hypotheses about the (parametric) relationship between inputs and outputs. Signatures of rule-based mechanisms can also be seen in theory-based RL[163–165], where agents generate hypotheses about the underlying rules governing its environment (e.g., “keys open doors, but only if the colors match”[147] in game environments) to inform learning and exploration. Rules thus represent different forms of structural regularities in the environment, with the essential characteristic that they can be compositionally combined to craft new hypotheses.

While rule-based mechanisms can facilitate rapid generalization (e.g., picking up the tipping customs or driving rules of a foreign country), rules can also be inflexible and difficult to learn. Inflexibility is most characteristic of early rule-based theories[24,30,59,64], which proposed a single hypothesized rule at a time. However, this critique largely vanishes when proposing multiple hypotheses or a distribution over rules within a Bayesian framework[2,5,37,60], although criticisms of tractability are substituted in its place[166]. Indeed, given the compositional and potentially infinite hypothesis space of potential rules that could be learned, it remains an area of intense interest to understand the mechanisms that allow humans to learn complex rules with cognitively limited resources[36,167–170]. Nevertheless, rule-based mechanisms may be a key component in understanding how humans infer structure from limited experience. We next

review candidate theories of structure learning and explore how to build the backbone for more complex and compositional theories of generalization.

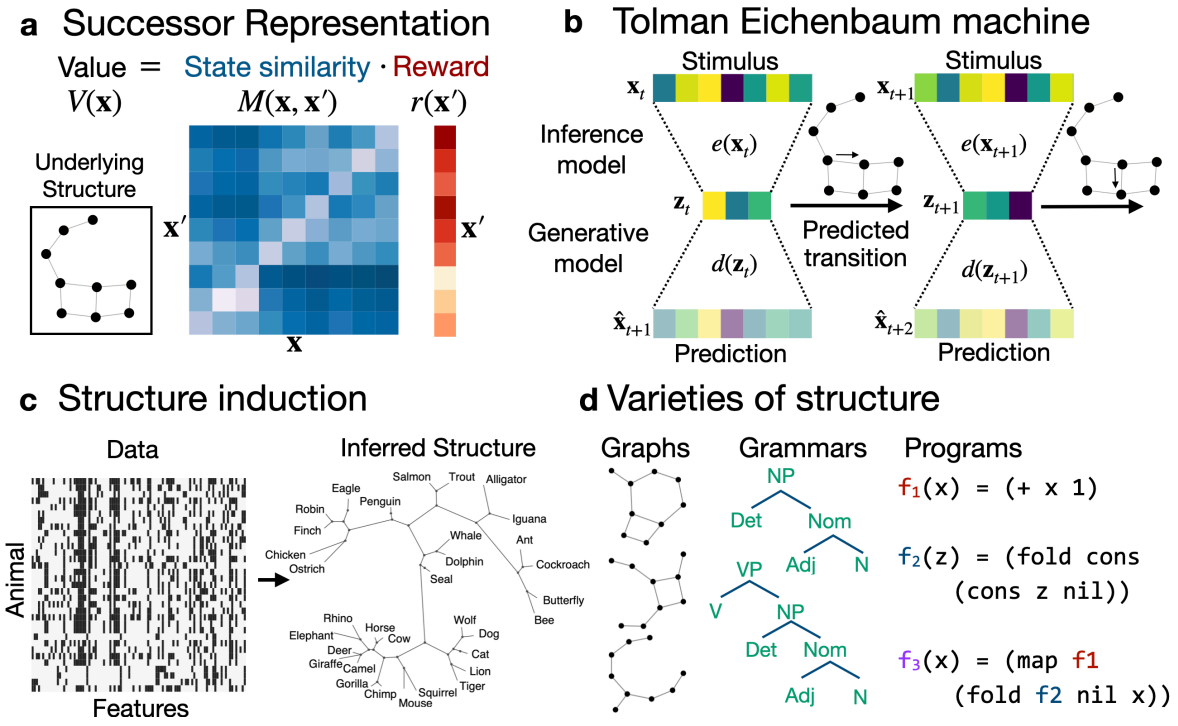

Fig. 5| **Inferring structure. a)** The Successor Representation[171] (SR) defines a decomposition of a TD-learning[15] value function V(**x**) into a similarity matrix *M*(**x**,**x**′) based on expected future state transitions, and the singular rewards of each state *r*(**x**′). The state similarities are a function of the underlying structure (left) and the agent's policy. **b)** The Tolman Eichenbaum machine[56] (TEM) uses a VAE framework (c.f., Fig. 4c) to learn latent structure by jointly training an inference model (encoder) and a generative model (decoder) to relate sensory data $\mathbf{x}_t$ to a set of latent variable $\mathbf{z}_t$ (differentiated as structural and conjunctive codes; see REF[56] for details). The latent variable $\mathbf{z}_t$ captures the temporal dynamics of the environment by implying hidden structure, and use path integration to predict future states $\mathbf{x}_{t+1}$ based on current sensory data $\mathbf{x}_t$and the inferred dynamics of $\mathbf{z}_t \rightarrow \mathbf{z}_{t+1}$. **c)** Bayesian structure induction[35] uses Baysian principles (Box. 1) to take relational data (e.g., animals and their shared features) in order to infer the underlying structure that gave rise to the data (e.g., a taxonomy). **d)** Structure induction is not limited to learning graph-structures, with similar frameworks applied towards learning grammatical structures[172] and programs[1,36,37].

Inferring structure

Similarity-based methods are capable of performing generalization over structured representations of the environment, for instance, using the Gaussian Process framework and a diffusion kernel over a proposed graph structure[45,158]. However, an important gap in our understanding is how structure can be learned through experience, rather than explicitly provided. In this section, we

review promising theories of latent structure learning and describe how they can be used in tandem with similarity-based mechanisms to describe human generalization in exceedingly complex problems.

Cognitive maps

One classic framework for representing environmental structure is by building a cognitive map[80,173,174] of the environment. In 1948, Tolman showed that rather than merely learning through stimulus-response associations, rats could rapidly generalize to new situations (e.g., choosing the second shortest path in a maze when the shortest path is blocked) and to new goals (e.g., efficiently navigating to food rewards placed in novel locations of a familiar maze)[173]. These results suggested the rats had internalized a representation of the environment. Today, the notion of a cognitive map is grounded in neural evidence (in humans and other animals) relating the activity of specialized cells in the hippocampal-entorhinal system to computations facilitating navigation and self-location, such as encoding spatial orientations, boundaries, and distance to objects (see REFS[175–177] for reviews). As Tolman originally speculated, cognitive maps are not only restricted to representing spatial structure. Rather, the same neural machinery used for spatial navigation also encodes relational and structural knowledge across a wide range of domains, including social relationships[108], smells[178], abstract visual features[179], and the connectivity of hidden graph structures[153].

One influential account of how the hippocampal-entorhinal system learns a cognitive map is the Successor Representation[16,154,171] (SR; Fig. 5a). The SR describes how the value function from a Temporal Difference (TD) learning[15] model can be decomposed into a similarity matrix and the singular rewards of each state. The similarity matrix quantifies the similarity between each pair of states based on expected future state transitions, influenced by both the connectivity of the environment and an agent's behavioral policy (i.e., how the agent moves around in the environment). The generalizations of the SR—taking the familiar form of a linear combination of state similarities and reward observations (not unlike Episodic RL; Fig. 3b)—can capture the underlying connectivity structure, with stronger generalizations between well-connected states. Related methods using kernel similarity[45,180] (Fig. 3d) rather than SR similarity, operate on similar principles, with exact equivalencies in special cases[109].

The SR offers a candidate mechanism for how the structure of the environment can be learned through experience (i.e., "on-policy" learning). This on-policy method learns the similarity matrix (encoding environmental structure) using the familiar computations of prediction-error learning[15,128], but rather than predicting rewards, future state transitions are predicted[171,181]. However, some of the most convincing demonstrations of the SR as a model of the hippocampal-entorhinal system are based on "off-policy" methods. These off-policy methods sidestep the problem of how humans learn latent relational structure by simply assuming a random policy operating over infinite time, allowing the SR similarity matrix to be computed analytically[16,182]. One motivation for using off-policy methods is that on-policy structure learning is slow, requiring exhaustive exploration of an environment before an accurate model of the environment can develop. In contrast, humans can leverage previously encountered structures to rapidly learn new structures with few experiences[155]. For instance, consider how you might deftly navigate a foreign airport based on intuitions about previously experienced airport layouts, or how you might transfer domain knowledge about bread baking to a new problem such as making steamed buns. In addition to these limitations in efficient structure learning, the SR also lacks access to active learning mechanisms that play a key role in human exploration (Box 2), since it only makes point estimates about expected reward, without quantifying uncertainty (but see REFS[155,183,184] for Bayesian extensions or REF[185] for count-based approximations of uncertainty). Altogether, the SR provides an elegant and simple theory of structure learning within the RL framework, but may fall short of explaining the full efficiency with which humans learn relational structure.

Another recent theory of relational structure learning is the Tolman Eichenbaum Machine[56] (TEM; Fig. 5b), which provides a model of many specialized cells in the hippocampal-entorhinal system. Put simply, TEM combines sensory data with internal representations of the environment to make predictions about where an agent will be in some future point in time, somewhat similar to how the SR represents structure using future state expectations. Key to the predictions made by TEM is the novel combination of path integration[186] and conjunctive memory[187]. Path integration allows the TEM to generalize about the structure of the environment, where prior experience within the same class of structures can inform predictions about novel state

visitations. For instance, we can recognize walking one block west then one block north will end up at the same location as walking one block north then one block west in a gridded city, even if traveling along novel streets. Conjunctive memory allows the TEM to develop factorizable encodings for spatial ("where am I?") and sensory ("what do I see?") representations, allowing for structural knowledge (but not irrelevant sensory representations) to be transferred across environments.

Formally, the TEM architecture resembles a VAE (Fig. 4c), combining a generative model of the environment with an inference model. The inference model learns to encode sensory observations as a latent variable (see REF[56] for details), while a generative model of state transitions, uses the dynamics of the latent variables to predict future observations. Just as the SR implies structure by representing similarity between states based on expected transitions, the TEM implies structure by representing the temporal dynamics of the environment, encoded as a latent variable.

While not a RL model, the TEM nevertheless offers a biologically plausible mechanism for unsupervised structure learning, which could inform downstream generalization of value or concepts by encoding a representation of the structure of the environment. And although the TEM is capable of transferring learned structure to new environments, it still lacks the ability to infer entirely novel structures. In contrast, humans are able to reason compositionally about new relational structures that they have never experienced before. Consider how you can imagine novel food combinations that have never been observed (e.g., tea-flavored jelly[188] or broccoli-flavored ice cream[180]) or novel configurations of previously encountered structures, such as predicting where your gate might be when racing through a foreign airport to catch a connecting flight. This ability to efficiently learn novel structures might require more explicit, rule-like mechanisms for generating hypotheses about structure, which we turn to next.

Structure induction

A different approach to structure learning is based on Bayesian structure induction[35,50,158] (Fig. 5c). Given some relational data, such as the features of different animals, Bayesian structure induction can infer the graph structure that gave rise to the data. The computations (see REF[35] for details) share a similar

mathematical framework as Bayesian concept learning (Box 1), based on describing a distribution of hypotheses, which adapts to the observed data. But instead of defining hypotheses about category boundaries, Bayesian structure induction operates on hypotheses about structural configurations (i.e., different graph structures). A prior over hypothesized graphs encodes a preference for simpler structures, with each hypothesis weighted by its likelihood of generating the observed data. Notably, a common method used to approximate the likelihood function is to sample from a Gaussian Process[35,158], where each candidate structure is used to parameterize a graph kernel[110] and a Gaussian Process is used to generate simulated observations (i.e., the matrix in Fig. 5c). High similarity between generated and observed data corresponds to a higher likelihood for the hypothesized structure. Thus, Bayesian structure induction frequently relies on the same computations using Gaussian Processes as used in function learning or value generalization[45,158].

A key advantage of performing computations over rule-like hypotheses about structure is the ability to reuse previously learned "schemas"[155,189,190], which provide scaffolding for learning new structures. For instance, given that all airports generally contain terminals, security gates, and waiting areas (among other components), we can more efficiently infer the likely structure of a novel airport, by recombining and reusing these previously learned substructures[36,55,191]. While the reuse of previously learned schemas was not included in the original Baysian structure induction framework[35], there have been several adaptations that have moved in this direction. Most notably, recent approaches to program induction[1,36,37] (e.g., learning programs for creating hand drawn sketches) have profited greatly by including mechanisms for reusing previously learned primitives to aid in compositionally generating new structures. Just as a graph structure provides a compositional representation of environmental structure (capturing diverse data such as taxonomies, floorplans of a building, or grammatical parse trees), programs offer yet another representation, where functions can be applied recursively to different variables to create new higher-level functions (Fig. 5d).

Combining rule-based structure learning with similarity-based generalization

Human generalization is much deeper than just comparing features on the basis of similarity. Rather, generalization also depends on relational structure and temporal dynamics of the environment, which are often hidden

and need to be inferred. Current theories of structure learning are only beginning to capture the flexibility and speed with which we can learn and leverage structure in the environment for generalization, despite possessing limited computational capacities. An important implication of these theories is that after learning a structured representation of the environment, we can use similarity-based computations over this structure to perform generalization in increasingly more complex problems. With rule-based mechanisms providing the tentpoles and similarity-based mechanisms providing the canvas, a picture of human generalization as both rigid and flexible begins to emerge—both compositional and efficient.

Summary and Future directions

We have shown how a Bayesian framework of function learning and value generalization using Gaussian Process regression[43,44,60] extends the spirit of previously successful models of concept learning[5], and explains human behavior in a wide range of supervised[45,49,60,91] (Fig. 1) and RL tasks[2,45,79–83] (Fig. 2). This Gaussian Process model uses kernel similarity—computed over features (Fig. 3c) or relational structures (Fig. 3d)—to generalize about novel outcomes and quantify uncertainty to guide exploration towards promising options. At the heart of this framework are two mechanisms: similarity for inferring novel outcomes and rules for learning structure.

Similarity provides a simple yet flexible mechanism for generalization through comparisons to previous experiences. Although similarity-based methods have historically been criticized as too flexible, recent theories of attention, contextual clustering, and compression help overcome these challenges by grounding similarity representations in a rational rather than arbitrary basis (Fig. 4). However, to explain the full power of human generalization, we may still need a more rigid framework for inferring hidden structure on which to support similarity-based computations.

Rules provide a powerful mechanism for extracting structure from the environment, unlocking the ability to compositionally combine multiple rules or substructures to generate an infinitely productive space of potential hypotheses. Yet, the infinite productivity of compositional representations also makes inference intractable for most interesting problems, leaving open questions

about the mechanisms that humans use to learn structure with limited computational resources (Fig. 5).

Having surveyed the past and present, we now turn our attention to the future. Psychology—like many sciences—is often at its best when combining different approaches. The ongoing success of hybrid models hints towards the importance of accommodating both rule- and similarity-based representations in describing human generalization. Yet, each approach makes computational commitments to a specific representational format, offering distinct advantages. Rules unlock compositionality, facilitating generalization and inference about novel structures, which is exemplified in Bayesian structure induction[35,51] and program induction[36,37]. Similarity provides a flexible and efficient approach to generalization, relating new situations to prior experiences[80,100,105], and leveraging relational knowledge when the underlying structure is known[45,158].

We propose that rather than accepting a duality of interpretation as the final synthesis of rule- and similarity-based mechanisms, future models of generalization could provide a more complete unification, using each mechanism to their strengths. We have advocated for Gaussian Process function learning as a candidate model of human value-generalization in many domains, where the kernel provides a similarity metric based on a given representation of the environment. Yet we currently lack a model that simultaneously infers structure while performing predictive generalization. Since Gaussian Processes play a key role in the computations of Bayesian structure induction[35], a future model could use Gaussian Processes to simultaneously perform inference over candidate structures and generate predictions about novel outcomes. Rule-based mechanisms can be used to propose hypotheses about structure (e.g., proposing different graph configurations by leveraging previously learned schemas[35,36,170,191,192]). Each hypothesized structure can then be used to parameterize a graph kernel, where similarity-based mechanisms can be used to both predict new outcomes and to evaluate the likelihood of a given hypothesis (as in REF[35]). Lastly, Bayesian principles will allow us to describe a distribution over hypotheses, which adapt to the observed data. With rules providing the structure and similarity providing the canvas, generalization combining both mechanisms can achieve both flexibility and efficiency.

One candidate algorithm to implement this proposal is a particle filter[193,194]. A particle filter uses a finite set of hypotheses (i.e., particles), which are refined and updated based on new data to provide an approximation to Bayesian inference. Here, particles can represent different hypotheses about latent structure—for instance, a specific graph configuration. When encountering new data, particles are reweighted by their likelihood (e.g., by generating simulated observations from a Gaussian Process[35]), and then resampled in proportion to this likelihood. Thus, inaccurate hypotheses die out, while more accurate hypotheses proliferate, and are refined. To take into account uncertainty about the underlying hypotheses (i.e., variance across particles), one could tap into the compositional nature of kernels[43,49], and combine all current hypotheses into a composite kernel (i.e., averaging across particles). Such an approach could propagate uncertainty about the underlying structure through to uncertainty about potential outcomes, facilitating active learning at both levels.

Since Tolman first proposed the notion of a cognitive map[173], psychologists and neuroscientists have sought to understand the mechanisms humans use to extract structure from the environment to inform generalization. Modern approaches like the SR[16,171] and TEM[56] offer promising explanations for the specialized cell-types that facilitate navigation and planning in spatial and non-spatial environments. Yet more studies are needed to demonstrate to what extent these models can match the human ability to compositionally reason about novel structures[1,36,37], by reusing[36,155] and recombining[180,188,195] previously learned schemas. Ultimately, future approaches will need to further integrate the compositional power of rule-based structure learning with the flexibility of similarity-based methods to generalize in both a structured and efficient manner.

Author contributions

All authors contributed to the conception and writing of the paper. CMW wrote the first draft and created the visualizations.

Competing financial interests

None of the authors reported competing financial interests.

Acknowledgements

CMW is supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A and funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy–EXC2064/1–390727645.

We thank Lee Sharkey, Tankred Saanum, and Philipp Schwartenbeck for feedback on an early draft of the manuscript. We are also grateful to Konstantin Genin, Ian Sellers, Mihály Bányai, Mani Hamidi, and Nikola Milosevic for helpful discussions.

References

1. Lake, B. M., Ullman, T. D., Tenenbaum, J. B. & Gershman, S. J. Building machines that learn and think like people. *Behav. Brain Sci.* **40**, e253 (2017).
   * A foundational paper that extends Bayesian concept learning to compositional program learning, capturing a key signature of human generalization
2. Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D. & Meder, B. Generalization guides human exploration in vast decision spaces. *Nat Hum Behav* **2**, 915–924 (2018).
   * An important demonstration of how Gaussian Process function learning captures human generalization in RL, featuring a large-scale model comparison and experiments using both generated and natural reward environments
3. Shepard, R. N. Toward a universal law of generalization for psychological science. *Science* **237**, 1317–1323 (1987).
   * A classic paper proposing a similarity-based account of generalization, and arguing that the first general law of psychology should be based on generalization
4. Chater, N. & Vitányi, P. M. B. The generalized universal law of generalization. *J. Math. Psychol.* **47**, 346–369 (2003).
5. Tenenbaum, J. B. & Griffiths, T. L. Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* **24**, 629–40; discussion 652–791 (2001).
   * This paper is an important milestone in Bayesian models of cognition and showed how a distributional approach to rule-based generalization produces behavior consistent with previous similarity-based accounts.
6. Taylor, J. E. *et al.* How do we generalize? *arXiv [q-bio.NC]* (2021).
7. Norbury, A., Robbins, T. W. & Seymour, B. Value generalization in human avoidance learning. *Elife* **7**, (2018).
8. Poggio, T. & Bizzi, E. Generalization in vision and motor control. *Nature* **431**, 768–774 (2004).
9. Zhang, C., Bengio, S., Hardt, M., Recht, B. & Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv [cs.LG]* (2016).
10. Jäkel, F., Schölkopf, B. & Wichmann, F. A. Generalization and similarity in exemplar models of categorization: insights from machine learning. *Psychon. Bull. Rev.* **15**, 256–271 (2008).
11. Geirhos, R. *et al.* Generalisation in humans and deep neural networks. *arXiv [cs.CV]* (2018).
12. Chater, N., Tenenbaum, J. B. & Yuille, A. Probabilistic models of cognition: conceptual foundations. *Trends Cogn. Sci.* **10**, 287–291 (2006).
13. Goodman, N. D., Tenenbaum, J. B., Feldman, J. & Griffiths, T. L. A rational analysis of rule-based concept learning. *Cogn. Sci.* **32**, 108–154 (2008).
14. Hahn, U. & Chater, N. Similarity and rules: distinct? Exhaustive? Empirically distinguishable? *Cognition* **65**, 197–230 (1998).

15. Sutton, R. S. & Barto, A. G. *Reinforcement Learning, second edition: An Introduction*. (MIT Press, 2018).
16. Stachenfeld, K. L., Botvinick, M. M. & Gershman, S. J. The hippocampus as a predictive map. *Nat. Neurosci.* **20**, 1643–1653 (2017).
    * An influential account of how the SR captures the predictive encodings of the hippocampus, where expected state transitions form the basis for relational structure learning
17. Botvinick, M. *et al.* Reinforcement Learning, Fast and Slow. *Trends Cogn. Sci.* **23**, 408–422 (2019).
18. Wimmer, G. E., Elliott Wimmer, G., Daw, N. D. & Shohamy, D. Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience* vol. 35 1092–1104 (2012).
19. Goodman, N. Seven Strictures on Similarity. in *Problems and Projects* (Bobs-Merril, 1972).
20. Medin, D. L., Goldstone, R. L. & Gentner, D. Respects for similarity. *Psychol. Rev.* **100**, 254–278 (1993).
21. Tversky, A. Features of similarity. *Psychol. Rev.* **84**, 327–352 (1977).
22. Silver, D. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* **529**, 484–489 (2016).
23. Tesauro, G. Temporal difference learning and TD-Gammon. *Commun. ACM* **38**, 58–68 (1995).
24. Smith, E. E. & Medin, D. L. Categories and Concepts. (1981) doi:10.4159/harvard.9780674866270.
25. Nosofsky, R. M. Attention, similarity, and the identification–categorization relationship. *J. Exp. Psychol. Gen.* **115**, 39–57 (1986).
26. Medin, D. L. & Schaffer, M. M. Context theory of classification learning. *Psychol. Rev.* **85**, 207–238 (1978).
27. Rosch, E. H. Natural categories. *Cogn. Psychol.* **4**, 328–350 (1973).
28. Erickson, M. A. & Kruschke, J. K. Rules and exemplars in category learning. *J. Exp. Psychol. Gen.* **127**, 107–140 (1998).
29. Ashby, F. G. & Maddox, W. T. Human category learning. *Annu. Rev. Psychol.* **56**, 149–178 (2005).
30. Ashby, F. G. & Gott, R. E. Decision rules in the perception and categorization of multidimensional stimuli. *J. Exp. Psychol. Learn. Mem. Cogn.* **14**, 33–53 (1988).
31. Nosofsky, R. M., Palmeri, T. J. & McKinley, S. C. Rule-plus-exception model of classification learning. *Psychol. Rev.* **101**, 53–79 (1994).
32. Tenenbaum, J. B. Rules and similarity in concept learning. *Adv. Neural Inf. Process. Syst.* **12**, 59–65 (2000).
33. Ashby, F. G., Alfonso-Reese, L. A., Turken, A. U. & Waldron, E. M. A neuropsychological theory of multiple systems in category learning. *Psychol. Rev.* **105**, 442–481 (1998).
34. Love, B. C., Medin, D. L. & Gureckis, T. M. SUSTAIN: a network model of category learning. *Psychol. Rev.* **111**, 309–332 (2004).
35. Kemp, C. & Tenenbaum, J. B. The discovery of structural form. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 10687–10692 (2008).
    * An important advancement in Bayesian modeling revealing how an entire universe of graph-structures can be inferred by aggregating over a distribution of hypothesized structures.
36. Ellis, K. *et al.* DreamCoder: Growing generalizable, interpretable knowledge with wake-sleep Bayesian program learning. *arXiv [cs.AI]* (2020).
37. Lake, B. M., Salakhutdinov, R. & Tenenbaum, J. B. Human-level concept learning through probabilistic program induction. *Science* **350**, 1332–1338 (2015).
38. Smith, J. D. & Minda, J. P. Prototypes in the mist: The early epochs of category learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **24**, 1411–1436 (1998).
39. Hahn, U. & Ramscar, M. *Similarity and categorization*. (Oxford University Press, 2001).
40. Tenenbaum, J. B. Bayesian modeling of human concept learning. *Adv. Neural Inf. Process. Syst.* 59–68 (1999).
41. Pothos, E. M. The rules versus similarity distinction. *Behav. Brain Sci.* **28**, 1–14; discussion 14–49 (2005).

42. Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B. & Griffiths, T. L. Structure and flexibility in Bayesian models of cognition. *Oxford handbook of computational and mathematical psychology* (2015).
43. Rasmussen, C. E. & Williams, C. K. I. Gaussian Processes for Machine Learning. (2005) doi:10.7551/mitpress/3206.001.0001.
44. Schulz, E., Speekenbrink, M. & Krause, A. A tutorial on Gaussian process regression: Modelling, exploring, and exploiting functions. *J. Math. Psychol.* **85**, 1–16 (2018).
45. Wu, C. M., Schulz, E. & Gershman, S. J. Inference and Search on Graph-Structured Spaces. *Computational Brain & Behavior* **4**, 125–147 (2021).
* An extension of the Gaussian Process framework to model human human function learning and RL behavior in graph-structured environments, where a diffusion kernel provides a similarity metric based on relational knowledge.
46. Kondor, R. I. & Lafferty, J. Diffusion kernels on graphs and other discrete input spaces. in *Proceedings of the 19th International Conference on Machine Learning* 3121–3322 (2002).
47. Smola, A. J. & Kondor, R. Kernels and Regularization on Graphs. *Learning Theory and Kernel Machines* 144–158 (2003) doi:10.1007/978-3-540-45167-9_12.
48. Duvenaud, D., Lloyd, J. R., Grosse, R., Tenenbaum, J. B. & Ghahramani, Z. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. *arXiv [stat.ML]* (2013).
49. Schulz, E., Tenenbaum, J. B., Duvenaud, D., Speekenbrink, M. & Gershman, S. J. Compositional inductive biases in function learning. *Cogn. Psychol.* **99**, 44–79 (2017).
50. Meder, B., Mayrhofer, R. & Waldmann, M. R. Structure induction in diagnostic causal reasoning. *Psychol. Rev.* **121**, 277–301 (2014).
* This paper shows that people go beyond the information they are given and can induce causal structures to make probabilistic inferences.
51. Griffiths, T. L. & Tenenbaum, J. B. Structure and strength in causal induction. *Cogn. Psychol.* **51**, 334–384 (2005).
52. Xu, F. & Tenenbaum, J. B. Word learning as Bayesian inference. *Psychol. Rev.* **114**, 245–272 (2007).
53. Dasgupta, I., Schulz, E. & Gershman, S. J. Where do hypotheses come from? *Cogn. Psychol.* **96**, 1–25 (2017).
54. Bonawitz, E., Denison, S., Gopnik, A. & Griffiths, T. L. Win-Stay, Lose-Sample: a simple sequential algorithm for approximating Bayesian inference. *Cogn. Psychol.* **74**, 35–65 (2014).
55. Dasgupta, I., Schulz, E., Goodman, N. D. & Gershman, S. J. Remembrance of inferences past: Amortization in human hypothesis generation. *Cognition* **178**, 67–81 (2018).
56. Whittington, J. C. R. *et al.* The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell* **183**, 1249–1263.e23 (2020).
* Here, the TEM is introduced, which can infer unobserved state-transitions through the mechanism of path integration.
57. van Rooij, I., Blokpoel, M., Kwisthout, J. & Wareham, T. *Cognition and Intractability: A Guide to Classical and Parameterized Complexity Analysis*. (Cambridge University Press, 2019).
58. Sanborn, A. N., Griffiths, T. L. & Navarro, D. J. Rational approximations to rational models: alternative algorithms for category learning. *Psychol. Rev.* **117**, 1144–1167 (2010).
59. Brehmer, B. Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organizational Behavior and Human Performance* vol. 11 1–27 (1974).
60. Lucas, C. G., Griffiths, T. L., Williams, J. J. & Kalish, M. L. A rational model of function learning. *Psychon. Bull. Rev.* **22**, 1193–1215 (2015).
61. Koh, K. & Meyer, D. E. Function learning: Induction of continuous stimulus-response relations. *Journal of Experimental Psychology: Learning, Memory, and Cognition* vol. 17 811–836 (1991).
62. Kalish, M. L., Griffiths, T. L. & Lewandowsky, S. Iterated learning: intergenerational knowledge transmission reveals inductive biases. *Psychon. Bull. Rev.* **14**, 288–294 (2007).

63. Busemeyer, J. R., Byun, E., Delosh, E. L. & McDaniel, M. A. Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. *Knowledge, concepts and categories.* **464**, 408–437 (1997).
64. Carroll, J. D. Functional learning: The learning of continuous functional mappings relating stimulus and response continua. *ETS Research Bulletin Series* **1963**, i–144 (1963).
65. Brehmer, B. Learning complex rules in probabilistic inference tasks. *Scand. J. Psychol.* **17**, 309–312 (1976).
66. DeLosh, E. L., Busemeyer, J. R. & McDaniel, M. A. Extrapolation: the sine qua non for abstraction in function learning. *J. Exp. Psychol. Learn. Mem. Cogn.* **23**, 968–986 (1997).
67. McDaniel, M. A. & Busemeyer, J. R. The conceptual basis of function learning and extrapolation: comparison of rule-based and associative-based models. *Psychon. Bull. Rev.* **12**, 24–42 (2005).
68. McClelland, J. L., Rumelhart, D. E., Group, P. R. & Others. *Parallel distributed processing*. vol. 2 (MIT press Cambridge, MA, 1986).
69. Busemeyer, J. R., Delosh, E., Choi, S. & McDaniel, M. A. Extrapolation: The Sine Qua Non of Abstraction. *PsycEXTRA Dataset* (1992) doi:10.1037/e665412011-109.
70. Kwantes, P. J. & Neal, A. Why people underestimate y when extrapolating in linear functions. *J. Exp. Psychol. Learn. Mem. Cogn.* **32**, 1019–1030 (2006).
71. Kalish, M. L., Lewandowsky, S. & Kruschke, J. K. Population of linear experts: knowledge partitioning and function learning. *Psychol. Rev.* **111**, 1072–1099 (2004).
72. Griffiths, T. L., Lucas, C. G., Williams, J. J. & Kalish, M. L. Modeling human function learning with Gaussian processes. in *Proceedings of the 21st International Conference on Neural Information Processing Systems* 553–560 (2008).
73. Schölkopf, B. *et al. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. (MIT Press, 2002).
74. Mercer, J. Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society A* **209**, 441–458 (1909).
75. Wang, R., Salakhutdinov, R. & Yang, L. F. Reinforcement Learning with General Value Function Approximation: Provably Efficient Approach via Bounded Eluder Dimension. *arXiv [cs.LG]* (2020).
76. Schölkopf, B. Learning to see and act. *Nature* **518**, 486–487 (2015).
77. Schulz, E. *et al.* Structured, uncertainty-driven exploration in real-world consumer choice. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 13903–13908 (2019).
78. Sutton, R. S., McAllester, D. A., Singh, S. P. & Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. in *Advances in neural information processing systems* 1057–1063 (2000).
79. Schulz, E., Wu, C. M., Huys, Q. J. M., Krause, A. & Speekenbrink, M. Generalization and Search in Risky Environments. *Cogn. Sci.* **42**, 2592–2620 (2018).
80. Wu, C. M., Schulz, E., Garvert, M. M., Meder, B. & Schuck, N. W. Similarities and differences in spatial and non-spatial cognitive maps. *PLoS Comput. Biol.* **16**, e1008149 (2020).
81. Stojić, H., Schulz, E., P Analytis, P. & Speekenbrink, M. It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. *J. Exp. Psychol. Gen.* **149**, 1878–1907 (2020).
82. Schulz, E., Wu, C. M., Ruggeri, A. & Meder, B. Searching for Rewards Like a Child Means Less Generalization and More Directed Exploration. *Psychol. Sci.* **30**, 1561–1572 (2019).
83. Meder, B., Wu, C. M., Schulz, E. & Ruggeri, A. Development of directed and random exploration in children. (2020) doi:10.31234/osf.io/7zhxv.
84. Schulz, E., Konstantinidis, E. & Speekenbrink, M. Putting bandits into context: How function learning supports decision making. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**, 927–943 (2018).
85. Settles, B. *Active Learning Literature Survey*. (2009).
86. Nelson, J. D. Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychol. Rev.* **112**, 979–999 (2005).

87. Schulz, E. & Gershman, S. J. The algorithmic architecture of exploration in the human brain. *Curr. Opin. Neurobiol.* **55**, 7–14 (2019).
88. Markant, D. B. & Gureckis, T. M. Is it better to select or to receive? Learning via active and passive hypothesis testing. *J. Exp. Psychol. Gen.* **143**, 94–122 (2014).
89. Nelson, J. D., Divjak, B., Gudmundsdottir, G., Martignon, L. F. & Meder, B. Children's sequential information search is sensitive to environmental probabilities. *Cognition* **130**, 74–80 (2014).
90. Meder, B., Nelson, J. D., Jones, M. & Ruggeri, A. Stepwise versus globally optimal search in children and adults. *Cognition* **191**, 103965 (2019).
91. Jones, A., Schulz, E., Meder, B. & Ruggeri, A. Active Function Learning. *bioRxiv* 262394 (2018) doi:10.1101/262394.
92. Borji, A. & Itti, L. Bayesian optimization explains human active search. *Adv. Neural Inf. Process. Syst.* (2013).
93. Mehlhorn, K. *et al.* Unpacking the exploration–exploitation tradeoff: A synthesis of human and animal literatures. *Decisions* **2**, 191–215 (2015).
94. Luthra, M., Izquierdo, E. J. & Todd, P. M. Cognition Evolves with the Emergence of Environmental Patchiness. in *Artificial Life Conference Proceedings* 450–458 (MIT Press One Rogers Street, Cambridge, MA 02142-1209 USA journals-info@ mit …, 2020).
95. Wilson, R. C., Geana, A., White, J. M., Ludvig, E. A. & Cohen, J. D. Humans use directed and random exploration to solve the explore-exploit dilemma. *J. Exp. Psychol. Gen.* **143**, 2074–2081 (2014).
96. Srinivas, N., Krause, A., Kakade, S. M. & Seeger, M. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. *arXiv [cs.LG]* (2009).
97. Wu, C. M., Schulz, E., Pleskac, T. J. & Speekenbrink, M. Time to explore: Adaptation of exploration under time pressure. (2021) doi:10.31234/osf.io/dsw7q.
98. Zajkowski, W. K., Kossut, M. & Wilson, R. C. A causal role for right frontopolar cortex in directed, but not random, exploration. *Elife* **6**, e27430 (2017).
99. Somerville, L. H. *et al.* Charting the expansion of strategic exploratory behavior during adolescence. *J. Exp. Psychol. Gen.* **146**, 155–164 (2017).
100. Gershman, S. J. & Daw, N. D. Reinforcement Learning and Episodic Memory in Humans and Animals: An Integrative Framework. *Annu. Rev. Psychol.* **68**, 101–128 (2017).
101. Torgerson, W. S. Multidimensional scaling: I. Theory and method. *Psychometrika* vol. 17 401–419 (1952).
102. Ekman, G. Dimensions of Color Vision. *J. Psychol.* **38**, 467–474 (1954).
103. Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika* vol. 27 219–246 (1962).
104. Kruskal, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika* vol. 29 115–129 (1964).
105. Jäkel, F., Schölkopf, B. & Wichmann, F. A. Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology* vol. 52 297–303 (2008).
***** An insightful discussion of similarity in psychology, with the intriguing argument that kernel similarity avoids some of the pitfalls of previous metric representations of similarity (e.g., Shepard's Law of Generalization).
106. Gershman, S. J. Deconstructing the human algorithms for exploration. *Cognition* **173**, 34–42 (2018).
107. Lau, T., Gershman, S. J. & Cikara, M. Social structure learning in human anterior insula. *Elife* **9**, (2020).
108. Tavares, R. M. *et al.* A Map for Social Navigation in the Human Brain. *Neuron* **87**, 231–243 (2015).
109. Machado, M. C. *et al.* Eigenoption Discovery through the Deep Successor Representation. in *International Conference on Learning Representations* (2018).
110. Zhu, X., Lafferty, J. & Ghahramani, Z. *Semi-Supervised Learning: From Gaussian Fields to Gaussian Processes*. (2003).

111. Niv, Y. *et al.* Reinforcement learning in multidimensional environments relies on attention mechanisms. *J. Neurosci.* **35**, 8145–8157 (2015).
112. Gershman, S. J. & Blei, D. M. A Tutorial on Bayesian Nonparametric Models. *arXiv [stat.ML]* (2011).
113. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv [stat.ML]* (2013).
114. Murphy, G. L. & Medin, D. L. The role of theories in conceptual coherence. *Psychol. Rev.* **92**, 289–316 (1985).
115. Hahn, U., Chater, N. & Richardson, L. B. Similarity as transformation. *Cognition* **87**, 1–32 (2003).
116. Gottlieb, J., Cohanpour, M., Li, Y., Singletary, N. & Zabeh, E. Curiosity, information demand and attentional priority. *Current Opinion in Behavioral Sciences* vol. 35 83–91 (2020).
117. Dayan, P., Kakade, S. & Montague, P. R. Learning and selective attention. *Nat. Neurosci.* **3 Suppl**, 1218–1223 (2000).
118. Reynolds, J. H. & Heeger, D. J. The normalization model of attention. *Neuron* **61**, 168–185 (2009).
119. Desimone, R. & Duncan, J. Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience* vol. 18 193–222 (1995).
120. Radulescu, A., Shin, Y. S. & Niv, Y. Human Representation Learning. *Annu. Rev. Neurosci.* (2021) doi:10.1146/annurev-neuro-092920-120559.
* A review of current theories of how humans learn the appropriate representations for generalization with a focus on selective attention.
121. Mikhael, J. G., Lai, L. & Gershman, S. J. Rational inattention and tonic dopamine. *PLoS Comput. Biol.* **17**, e1008659 (2021).
122. Leong, Y. C., Radulescu, A., Daniel, R., DeWoskin, V. & Niv, Y. Dynamic Interaction between Reinforcement Learning and Attention in Multidimensional Environments. *Neuron* **93**, 451–463 (2017).
123. Schlegelmilch, R. & von Helversen, B. The influence of reward magnitude on stimulus memory and stimulus generalization in categorization decisions. *J. Exp. Psychol. Gen.* **149**, 1823–1854 (2020).
124. Gottlieb, J. Attention as rational choice. in *The Drive for Knowledge: the Science of Human Information-Seeking* (eds. Cogliati-Dezza, I., Schulz, E. & Wu, C. M.) (Cambridge University Press, 2022).
125. Shenhav, A. *et al.* Toward a Rational and Mechanistic Account of Mental Effort. *Annu. Rev. Neurosci.* **40**, 99–124 (2017).
126. Bhui, R., Lai, L. & Gershman, S. J. Resource-rational decision making. *Current Opinion in Behavioral Sciences* **41**, 15–21 (2021).
127. Grujic, N., Brus, J., Burdakov, D. & Polania, R. Rational inattention in mice. *bioRxiv* (2021).
128. Rescorla, R. A. & Wagner, A. R. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. in *Classical conditioning: current research and theory* vol. 2 (1972).
129. Barsalou, L. W. Ad hoc categories. *Mem. Cognit.* **11**, 211–227 (1983).
130. Gershman, S. J. & Niv, Y. Learning latent structure: carving nature at its joints. *Curr. Opin. Neurobiol.* **20**, 251–256 (2010).
131. Radvansky, G. A. & Zacks, J. M. *Event Cognition*. (Oxford University Press, 2014).
132. Butz, M. V., Achimova, A., Bilkey, D. & Knott, A. Event-predictive cognition: A root for conceptual human thought. *Top. Cogn. Sci.* **13**, 10–24 (2021).
133. Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M. & Gershman, S. J. Structured Event Memory: A neuro-symbolic model of event cognition. *Psychol. Rev.* **127**, 327–361 (2020).
134. Gershman, S. J., Blei, D. M. & Niv, Y. Context, learning, and extinction. *Psychol. Rev.* **117**, 197–209 (2010).
135. Sanders, H., Wilson, M. A. & Gershman, S. J. Hippocampal remapping as hidden state inference. *Elife* **9**, (2020).
136. Mack, M. L., Preston, A. R. & Love, B. C. Ventromedial prefrontal cortex compression during concept learning. *Nat. Commun.* **11**, 46 (2020).

137. Bates, C. J. & Jacobs, R. A. Efficient data compression in perception and perceptual memory. *Psychol. Rev.* (2020) doi:10.1037/rev0000197.
138. Sims, C. R. Efficient coding explains the universal law of generalization in human perception. *Science* **360**, 652–656 (2018).
139. Austerweil, J. L. & Griffiths, T. L. A nonparametric Bayesian framework for constructing flexible feature representations. *Psychol. Rev.* **120**, 817–851 (2013).
140. Flesch, T., Juechems, K., Dumbalska, T. & Saxe, A. Rich and lazy learning of task representations in brains and neural networks. *bioRxiv* (2021).
141. Berger, T. *Rate Distortion Theory: A Mathematical Basis for Data Compression*. (Prentice-Hall, 1971).
142. Sims, C. R. Rate-distortion theory and human perception. *Cognition* **152**, 181–198 (2016).
143. Dasgupta, I. & Gershman, S. J. Memory as a Computational Resource. *Trends Cogn. Sci.* **25**, 240–251 (2021).
144. Gershman, S. J. The rational analysis of memory. in *Oxford handbook of human memory* (Oxford University Press, 2021).
145. Cushman, F. Rationalization is rational. *Behav. Brain Sci.* **43**, e28 (2020).
146. Shannon, C. E. & Others. Coding theorems for a discrete source with a fidelity criterion. *IRE Nat. Conv. Rec* **4**, 1 (1959).
147. Pouncy, T., Tsividis, P. & Gershman, S. J. What is the model in model-based planning? *Cogn. Sci.* **45**, e12928 (2021).
148. Navarro, D. J. & Griffiths, T. L. Latent features in similarity judgments: a nonparametric bayesian approach. *Neural Comput.* **20**, 2597–2628 (2008).
149. Sheahan, H., Luyckx, F., Nelli, S., Teupe, C. & Summerfield, C. Neural state space alignment for magnitude generalization in humans and recurrent networks. *Neuron* **109**, 1214–1226.e8 (2021).
150. Nagy, D. G., Török, B. & Orbán, G. Optimal forgetting: Semantic compression of episodic memories. *PLoS Comput. Biol.* **16**, e1008367 (2020).
151. Malloy, T. & Sims, C. Modelling Visual Decision Making Using a Variational Autoencoder. in *Proceedings of the 43rd Annual Conference of the Cognitive Science Society* (eds. Fitch, T., Lamm, C., Leder, H. & Teßmar-Raible, K.) (2021).
152. Halford, G. S., Wilson, W. H. & Phillips, S. Relational knowledge: the foundation of higher cognition. *Trends Cogn. Sci.* **14**, 497–505 (2010).
153. Garvert, M. M., Dolan, R. J. & Behrens, T. E. J. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *Elife* **6**, e17086 (2017).
154. Momennejad, I. Learning Structures: Predictive Representations, Replay, and Generalization. *Current Opinion in Behavioral Sciences* **32**, 155–166 (2020).
155. Mark, S., Moran, R., Parr, T., Kennerley, S. W. & Behrens, T. E. J. Transferring structural knowledge across cognitive maps in humans and models. *Nat. Commun.* **11**, 4783 (2020).
* An empirical demonstration of how structural knowledge (e.g., "schemas") can be transferred across domains, supporting more efficient learning.
156. Palmeri, T. J. & Blalock, C. The role of background knowledge in speeded perceptual categorization. *Cognition* **77**, B45–57 (2000).
157. Lin, E. L. & Murphy, G. L. Effects of background knowledge on object categorization and part detection. *Journal of Experimental Psychology: Human Perception and Performance* vol. 23 1153–1169 (1997).
158. Kemp, C. & Tenenbaum, J. B. Structured statistical models of inductive reasoning. *Psychol. Rev.* **116**, 20–58 (2009).
159. Simon, H. A. & Lea, G. Problem solving and rule induction: A unified view. in *Knowledge and cognition* (ed. Gregg, L. W.) vol. 321 (Lawrence Erlbaum, 1974).
160. Bruner, J. S., Goodnow, J. J. & Austin, G. A. A study of thinking. *AIBS bull.* **7**, 40 (1957).
161. Reber, A. S. & Lewis, S. Implicit learning: An analysis of the form and structure of a body of tacit knowledge. *Cognition* **5**, 333–361 (1977).

162. Newell, A. & Simon, H. A. Computer science as empirical inquiry: Symbols and search. *Commun. ACM* **19**, 113–126 (1976).
163. Tsividis, P. A. *et al.* Human-Level Reinforcement Learning through Theory-Based Modeling, Exploration, and Planning. *arXiv [cs.AI]* (2021).
164. Deisenroth, M. & Rasmussen, C. E. PILCO: A model-based and data-efficient approach to policy search. in *Proceedings of the 28th International Conference on machine learning (ICML-11)* 465–472 (Citeseer, 2011).
165. Allen, K. R., Smith, K. A. & Tenenbaum, J. B. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 29302–29310 (2020).
166. Kwisthout, J. & van Rooij, I. Computational Resource Demands of a Predictive Bayesian Brain. *Computational Brain & Behavior* **3**, 174–188 (2020).
167. Nye, M. I., Solar-Lezama, A., Tenenbaum, J. B. & Lake, B. M. Learning Compositional Rules via Neural Program Synthesis. *arXiv [cs.AI]* (2020).
168. Lake, B. M., Linzen, T. & Baroni, M. Human few-shot learning of compositional instructions. *arXiv [cs.CL]* (2019).
169. Schwartenbeck, P., Baram, A., Liu, Y., Mark, S. & Muller, T. Generative replay for compositional visual understanding in the prefrontal-hippocampal circuit. *bioRxiv* (2021).
170. Fränken, J.-P., Theodoropoulos, N. C. & Bramley, N. R. Algorithms of Adaptation in Inductive Inference. (2021) doi:10.31234/osf.io/ysndt.
171. Dayan, P. Improving Generalization for Temporal Difference Learning: The Successor Representation. *Neural Computation* vol. 5 613–624 (1993).
172. O'Donnell, T. J. *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. (MIT Press, 2015).
173. Tolman, E. C. Cognitive maps in rats and men. *Psychological Review* vol. 55 189–208 (1948).
174. Behrens, T. E. J. *et al.* What Is a Cognitive Map? Organizing Knowledge for Flexible Behavior. *Neuron* **100**, 490–509 (2018).
175. Moser, E. I. *et al.* Grid cells and cortical representation. *Nat. Rev. Neurosci.* **15**, 466–481 (2014).
176. Peer, M., Brunec, I. K., Newcombe, N. S. & Epstein, R. A. Structuring Knowledge with Cognitive Maps and Cognitive Graphs. *Trends Cogn. Sci.* **25**, 37–54 (2021).
177. Epstein, R. A., Patai, E. Z., Julian, J. B. & Spiers, H. J. The cognitive map in humans: spatial navigation and beyond. *Nat. Neurosci.* **20**, 1504–1513 (2017).
178. Bao, X. *et al.* Grid-like Neural Representations Support Olfactory Navigation of a Two-Dimensional Odor Space. *Neuron* **102**, 1066–1075.e5 (2019).
179. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
180. Gershman, S. J., Malmaud, J. & Tenenbaum, J. B. Structured representations of utility in combinatorial domains. *Decisions* **4**, 67–86 (2017).
181. Russek, E. M., Momennejad, I., Botvinick, M. M., Gershman, S. J. & Daw, N. D. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLoS Comput. Biol.* **13**, e1005768 (2017).
182. Momennejad, I. *et al.* The successor representation in human reinforcement learning. *Nat Hum Behav* **1**, 680–692 (2017).
183. Geerts, J. P., Stachenfeld, K. L. & Burgess, N. Probabilistic Successor Representations with Kalman Temporal Differences. *arXiv [q-bio.NC]* (2019).
184. Madarasz, T. & Behrens, T. Better Transfer Learning with Inferred Successor Maps. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) vol. 32 (Curran Associates, Inc., 2019).
185. Machado, M. C., Bellemare, M. G. & Bowling, M. Count-Based Exploration with the Successor Representation. *arXiv [cs.LG]* (2018).

186. Mittelstaedt, M.-L. & Mittelstaedt, H. Homing by path integration in a mammal. *Naturwissenschaften* vol. 67 566–567 (1980).
187. Manns, J. R. & Eichenbaum, H. Evolution of declarative memory. *Hippocampus* **16**, 795–808 (2006).
188. Barron, H. C., Dolan, R. J. & Behrens, T. E. J. Online evaluation of novel choices by simultaneous representation of multiple memories. *Nat. Neurosci.* **16**, 1492–1498 (2013).
189. Zhou, J. *et al.* Evolving schema representations in orbitofrontal ensembles during learning. *Nature* **590**, 606–611 (2021).
190. Schank, R. C. & Abelson, R. P. *Scripts, Plans, Goals, and Understanding: An Inquiry Into Human Knowledge Structures*. (Routledge, 1977).
191. Wingate, D., Diuk, C., O'Donnell, T., Tenenbaum, J. & Gershman, S. *Compositional Policy Priors*. (2013).
192. Le, T. A. *et al.* Hybrid Memoised Wake-Sleep: Approximate Inference at the Discrete-Continuous Interface. *arXiv [cs.CV]* (2021).
193. Doucet, A., Johansen, A. M. & Others. A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of nonlinear filtering* **12**, 3 (2009).
194. Speekenbrink, M. A tutorial on particle filters. *J. Math. Psychol.* **73**, 140–152 (2016).
195. Schwartenbeck, P. *et al.* Compositional Neural Representations in the Hippocampal Formation and Prefrontal Cortex Underlie Visual Construction and Planning.