ANNUAL REVIEWS

Annual Review of Psychology

Unifying Principles of Generalization: Past, Present, and Future

Charley M. Wu,[1,2,3] Björn Meder,[4] and Eric Schulz[5]

[1]Human and Machine Cognition Lab, University of Tübingen, Tübingen, Germany; email: charley.wu@uni-tuebingen.de

[2]Department of Computational Neuroscience, Max Planck Institute of Biological Cybernetics, 72074 Tübingen, Germany

[3]Center for Adaptive Rationality, Max Planck Institute for Human Development, Berlin, Germany

[4]Institute for Mind, Brain and Behavior, Department of Psychology, Health and Medical University Potsdam, Potsdam, Germany

[5]Helmholtz Institute for Human-Centered AI, Helmholtz Zentrum München, Munich, Germany

Annu. Rev. Psychol. 2025. 76:275–302

First published as a Review in Advance on October 16, 2024

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-021524-110810

Keywords

generalization, concept learning, function learning, value approximation, reinforcement learning, structure induction

Abstract

Generalization, defined as applying limited experiences to novel situations, represents a cornerstone of human intelligence. Our review traces the evolution and continuity of psychological theories of generalization, from its origins in concept learning (categorizing stimuli) and function learning (learning continuous input-output relationships) to domains such as reinforcement learning and latent structure learning. Historically, there have been fierce debates between approaches based on rule-based mechanisms, which rely on explicit hypotheses about environmental structure, and approaches based on similarity-based mechanisms, which leverage comparisons to prior instances. Each approach has unique advantages: Rules support rapid knowledge transfer, while similarity is computationally simple and flexible. Today, these debates have culminated in the development of hybrid models grounded in Bayesian principles, effectively marrying the precision of rules with the flexibility of similarity. The ongoing success of hybrid models not only bridges past dichotomies but also underscores the importance of integrating both rules and similarity for a comprehensive understanding of human generalization.

Erratum >

Contents

1. INTRODUCTION

Generalization: the process of applying previously acquired knowledge to new, unfamiliar situations

Concept learning: learning to apply discrete category labels to objects or events

Function learning: learning to understand and predict the continuous relationship between input and output variables

Rules: explicit hypotheses about the structure of the environment that can guide generalization

Similarity: a comparison of new situations to previous experiences as a basis for generalization

In the unending flux and flow of new experiences, the study of how people generalize past experiences to novel situations is a testament to the flexibility of human intelligence (Lake et al. 2017, Chollet 2019). Thus, it is no surprise that generalization has occupied a central role in psychology (Shepard 1987, Tenenbaum & Griffiths 2001, Chater & Vitányi 2003, Wu et al. 2018), neuroscience (Norbury et al. 2018, Taylor et al. 2021), and machine learning (Zhang et al. 2016, Geirhos et al. 2018). Here, we bridge traditional psychological theories with modern computational approaches, providing new perspectives for both old problems and enduring challenges. While the computational methods are certainly new, their theoretical underpinnings and core questions are deeply familiar to psychology and can be traced back to foundational research in concept and function learning.

Over the years, debates about the mechanisms underlying human generalization have spanned multiple domains. Research in concept learning has studied how people generalize learned category labels when asked to classify new instances—for example, 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 within and extrapolation beyond observed data, such as predicting how much study time is needed to pass a test or anticipating how much one will enjoy a new menu item at their 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.

While fierce historical debates have raged over which ingredient is more central, today these arguments have largely been settled in favor of hybrid models, which have both rule- and similarity-based interpretations and are frequently based on Bayesian principles (Tenenbaum &

Griffiths 2001, Lucas et al. 2015). While a duality of interpretations suggests an exchangeability between rule- and similarity-based representations (Goodman et al. 2008), the computations used by hybrid models typically operate over either hypothesized rules or representations of similarity, each conferring distinct advantages. Rules unlock compositionality and rapid transfer, while similarity is easy to compute and can flexibly capture various relationships in the environment.

In this review, we revisit the distinction between similarity- and rule-based mechanisms of generalization. Our approach seeks to bridge the past and the present by emphasizing the continuity of these two mechanisms in theories of generalization. We first explore the development of theories of generalization in concept learning and function learning. Each domain has seen converging trajectories toward hybrid models that integrate both rule-based and similarity-based approaches. Second, we connect classic theories of function learning with contemporary methods for value generalization in reinforcement learning (RL), thus integrating a new dimension of uncertainty-directed exploration to guide adaptive learning when acting on the world. Third, we highlight inherent relations between Bayesian concept learning and theories of structure induction, which support generalization by inferring hidden environmental structure. We conclude by proposing new directions for further integrating similarity and rules, combining their relative advantages to unlock faster and more efficient generalization in increasingly complex problems.

Reinforcement learning (RL): a framework for understanding learning through trial-and-error feedback from the environment

Uncertainty-directed exploration: a strategic approach in learning or decision making where exploration is guided by specific goals or hypotheses, such as reducing subjective uncertainty

Bayesian concept learning: a probabilistic approach to concept learning, using a distribution over rule-like hypotheses about concept boundaries and producing similarity-like generalization patterns

Structure induction: the process of inferring underlying structure from observed data, often using Bayesian principles

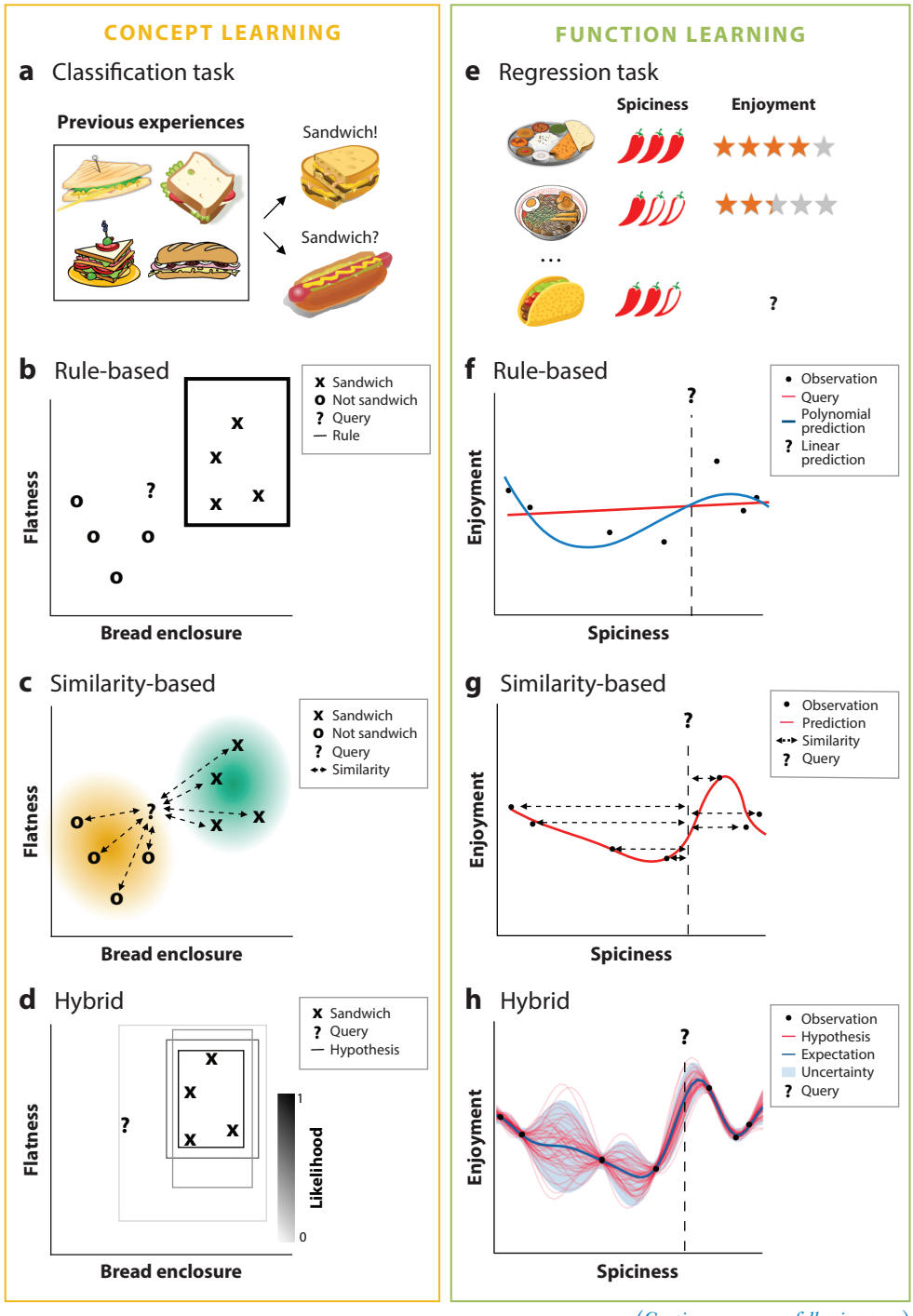
2. COMMON PRINCIPLES FOR GENERALIZATION

We first review foundational theories of generalization in concept learning and function learning, which broadly map onto the distinction between classification and regression problems, as they are commonly referred to in statistics and machine learning. A child distinguishing dogs from cats based on characteristics like barking or meowing is a type of classification problem used in concept learning, while a teacher predicting students' test scores based on study habits and past performance is a type of regression problem used in function learning. Research in these two domains has largely progressed in distinct, parallel tracks. Yet they share a similar historical trajectory of debates about the main mechanisms supporting generalization. The proposed mechanisms can be categorized as rule-based approaches, which focus on extracting regularities or generic rules from the environment, and similarity-based approaches, which compare new situations to past examples.

In this section, we examine the evolution of theories about generalization across concept learning and function learning. In both domains, these theories have largely culminated in hybrid models, often using Bayesian principles to unify rule-based and similarity-based approaches. We then show how these hybrid approaches provide the foundations for scaling up to increasingly more complex and real-world problems, drawing connections between theories of function learning and modern approaches to value generalization in RL, and linking Bayesian concept learning to theories of structure learning.

2.1. Concept Learning

A chief aim of psychology has been to understand how individuals categorize and differentiate between different elements of the "blooming and buzzing confusion" (James 1890, p. 488) of the environment. Research in the domain of concept learning has long used classification problems with discrete stimuli as a means to study generalization (Rosch 1973, Medin & Schaffer 1978). An example would be 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 (**Figure 1*a***). Important debates in this field have concerned



(*Caption appears on following page*)

Figure 1 (*Figure appears on preceding page*)

Generalization in concept learning and function learning. (*a*) Concept learning is often studied based on classifying discrete stimuli (e.g., sandwich versus not sandwich). (*b*) Rule-based methods describe explicit category boundaries (*rectangle*), while (*c*) similarity-based methods utilize similarity (*arrows*) to previous exemplars (*data points*) or learned prototypes (*centroids of colored ovals*). (*d*) Bayesian concept learning (Tenenbaum & Griffiths 2001) provides a hybrid approach by defining a distribution over rules (*rectangles*), which yields similarity-like patterns of generalization (Tversky 1977, Shepard 1987). The likelihood favors narrower hypotheses (*shading of lines*). (*e*) Function learning is often studied based on predicting outputs (e.g., enjoyment) given some input (e.g., spiciness). (*f*) Rule-based methods describe specific parametric families of functions (e.g., linear or polynomial), while (*g*) similarity-based methods often use artificial neural networks to approximate nonlinear functions, where the influence of each of the data points is proportional to their similarity (i.e., inverse distance) (*arrows*). (*h*) Gaussian Process regression provides a hybrid approach using kernel similarity to describe a distribution over hypothesized functions (*red lines*), which are summarized in terms of an expectation (*blue line*) and uncertainty (*blue ribbon*). Food images are from OpenClipArt under CC0 1.0.

which representations are learned and the mechanisms used for generalizing about novel stimuli (Erickson & Kruschke 1998, Hahn & Chater 1998, Bowman et al. 2020). Here, we broadly categorize different influential approaches into rule-based and similarity-based approaches.

2.1.1. Rule-based concept learning. One influential class of theories proposes that concepts are defined based on rules that describe the explicit boundaries of category membership (Bruner et al. 1956, Ashby & Gott 1988, Rouder & Ratcliff 2006) (see **Figure 1*b***). For instance, one might describe the necessary and sufficient features (Smith & Medin 1981) 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 (Goodman et al. 2008).

Yet, for the same reasons, rules are inflexible (e.g., 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 for added flexibility (Nosofsky et al. 1994), rule-based methods only seem to offer partial explanations of human category learning (Tenenbaum & Griffiths 2001) and perform best when paired with other learning mechanisms (Ashby et al. 1998, Erickson & Kruschke 1998, Love et al. 2004). Nevertheless, the basic mechanisms of rule-based generalization (i.e., proposing explicit hypotheses) play an important role in modern theories of structure learning (Kemp & Tenenbaum 2008) and program induction (Lake et al. 2015, Rule et al. 2020), which use a probabilistic framework to add flexibility to the rigid structure of rules.

2.1.2. Similarity-based concept learning. Another class of theories uses similarity-based methods for predicting the categories of novel stimuli (**Figure 1*c***). Early theories introduced the notion of a psychological space (Torgerson 1952, Ekman 1954), 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 (1987) Universal Law of Generalization, 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 (Shepard 1962, Kruskal 1964). Intuitively, 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 to do so as their distance increases (**Figure 2*a***).

Universal Law of Generalization: the probability of a response for one stimulus being generalized to another as a function of the distance between the two stimuli in a psychological space

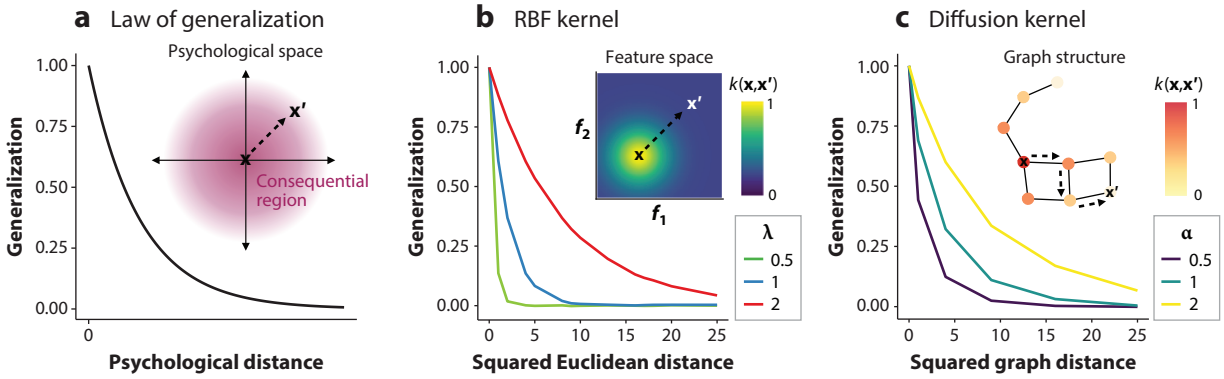


Figure 2

Principles of generalization. (*a*) Shepard's (1987) Law of Generalization describes generalization as a function of distance (*dashed arrow*) between stimuli in psychological space (*inset*). The gradient of generalization arises due to uncertainty about the extent of a consequential region, with more distant stimuli being less likely to belong to the same region. (*b*) A radial basis function (RBF) kernel provides a similarity metric based on the distance between stimuli in feature space (*inset*, *dashed arrow*), producing similar generalization gradients as Shepard's model (here quantified using Pearson correlation between outputs sampled from a Gaussian Process prior). The length scale λ governs the rate at which generalization decays as a function of distance. (*c*) In structured environments, a diffusion kernel (Kondor & Lafferty 2002, Wu et al. 2021) offers an analogous similarity metric based on the connectivity structure of a graph, where the diffusion parameter α governs the rate at which previous observations diffuse over the graph.

At the core of Shepard's theory is the assumption that representations about categories correspond to a consequential region in psychological space (**Figure 2*a***). 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. Other similarity-based approaches are consistent with this notion of a psychological space, where comparison to either previously encountered exemplars (Medin & Schaffer 1978, Nosofsky 1986) or to a learned prototype (Rosch 1973, Smith & Minda 1998) aggregated over multiple experiences provides the basis for generalization (see **Figure 1*c***).

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 (Goodman 1972, Medin et al. 1993, Hahn & Ramscar 2001). Modern theories address this challenge by providing new approaches for describing the psychological mechanisms people use to construct context-relevant similarity representations (for a review, see Radulescu et al. 2021), forming a rational rather than an arbitrary basis for computing similarity. Furthermore, advances in similarity-based approaches to generalization are now able to capture rich relational structure (Whittington et al. 2020, Wu et al. 2021) and to represent the temporal dynamics of the environment (Stachenfeld et al. 2017, Garvert et al. 2023).

Bayesian size principle: the preference for smaller, more specific hypotheses over broader ones, given consistent evidence

2.1.3. Hybrid concept learning using Bayesian principles. Today, the most prolific theories of concept learning are hybrids that have a duality of both rule- and similarity-based interpretations (Pettine et al. 2023). One influential example is Bayesian concept learning (Tenenbaum & Griffiths 2001), which uses a distribution over hypothesized category boundaries (see **Figure 1*d***) to categorize novel stimuli (see the sidebar titled Bayesian Concept Learning).

A key concept is the Bayesian size principle (Tenenbaum & Griffiths 2001), whereby under the assumption of strong sampling, greater likelihoods are assigned to narrower hypotheses consistent with the data (see **Figure 1*d***). Strong sampling assumes that rather than being completely random, the data $\mathcal{X}$ are explicitly sampled from positive examples of the category C, as is commonly the case

BAYESIAN CONCEPT LEARNING

Bayes' rule is used to describe the posterior probability that each hypothesis h captures category C given a set of positive observations $\mathbf{x}_i \in \mathcal{X}$:

$$p(h|\mathcal{X}) \propto p(h)p(\mathcal{X}|h). \quad 1.$$

This posterior integrates prior beliefs $p(h)$ and the likelihood of the data $p(\mathcal{X}|h)$, where the prior is usually assumed to be uniform, while the likelihood makes use of the Bayesian size principle (Tenenbaum & Griffiths 2001) to favor narrower hypotheses that are still consistent with the data:

$$p(\mathcal{X}|h) = \begin{cases} \frac{1}{|h|^n} & \text{if } \mathbf{x}_1, \ldots, \mathbf{x}_n \in h \\ 0 & \text{otherwise} \end{cases}. \quad 2.$$

Having defined the posterior probability of a single hypothesis h, the goal is to predict whether a novel stimulus $\mathbf{x}_*$ falls within the same category C as previously observed examples $\mathcal{X}$. Bayesian concept learning defines this probabilistically, by aggregating over all hypotheses h (i.e., category boundaries) consistent with $\mathbf{x}_*$ belonging to C:

$$p(\mathbf{x}_* \in C|\mathcal{X}) = \sum_{h:\mathbf{x}_* \in C} p(h|\mathcal{X}). \quad 3.$$

This represents a sum of posterior probabilities $p(h|\mathcal{X})$ for different hypotheses that encapsulate $\mathbf{x}_*$, where the contribution of each hypothesis is weighted by the size principle (Equation 2).

in pedagogical settings (Csibra & Gergely 2009), where a parent or a teacher provides informative examples of categories such as "plane," "dog," or "sandwich." Consequently, the distribution of the observed data $\mathcal{X}$ is expected to reflect the range of the category boundaries, thus preferring narrower hypotheses consistent with the data, where the strength of this preference increases with more observations.

Bayesian concept learning thus uses computations over rule-like category boundaries, yet it replicates behavioral patterns of similarity-based theories, such as Shepard's (1987) generalization gradient, and is equivalent to a special case of Tversky's set-theoretic model (Tversky 1977). While other hybrid models advocate for a "separate-but-equal" approach (Erickson & Kruschke 1998) by incorporating rules and similarity as separate mechanisms, Bayesian concept learning represents a unified approach where rules and similarity are seen as two sides of the same coin (Pothos 2005, Goodman et al. 2008, Austerweil et al. 2015). 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 (Griffiths & Tenenbaum 2005, 2009; Meder et al. 2014), word learning (Xu & Tenenbaum 2007), structure induction (Kemp & Tenenbaum 2008), and the learning of compositional programs (Lake et al. 2017, Fränken et al. 2022, Ellis et al. 2023, Zhao et al. 2024). A distinct advantage of operating over rule-based representations is that it affords the ability to reason compositionally by syntactically manipulating and combining multiple rules (Piantadosi et al. 2016). Yet, given an expressive hypothesis space, exact Bayesian inference is usually intractable, with most approaches relying on sample-based (Tenenbaum & Griffiths 2001, Kemp & Tenenbaum 2008, Ellis et al. 2023) or variational (Dasgupta et al. 2020) approximations. Thus, it remains an open question how humans achieve the power and productivity of rule learning with limited cognitive resources (Sanborn et al. 2010, Rubino et al. 2023, Zhou et al. 2024b).

2.2. Function Learning

Gaussian Process regression: a probabilistic function learning method using a distribution over hypothesized functions with both rule- and similarity-based interpretations

Beyond discrete category membership, generalization has also been studied in the domain of function learning (**Figure 1*e***) based on inferring a continuous relationship between inputs and outputs (Carroll 1963, Brehmer 1974, Lucas et al. 2015). An example would be learning how spiciness (input) relates to one's enjoyment of a meal (output) or how the amount of time spent studying (input) predicts test scores (output). Pioneering research by Carroll (1963) used function learning to show that human generalization goes beyond merely predicting previously observed outcomes, in contrast to the domain of concept learning, where participant responses are typically limited to previously learned category labels, even if the stimuli presented are novel. Rather, Carroll's (1963) work on function learning showed that people can extrapolate beyond their past experiences, not only generalizing to new inputs but also predicting new outputs (e.g., an off-the-charts food experience). While largely operating along a separate research tradition, the domain of function learning is also characterized by a parallel debate between rule- and similarity-based theories, which has culminated in hybrid formalizations (Busemeyer et al. 1997, Kalish et al. 2007, Lucas et al. 2015).

2.2.1. Rule-based function learning. Early research on function learning proposed rule-based theories, assuming people use a specific parametric model (e.g., a linear or polynomial function) and then learn by optimizing the parameters to best explain the data (Carroll 1963, Brehmer 1976) (see **Figure 1*f***). In function learning, rules correspond to a hypothesized relationship between variables, like the law of gravity describing a polynomial relationship between mass and distance, or the linear assumptions of a linear regression. While rule-based methods capture the systematicity of human extrapolation patterns (i.e., strong linear assumptions; DeLosh et al. 1997), they lack the flexibility of humans, who can learn to interpolate almost any function with enough training (McDaniel & Busemeyer 2005).

2.2.2. Similarity-based function learning. To better account for the flexibility of human generalization, similarity-based models (**Figure 1*g***) of function learning were developed, often using artificial neural networks (ANNs) to encode the generic principle that similar inputs produce similar outputs (McClelland et al. 1986, Busemeyer et al. 1997). The influence of previous observations decreases as a function of distance (see **Figure 1*g***) to a given input, with nearby observations exerting a larger influence. ANNs are universal function approximators (Cybenko 1989) and can approach arbitrarily low error in fitting a given function, given sufficient neurons in the hidden layers. However, while this flexibility aligns with similar human capabilities in interpolation tasks, ANNs fail to match the specific inductive biases humans exhibit when extrapolating beyond observed data (Schulz et al. 2017). For instance, humans tend to extrapolate functions with strong linear expectations (Kalish et al. 2004), a tendency not inherently captured by standard ANNs. This distinction underscores the need to further refine neural network models to more accurately mirror human cognitive processes in both interpolation and extrapolation.

2.2.3. Hybrid function learning using Bayesian principles. 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 (Rasmussen & Williams 2005) (see **Figure 1*b***), which can account for many empirical patterns of human function learning (Lucas et al. 2015) while using similar Bayesian computations as hybrid models of concept learning (see the sidebar titled Bayesian Function Learning).

Gaussian Process regression provides a Bayesian approach to function learning (for a tutorial, see Schulz et al. 2018) based on a distribution over hypothesized functions that explain the data (see **Figure 1*b***). In contrast to Bayesian concept learning, Gaussian Process regression is analytically

BAYESIAN FUNCTION LEARNING

Gaussian Process regression (Rasmussen & Williams 2005) provides a Bayesian approach to function learning by mapping inputs $\mathcal{X}$ to real-valued outputs y through a distribution over hypothesized functions h. A prior over functions takes the form of a multivariate Gaussian distribution,

$$p(h) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right), \quad 4.$$

defined by a prior mean $m(\mathbf{x})$, which is typically set to 0 without loss of generality (Rasmussen & Williams 2005), and a covariance function $k(\mathbf{x}, \mathbf{x}')$, which is defined by a choice of kernel. A common choice is the radial basis function (RBF) kernel,

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(\frac{-||\mathbf{x} - \mathbf{x}'||^2}{2\lambda^2}\right), \quad 5.$$

capturing the inductive bias that similar inputs are expected to produce similar outputs, with similarity defined as an exponentially decaying function of distance (scaled by λ) in feature space (**Figure 2*b***). The posterior distribution is then defined by conditioning on observed data $\mathcal{D} = \{\mathcal{X}, \mathbf{y}\}$ of encountered inputs $\mathbf{x}_i \in \mathcal{X}$ and outputs $y_i \in \mathbf{y}$. The posterior is also Gaussian, with predictions for any input $\mathbf{x}_*$ characterized by posterior mean $m(\mathbf{x}_*|\mathcal{D})$ and variance $v(\mathbf{x}_*|\mathcal{D})$:

$$p\left(h(\mathbf{x}_*)\,|\mathcal{D}\right) \sim \mathcal{GP}\left(m(\mathbf{x}_*|\mathcal{D}), v(\mathbf{x}_*|\mathcal{D})\right). \quad 6.$$

tractable, with a Gaussian posterior distribution characterized by a mean (i.e., expected outcome) and variance (i.e., uncertainty) (see **Figure 1*b***). These analytically tractable computations have exact equivalencies to ANNs in the limit of an infinite number of hidden units (Neal 1996).

A key ingredient in Gaussian Process regression is the choice of kernel function, which provides an explicit similarity metric between any pair of inputs $\mathbf{x}$ and $\mathbf{x}'$ with desirable mathematical properties (Schölkopf & Smola 2002). This can capture inductive biases present in similarity-based theories (**Figure 1*g***), such as the common radial basis function (RBF) kernel (Equation 5), which assumes that similar inputs are likely to produce similar outcomes (**Figure 2*b***). However, there is a rich set of kernel functions to choose from, capturing different forms of inductive biases (Duvenaud et al. 2013). For instance, linear kernels make strong assumptions about linear relationships, periodic kernels encode cyclical trends, and graph kernels capture relational structure between discrete nodes on a graph (e.g., a diffusion kernel; Wu et al. 2021, Kondor & Lafferty 2002) (see **Figure 2*c***).

Gaussian Process regression can also be considered a hybrid model due to both similarity- and rule-based interpretations (Austerweil et al. 2015, Lucas et al. 2015). The similarity-based interpretation is straightforward, since the kernel explicitly encodes similarity between data points. However, the framework also lends support to two rule-based interpretations. The first is based on a mathematical property known as Mercer's (1909) theorem, describing how any kernel can be decomposed into a combination of basis functions (Austerweil et al. 2015, Lucas et al. 2015), each corresponding to an abstract rule. Just as any color can be decomposed into RGB 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. A second rule-based interpretation is based on the compositionality of Gaussian Process kernels (Duvenaud et al. 2013, Schulz et al. 2017). Multiple 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 function: a similarity metric defined for any pair of stimuli, used in Gaussian Processes

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. Composing multiple kernels thus allows for aggregating multiple hypotheses about the hidden structure of the environment. The Gaussian Process framework further formalizes this idea and injects the ability to reason about compositional rules as well. Thus, inversely analogous to Bayesian concept learning, Gaussian Processes operate on similarity-based computations but provide equivalent rule-based interpretations.

Value function approximation: a key method for generalization in reinforcement learning based on the expected value of different states or actions

2.3. Converging Historical Traditions

Given their similar historical trajectories, there is much to be gained from integrating theories of generalization across domains. In concept learning (**Figure 1*a–d***), Shepard's (1987) Universal Law of Generalization provides an influential similarity-based approach, where generalization is characterized as distance in psychological space, with stimuli embedded at closer distances more likely to produce the same responses (**Figure 2*a***). Yet, through a probabilistic application of rule-based mechanisms, a hybrid Bayesian concept learning framework (Tenenbaum & Griffiths 2001) (see **Figure 1*d***) can reproduce the same smooth gradient of generalization, showing how rule- and similarity-based mechanisms can be interpreted as two sides of the same coin (Pothos 2005, Austerweil et al. 2015). In function learning, there was an analogous trajectory of rule- and similarity-based theories culminating in hybrid approaches using Bayesian principles (**Figure 1*e–h***). Current hybrid theories of function learning based on Gaussian Process regression utilize similarity-based mechanisms implemented through kernel functions that capture inductive biases (e.g., the expectation that stimuli with similar features will yield similar outputs), yet they also provide rule-based interpretations and allow for compositional operations over different kernels (Lucas et al. 2015, Schulz et al. 2017). Ultimately, this convergence of concept and function learning has leveraged the strengths of both rule- and similarity-based approaches to illuminate the rich tapestry of human generalization. Building on these historical developments, we now turn to examining how principles of concept learning and function learning have informed new domains of generalization, tackling increasingly complex and challenging domains.

3. FROM LEARNING FUNCTIONS TO ACTING ON THE WORLD

Function learning has received relatively less attention and produced fewer experiments compared to concept learning. Yet there has been a revival of interest given the importance of value function approximation for generalization in RL problems (Sutton & Barto 2018). RL provides a computational framework for understanding learning in both biological and artificial systems, tracing its origins to early research on associative and instrumental learning (Thorndike 1911, Pavlov 1927, Skinner 1938). In a process framed as trial and error, RL agents learn to associate actions with expectations of reward through feedback from the environment, leading to gradual improvements in selecting reward-maximizing behaviors. However, no biological or artificial agent can try every possible action in most real-world settings, highlighting a growing impetus to better understand how humans generalize in real-world contexts, where the space of possible outcomes is too vast to be experienced exhaustively (Wu et al. 2018).

RL researchers have long grappled with the need to generalize past feedback to novel actions and states (Tesauro 1995). In complex games such as Go (Silver et al. 2016), the number of possible game states vastly outnumbers the number of atoms in the known universe. Thus, in scaling up to solve increasingly complex tasks, most modern RL algorithms commonly infer a value function mapping a vast space of potential actions or states to expectations of reward (i.e., value) (Sutton & Barto 2018). 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, guiding efficient exploration and action selection. Here, we review recent advances in understanding human generalization in RL settings that do not permit exhaustive exploration and we connect these findings to theories from function learning.

Bandit problem: experimental paradigm to investigate the trade-off between exploration and exploitation, in which decision makers repeatedly choose among options that yield probabilistic rewards to accumulate payoffs

3.1. Generalization in Reinforcement Learning Using Function Approximation

Several recent studies have investigated human generalization in RL problems with structured rewards (Wimmer et al. 2012, Norbury et al. 2018, Wu et al. 2018, Stojić et al. 2020). Rather than each option having independent reward distributions, as is commonly the case in bandit tasks, here structured bandit problems use correlated rewards (**Figure 3*a***). Thus, options in similar spatial locations (Wu et al. 2018) with similar abstract features (Norbury et al. 2018, Stojić et al. 2020, Wu et al. 2020), or nodes that are well connected on a graph (Wu et al. 2021), will yield similar rewards. This correlated reward structure provides traction for generalization, allowing participants to guide the selection of actions toward promising regions of the search space. In

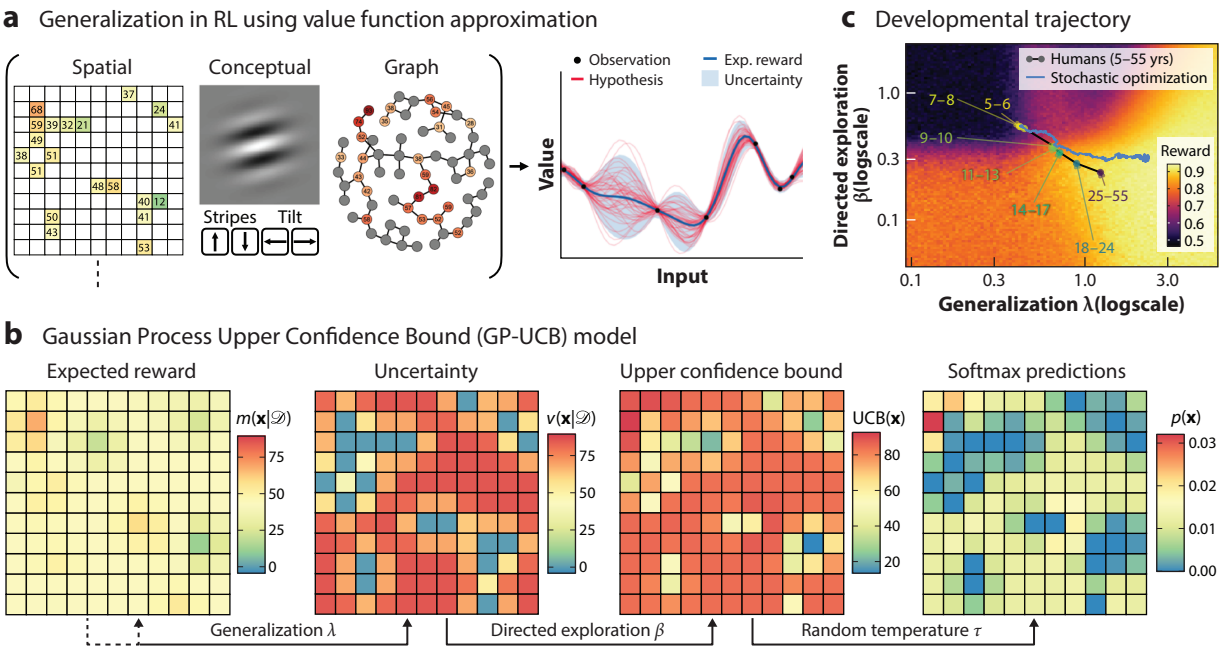


Figure 3

Generalization in reinforcement learning (RL). (*a*) Generalization using value function approximation. (*Left*) Bandit tasks with structured rewards, where similar locations, feature combinations, or connected nodes generate similar rewards. (*Right*) Generalization modeled using Gaussian Process regression to infer a value function, mapping a potentially infinite range of actions or states to probabilistic predictions about expected reward and subjective uncertainty. (*b*) Overview of the Gaussian Process Upper Confidence Bound (GP-UCB) model in a spatial bandit task. Conditioned on the observations in panel *a* (*left*), the Gaussian Process model makes predictions about expected reward $m(\mathbf{x})$ and uncertainty $v(\mathbf{x})$, where the parameter λ governs the extent that past observations generalize to new options. UCB sampling combines the expected rewards $m(\mathbf{x}|\mathcal{D})$ and uncertainty $v(\mathbf{x}|\mathcal{D})$ using a weighted sum, where the parameter β defines the value of exploring uncertain options relative to exploiting high-reward expectations. $\mathcal{D}$ stands for observed data. Lastly, UCB values are transformed into probabilistic predictions of where the participant will search next using a softmax function, where the temperature τ governs the amount of random exploration. (*c*) The developmental trajectory of human learners (5–55 years of age) resembles stochastic optimization over GP-UCB parameters. The labeled dots are the median parameter estimates from human subjects, while the blue line is the trajectory of the best-performing stochastic optimization algorithm (Giron et al. 2023).

this body of research, human generalization is typically best characterized by the same Gaussian Process model as in traditional function learning tasks, but based on the implicit learning of a value function that is then used to predict actions (**Figure 3*b***).

One important distinction between generalization in RL settings and in traditional function learning is that the goal in RL is not necessarily to learn the true underlying value function. Rather, one needs to balance the explore–exploit dilemma (Mehlhorn et al. 2015) by both exploring uncertain options to acquire information and exploiting options with high expectations of reward to maximize immediate gains. Thus, a fundamental challenge in RL is to determine what information should be acquired given current beliefs (i.e., active learning; Nelson 2005). Two prominent mechanisms of exploration are random exploration (e.g., flipping a coin to decide) and uncertainty-directed exploration (i.e., curiosity toward subjective uncertainty), which play a dissociable role in human exploration (Wilson et al. 2014; Wu et al. 2018, 2022a; Cogliati Dezza et al. 2019) and have different neural signatures (Zajkowski et al. 2017). In RL settings, the Gaussian Process model addresses the explore–exploit dilemma by making Bayesian predictions about expected rewards, including quantifications of uncertainty (see the right side of **Figure 3*a***). These two components can be used to implement policies that both balance the exploration–exploitation dilemma and predict human behavior.

Figure 3*b* illustrates how the reward expectations and uncertainty estimates of a Bayesian function learning model are combined to predict choices in a spatially correlated bandit problem (Wu et al. 2018). The best account of human choice behavior combines Gaussian Process (GP) regression as a model of value generalization with Upper Confidence Bound (UCB) sampling (Auer 2002), which quantifies the value of a choice option by adding an uncertainty-based "exploration bonus" to reward expectations. Taken as a whole, the GP-UCB model demonstrates how generalization and exploration mechanisms interact to guide decision making in RL. Furthermore, additional studies have fit the GP-UCB model on choices and then used it to predict out-of-sample judgments that participants made about the expected reward of unexplored options along with subjective confidence ratings (Wu et al. 2020, 2021). Thus, not only do participants select actions "as-if" they are using a form of Bayesian value function approximation but also the same computations can be used to predict their judgments, showing a correspondence between implicit value generalization in RL and explicit function learning in psychology.

3.2. Developmental Changes in Generalization and Exploration

By casting generalization as Bayesian function learning, the GP-UCB model has provided a powerful lens for understanding developmental changes in learning. Human development is often likened to a "cooling off" process (Gopnik et al. 2015), in analogy to mechanisms of stochastic optimization used in modern machine learning models. Like a heated piece of metal that becomes harder to manipulate as it cools off, stochastic optimization starts off highly flexible and open to solutions that might not seem very good at first. As it cools down, however, the algorithm becomes less flexible and more selective in favoring only local improvements. This analogy is appealing, since children are highly stochastic and flexible learners, with the randomness of their choices (Bonawitz et al. 2014) and hypotheses (Buchsbaum et al. 2012, Denison et al. 2013, Lucas et al. 2014, Gopnik et al. 2017) gradually diminishing over the lifespan.

Yet there is ambiguity in how to interpret this verbal analogy. The most common interpretation of cooling off is as a unidimensional transition from exploration to exploitation (Gopnik 2020), focusing primarily on a reduction in random exploration. While past work has indeed found differences in random exploration among different age groups (Somerville et al. 2017, Schulz et al. 2019, Blanco & Sloutsky 2021, Meder et al. 2021), this seems to be only part of the picture.

Developmental changes are also found in more systematic, uncertainty-directed exploration (Somerville et al. 2017, Schulz et al. 2019, Blanco & Sloutsky 2021), along with aspects of belief integration and generalization about novel options (Van den Bos et al. 2012, Blanco et al. 2016).

To provide a concrete test of the cooling off analogy, Giron et al. (2023) directly compared the trajectory of human learners (aged 5 to 55) with that of various optimization algorithms (**Figure 3*c***). The results showed that cooling off does not apply only to the single dimension of randomness. Rather, development resembles an optimization process in the space of learning strategies: What begins as large tweaks in the parameters that define learning during childhood plateaus and converges in adulthood. While the developmental trajectory of human learning strategies is strikingly similar to the best-performing algorithms (**Figure 3*c***), none of the tested algorithms discovered reliably better regions of the strategy space than adult participants, suggesting a remarkable efficiency of human development.

In sum, integrating principles of generalization with the exploration–exploitation dilemma from RL has proven to be a productive approach for understanding behavior in increasingly complex scenarios. The use of structured reward environments constitutes an important step toward characterizing the psychological mechanisms of adaptive behavior in more complex and naturalistic settings (Wise et al. 2024), where the environment does not permit exhaustive experience of all options. Here, generalization via Bayesian function learning shines a guiding light by predicting both where to exploit and where to explore.

Cognitive map: a mental representation of the structure of the environment used for navigation, learning, and generalization

4. FROM LEARNING CONCEPTS TO LEARNING STRUCTURE

Human generalization is much deeper than just comparing features at face value. Rather, generalization also depends on the relational structure and temporal dynamics of the environment, which are often hidden and need to be inferred. Research on structure learning can be broadly divided into two traditions. The first originates from Tolman's (1948) pioneering notion of a cognitive map. Research in this domain has extensively studied spatial navigation in the hippocampal-entorhinal system (Moser et al. 2014, Epstein et al. 2017, Whittington et al. 2022), which has since been extended to a wide range of nonspatial modalities and domains (Behrens et al. 2018). The second tradition, known as Bayesian structure induction (Kemp & Tenenbaum 2008), builds on a similar formalism as Bayesian concept learning (Tenenbaum & Griffiths 2001), where explicit, rule-like hypotheses about structure can be inferred from observed data, reflecting our ability to discern patterns and regularities in the environment. While these traditions are based on different theoretical foundations, here we show that they share a common framework of similarity-based mechanisms for learning rule-like hypotheses about structure.

4.1. Cognitive Maps

Originally proposing this framework as an alternative to stimulus-response learning, Tolman (1948) found that rats could rapidly adapt to new situations (e.g., choosing the second shortest path in a maze when the shortest path was blocked) and to new goals (e.g., efficiently navigating to food rewards placed in novel locations of a familiar maze). These results suggested the rats had generalized their experiences based on establishing a "field map of the environment" (Tolman 1948). Today, this 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 the encoding of spatial orientations, boundaries, and distance to objects (for reviews, see Moser et al. 2014, Epstein et al. 2017, Peer et al. 2021). As Tolman originally speculated, cognitive maps are not restricted to representing spatial structure. Rather, the same neural machinery used for spatial navigation also encodes

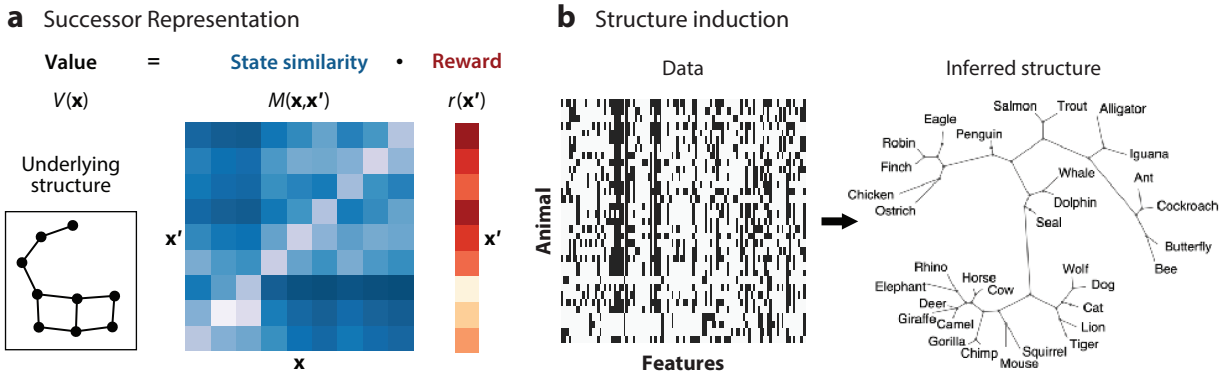


Figure 4

Structure learning. (*a*) The Successor Representation (Dayan 1993) defines a decomposition of a Temporal Difference–learning (Sutton & Barto 2018) value function $V(\mathbf{x})$ into a similarity matrix $M(\mathbf{x}, \mathbf{x}')$ based on expected future state transitions and the singular rewards of each state $r(\mathbf{x}')$. The state similarities are a function of the underlying structure (*left*) and the agent's policy, and they allow for generalization via a linear form of value function approximation. (*b*) Bayesian structure induction (Kemp & Tenenbaum 2008) uses Bayesian principles to infer the underlying structure (e.g., a taxonomy) that gave rise to the observed relational data (e.g., animals and their shared features).

relational and structural knowledge across a wide range of domains, including social relationships (Tavares et al. 2015), smells (Bao et al. 2019), abstract visual features (Constantinescu et al. 2016), and the connectivity of hidden graph structures (Garvert et al. 2017).

One influential account of structure learning in the hippocampal-entorhinal system is the Successor Representation (SR; Dayan 1993, Stachenfeld et al. 2017). Originally developed as a method to improve the generalization of Temporal Difference learning (Sutton & Barto 2018), the SR describes a decomposition of the value function into a similarity matrix and the singular rewards of each state (**Figure 4*a***). The similarity matrix quantifies the similarity between each pair of states based on expected future state transitions, influenced by both the structure of the environment and the agent's behavioral policy (i.e., how the agent moves around in the environment), and thus corresponds to an explicit, graph-like representation of the environment (Peer et al. 2021). The value generalizations predicted by the SR—taking the form of a linear function of state similarities and reward observations—capture the underlying transition dynamics and connectivity structure of the environment, with stronger generalizations between well-connected states. Related methods using kernel similarity (Gershman et al. 2017, Wu et al. 2021) (see **Figure 2*c***) rather than SR similarity operate on similar principles, with exact equivalencies in special cases (Machado et al. 2018). For example, Garvert et al. (2023) showed that a Gaussian Process kernel can be approximated by the SR of visited states in an open environment and then used to successfully predict human choices in a bandit task, illustrating a continuity between cognitive maps and value generalization using function learning.

Successor Representation (SR): a reinforcement learning model using anticipated future states of the environment for predictive generalization

Overall, the SR provides an elegant and simple theory of structure learning within the RL framework, where similarity-based representations acquired through associative learning enable structure-informed generalization via value function approximation. However, the slowness of this learning process may fall short of explaining the full efficiency with which humans learn relational structure. Other recent theories of cognitive map learning such as the Tolman-Eichenbaum Machine (TEM; Whittington et al. 2020) combine path integration with conjunctive memory to more efficiently learn latent structure. While the TEM is capable of transferring learned structures to new environments, it still cannot infer entirely novel structures. In contrast, humans can

reason compositionally about new relational structures that they have never experienced before (Mark et al. 2020, Rubino et al. 2023). Consider how you can imagine novel food combinations that have never been observed [e.g., tea-flavored jelly (Barron et al. 2013) or broccoli-flavored ice cream (Gershman et al. 2017)] 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 highlights the necessity for more explicit, rule-based theories of cognitive map learning, which is an area ripe for further exploration.

4.2. Structure Induction

Structure induction (Kemp & Tenenbaum 2008, Meder et al. 2014, Lynn & Bassett 2020, Zhou et al. 2024a) provides an alternative approach to inferring the underlying structure or organizational pattern in a set of observations or data—for example, inferring the taxonomy of different animals based on their shared features (**Figure 4*b***). Animals with similar features can be expected to occupy closely connected positions in a taxonomy, yet there is a large hypothesis space of possible configurations. Here, Bayesian structure induction (Kemp & Tenenbaum 2008) uses a similar mathematical formalism as Bayesian concept learning (Tenenbaum & Griffiths 2001), based on describing a distribution of rule-like hypotheses, which are evaluated based on their similarity to the observed data. Instead of defining hypotheses about category boundaries, structure induction defines an inference process operating 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 according to its likelihood of generating the observed data (Kemp & Tenenbaum 2008).

Although Bayesian inference about latent structure is often intractable when scaled to complex problems, hybrid models of structure induction circumvent this problem by incorporating similarity-based mechanisms. One notable method (Kemp & Tenenbaum 2008, 2009) used to evaluate the likelihood of each candidate hypothesis is identical to a Gaussian Process.[1] Here, each candidate hypothesis is used to parameterize a graph kernel (Zhu et al. 2003) (**Figure 2*c***), and simulated observations are sampled from the Gaussian Process prior (Equation 4). Higher similarity between generated and observed data corresponds to a higher likelihood for the hypothesized structure.

Thus, structure learning and function learning can be seen as complementary problems, with a hybrid approach to Bayesian structure induction relying on the same Gaussian Process computations as in function learning (Lucas et al. 2015) and value generalization settings (Wu et al. 2021). Hypotheses about rule-like structures are used to define a similarity metric based on a Gaussian Process kernel and to simulate data. Comparisons between the simulated and observed data facilitate inference about which structures are most likely. Once the structure has been inferred, these same computations can be reused to generalize about novel outcomes (Kemp & Tenenbaum 2008) and guide exploration (Wu et al. 2021) in structured environments. For instance, this complementary relationship has also been leveraged to generalize about novel properties of the data (i.e., property induction; Kemp & Tenenbaum 2009). Given a set of binary features of various animals (**Figure 4*b***), structure induction can be used to infer the underlying taxonomy structure. Once a posterior distribution over structure has been defined, the same Gaussian Process function learning approach (with an additional binarization of outcome variables) is used to infer the probability of novel features. If one were to learn a new fact about squirrels (e.g.,

[1]The authors refer to this as a Gaussian Markov Random Field (Zhu et al. 2003), which is a multivariate Gaussian distribution identical to a Gaussian Process prior (Equation 4).

their front teeth never stop growing), one might be more likely to generalize this fact to similar animals, such as mice, but less likely to generalize it to more dissimilar animals, such as penguins.

In summary, structure induction offers a prime example of the complementarity between rule- and structure-based mechanisms. Rule-based computations over a distribution of hypothesized structures offer the possibility of rapid generalization. Yet the intractability of Bayesian inference can be side-stepped through sample-based approximations, using Bayesian function learning operating over similarity-based computations. Together, these complementary approaches to generalization support both the inference of latent structure and the use of this structure to infer new features and outcomes.

5. GENERAL DISCUSSION

We have traced the development of psychological theories of generalization, from foundational research on concept learning and function learning to more modern domains of RL and latent structure induction. Throughout this long history, continued debates between rule- and similarity-based theories have been reconciled through the development of hybrid models, often based on Bayesian principles. The ongoing success of hybrid models suggests that accommodating both rule- and similarity-based representations is central to explaining human generalization.

Yet, each approach makes computational commitments to a specific representational format, offering distinct advantages. Similarity provides a flexible and efficient approach to generalization, relating new situations to prior experiences and leveraging relational knowledge when the underlying structure is known. In turn, rules unlock compositionality, facilitating generalization and inference about novel structures, which is exemplified in Bayesian structure induction. However, there may also be exchangeability between rule-based and similarity-based mechanisms of generalization, suggesting a dynamic interplay that enables adaptive learning through hybrid approaches that blend both strategies. We first explore these themes before plotting out a trajectory for the future of research on generalization.

5.1. Rules Unlock Compositionality but Are Challenging to Learn

Rule-based mechanisms are foundational to our understanding of generalization, drawing upon a rich history of theoretical and empirical research (Bruner et al. 1956, Ashby & Gott 1988; for a review, see Ashby & Maddox 2005). These mechanisms are particularly effective in structured domains, where the precision of rules facilitates rapid, one-shot generalization (Dasgupta et al. 2022). Whether taught pedagogically (e.g., "i before e except after c") or learned through experience (e.g., "talking loudly in the library is forbidden"), rules represent explicit hypotheses about regularities of the environment extracted from data (Reber & Lewis 1977). 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 (Allen et al. 2020, Tsividis et al. 2021), where agents generate hypotheses about the underlying rules governing the environment to inform learning and exploration (e.g., "keys open doors, but only if the colors match" in game environments; Pouncy et al. 2021).

This ability to reason about and use rules thus unlocks an unrivaled capacity of human intelligence, since rules allow for compositional and syntactic manipulation (Piantadosi et al. 2016, Dehaene et al. 2022). Indeed, the power of logic and mathematics can be thought of as nothing more than the manipulation of syntactic rules (Newell & Simon 1976). Thus, rule-based mechanisms unlock the ability to compositionally combine multiple rules or substructures to generate an infinitely productive space of potential hypotheses. Recent advances in program induction (Ellis

et al. 2023)—using similar computations as Bayesian concept learning (Tenenbaum & Griffiths 2001) and structure induction (Kemp & Tenenbaum 2008)—indicate a promising framework for modeling how humans infer generative rule-like structure from data, providing a modern interpretation of Fodor's (1975) Language of Thought. However, the compositionality of rules also creates a combinatorial explosion of possible hypotheses, making search and inference increasingly difficult (Fränken et al. 2022, Zhou et al. 2024b). Thus, despite the utility of rule-based mechanisms, open challenges lie in their complexity and in the demands they place on cognitive resources for generating and testing new hypotheses (Rubino et al. 2023).

5.2. Similarity Is Flexible but Can Be Arbitrary

Similarity-based mechanisms for generalization are ubiquitous in psychology (Tversky 1977, Shepard 1987, Tenenbaum & Griffiths 2001, Chater & Vitányi 2003). 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 is a powerful principle of generalization, and it can be flexibly applied to a wide range of domains. While such notions are historically defined based on feature comparisons or by appealing to some abstract psychological space (Shepard 1987), recent advances have expanded these mechanisms to capture rich relational structures (Wu et al. 2021) based on network connections (Tavares et al. 2015, Lau et al. 2020) or environmental dynamics (Stachenfeld et al. 2017, Garvert et al. 2023). Thus, similarity-based theories of generalization are being extended to increasingly structured environments.

However, these mechanisms are not without drawbacks. It is far from straightforward to 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 (Goodman 1972). 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 (Medin et al. 1993). 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. Thus, the endless ways in which different stimuli can be compared has led to the criticism that similarity is too flexible (Murphy & Medin 1985), potentially undermining its utility as a concept in psychology. The context-dependent nature of human similarity judgments can also lead to paradoxical conclusions, as illustrated by violations of logical axioms like the law of triangle inequality (Tversky 1977). Moreover, while recent theories of rational attention have proposed associative learning mechanisms for gradually ignoring reward-irrelevant features (Radulescu et al. 2021), this approach is only feasible for simple stimuli with a handful of predefined features. In more naturalistic settings, stimuli may have a potentially innumerable set of features, making it infeasible to gradually prune irrelevant features from an infinite set. These complexities illustrate the nuanced and sometimes contradictory nature of similarity-based generalization in human cognition.

Similar challenges also apply when defining similarity representations over latent structure that share context- and goal-dependent assumptions about which features are relevant. For instance, the development of Darwin's Tree of Life was rooted in targeted observations about features that were shared or differed between species (Doolittle & Bapteste 2007), and dimensional accounts of psychopathology similarly aim to capture shared symptom patterns across mental illnesses based on a targeted subset of features (Kotov et al. 2021). Thus, while latent structure plays a pivotal role in generalization by complementing similarity-based inferences, it shares some of the very same challenges that arise in defining relevant features for computing similarity. This intertwined

nature of similarity and structure highlights both their importance and the enduring challenges concerning their role in human cognition and generalization.

Model-free reinforcement learning: category of reinforcement learning methods using reward outcomes to learn a behavioral policy and value function without simulating future scenarios

Model-based reinforcement learning: more complex form of reinforcement learning, which builds a model of the environment to simulate and plan future actions

5.3. Integrating Rules and Similarity

We have highlighted the relative advantages and disadvantages of rule- and similarity-based mechanisms of generalization. However, the success of hybrid approaches suggests it is not one or the other. Rather, there is likely a degree of exchangeability between rules and similarities, involving transformations from one currency to the other (Cushman 2020). This is not a new concept. In RL, model-based representations of the environment can be used to rationally plan out actions (Miller et al. 2017), but in the process, new value and policy representations are constructed, supporting future model-free action selection (Kool et al. 2018). In social learning, observed actions can be unpacked via inverse RL (IRL; Jara-Ettinger 2019) to infer latent model-free and model-based representations assumed to have generated the behavior (Wu et al. 2022b). Thus, the caching of past computations (i.e., amortization; Dasgupta et al. 2018) and inference via IRL provide two mechanisms by which the representations involved in model-free and model-based RL are exchanged and combined with one another (Cushman 2020, Wu et al. 2022b). Our current theories in this domain suggest that we use a mixture of strategies, composing elements from each mechanism into an adaptive mixture of representations (Huys et al. 2015, Keramati et al. 2016, Russek et al. 2017).

Are rule- and similarity-based representations exchangeable in a similar sense? Rule-based representations about category boundaries, functional forms, or the structure of the environment can inform or be directly used to define similarity representations. We have shown how rule-like hypotheses about the structure of some latent graph can be used to define a similarity matrix using a graph kernel (**Figure 2*c***) to infer rule-like representations about the latent structure (Kemp & Tenenbaum 2008), predict novel features outside of the training data (Kemp & Tenenbaum 2009), or perform value generalization in an RL setting (Wu et al. 2021). In the other direction, we have also shown how similarity-based representations support the inference of rule-like hypotheses about latent structure. The SR (Dayan 1993) leverages simple associative learning mechanisms to learn a similarity matrix, corresponding to a rule-like hypothesis about the underlying latent structure of the environment. Even more directly, hybrid models of Bayesian structure induction (Kemp & Tenenbaum 2008) have relied on similarity-based computations using Gaussian Process kernels to simulate data under each hypothesized graph structure. Thus, learned rules can be cached as similarity representations, facilitating rapid and efficient generalization. Meanwhile, inferring rules and structure can be supported by sample-based approximations, where each candidate hypothesis can be used to construct a similarity representation to perform tractable inference.

5.4. The Future of Generalization

Having surveyed the past and present, we now turn our attention to the future. First, we propose a new integration of rule- and similarity-based mechanisms for structure learning in RL settings, combining their relative strengths and leveraging the exchangeability of representations to achieve a more comprehensive framework of generalization. Second, we point out fundamental connections between Gaussian Process regression and theories of episodic memory, which suggest the potential for developing boundedly rational models of generalization to account for cognitive limitations. Third, there is still a need to explore generalization in environments that more closely resemble real-world conditions, requiring the integration of individual and social information. By addressing these issues, future research will continue a long and always central line

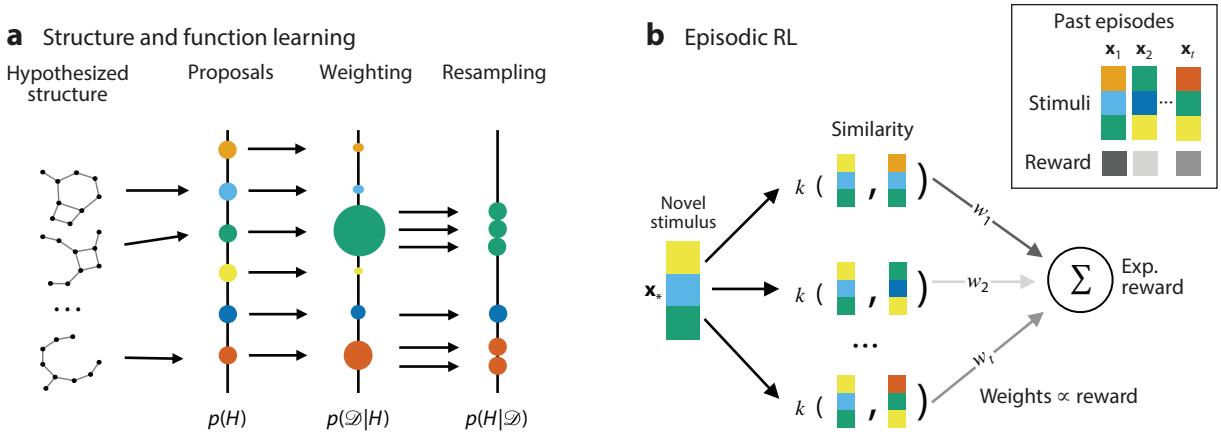


Figure 5

Future directions. (*a*) Integrating structure learning and function learning under a common framework. Prior hypotheses $p(H)$ about structure are reweighted by the likelihood of the data $p(\mathcal{D}|H)$. The posterior $p(H|\mathcal{D})$ is then defined by resampling hypotheses proportional to their likelihood, which allows for predictive generalization based on a sequentially refined distribution over hypothesized structure. (*b*) The episodic reinforcement learning (RL) framework provides a different conceptualization of the computations in a Gaussian Process, with exact equivalencies to expected reward predictions (but not uncertainty). Here, similarity is computed between new stimuli and past episodes, which are then weighted by rewards and summed up.

of inquiry seeking to understand how humans adapt and continually improvise to adapt to novel situations.

5.4.1. Combining structure learning and function learning. Rather than accepting a duality of interpretation as the final synthesis of rule- and similarity-based mechanisms, we propose that future models of generalization could provide a more complete unification, utilizing the strengths of each mechanism. 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 infers structure while simultaneously performing predictive generalization.

Since Gaussian Processes play a key role in the computations of Bayesian structure induction (Kemp & Tenenbaum 2008), a future model (**Figure 5*a***) could 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; Kemp & Tenenbaum 2008, Wingate et al. 2013, Fränken et al. 2022, Ellis et al. 2023, Rubino et al. 2023). Each hypothesized structure can then be used to parameterize a graph kernel (**Figure 2*c***), where a Gaussian Process using similarity-based mechanisms can be used to both predict new outcomes and evaluate the likelihood of a given hypothesis (as in Kemp & Tenenbaum 2008). Both sample-based (Doucet et al. 2011, Speekenbrink 2016) and neural network (Deleu et al. 2022) approximations provide tractable computations of a Bayesian posterior distribution over hypotheses (about structure), which adapt to new data and facilitate active learning. With rules providing the structure and similarity providing the canvas, generalization combining both mechanisms can achieve both flexibility and efficiency.

5.4.2. Generalization with limited resources. Although originating as a machine learning technique, Gaussian Process regression has direct links to psychological theories integrating RL mechanisms with episodic memory (Lengyel & Dayan 2007, Gershman & Daw 2017). In this

light, Gaussian Processes can be understood as a Bayesian extension of episodic RL (Gershman & Daw 2017, Botvinick et al. 2019). In episodic RL (**Figure 5*b***), an agent stores episodic memories about previously encountered stimuli and their associated rewards. To predict the value of some novel stimuli, one first computes similarity to each previously encountered episode; then, the reward value for each episode is multiplied by its similarity to the novel stimuli and then summed up. In other words, generalization is performed through inferring similarity-weighted expectations, where more similar episodes exert more influence on how their rewards generalize to the novel stimuli (reminiscent of classic exemplar-based theories of concept learning; Nosofsky 1986, Kruschke 1992).

Episodic reinforcement learning: a framework for value function approximation that compares novel states, actions, or stimuli to previously encountered episodes

When using a kernel function to compute similarity (Gershman & Daw 2017), episodic RL is equivalent to the posterior mean of a Gaussian Process (Jäkel et al. 2008, Wu et al. 2021). Furthermore, when using an RBF kernel (**Figure 2*b***) as the similarity metric, episodic RL is equivalent to an RBF network, which has featured prominently in machine learning approaches to value function approximation (Jäkel et al. 2008, Sutton & Barto 2018) and as a theory of human generalization in the visual and motor systems (Poggio & Bizzi 2004). Thus, while the mathematics of Gaussian Process regression may seem unfamiliar to psychology, the underlying computations reoccur in numerous psychological theories of learning and generalization. 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 exploration (Wilson et al. 2014, Gershman 2018, Wu et al. 2018, Giron et al. 2023) and subjective confidence judgments (Wu et al. 2020, 2021).

The relationships between Gaussian Process regression and episodic RL provide pathways for further integrating psychological and computational theories. For instance, to investigate the role of memory limitations in value generalization, one can induce memory load by removing information about previous choices and their outcomes (e.g., withholding observations from the grid shown in **Figure 3*a***; Breit et al. 2022). In this case, learners would be reliant on episodic memory of past choices to generalize previous experiences, thus offering opportunities for studying how forgetting distorts our patterns of generalization.

5.4.3. What is still missing? Here, we have explored an expanding core of research on human generalization. From early work studying stimulus categorization and function learning, we have traced a continuity of mechanisms to new domains, such as active learning in RL and latent structure learning. However, the full scope of human generalization is still clearly beyond our current theories. Consider a chef figuring out how to substitute a missing ingredient in a recipe or a biologist identifying new species in an unexplored habitat. Generalization in both settings is informed by an interplay of rules and similarity—concerning the interactions between different foods and cooking techniques or the interplay of biological traits, ecological niches, and reproductive success. Yet, the open-ended complexity of features to evaluate (Wise et al. 2024) and actions to consider (Moskvichev et al. 2023) presents open challenges for our current theories. Additionally, chefs, biologists, and humans in all walks of life primarily learn from one another. While psychological research has often focused on studying isolated individuals learning from the environment (imagine a Skinner box as a canonical example), there is evidence of distinct mechanisms when learning from other people (Ho et al. 2017) compared to learning from the environment. Thus, future theories of human generalization must also account for more open-ended and socially embedded problems.

On one hand, psychological research has been continually expanding to investigate learning and generalization in more complex and open-ended problems. For instance, there is currently great interest in studying generalization in the Abstraction and Reasoning Corpus (ARC; Chollet

2019). The ARC challenge is comprised of visual grids representing an abstract concept (input), with the decision maker tasked with constructing an output grid corresponding to the input. This can be seen as a type of function learning problem requiring strong inductive biases about the generative process, since one needs to generate solution grids instead of only selecting from possible answers. These challenges may play a key role in explaining why artifical intelligence (AI) approaches, including Large Language Models, have yet to come close to human performance (Moskvichev et al. 2023). Thus, there is a promising future for efforts directed toward studies of generalization that integrate the complexity and open-endedness inherent in real-world decision-making environments.

On the other hand, a promising yet underexplored area is the integration of individual and social generalization mechanisms (Witt et al. 2024, Wu et al. 2023). Far from being a peripheral feature, the capacity for social learning is often proposed as being the defining characteristic of human intelligence (Henrich 2016, Heyes 2018), differentiating us from other animals and AI (Wu et al. 2022b). Yet, research on generalization has commonly focused on individual learning in a vacuum. In many real-world contexts, however, we are surrounded by social information, which can greatly inform our generalization and decision-making processes. For instance, we could observe which menu items other customers order in a restaurant and use that to inform our own choices. In such scenarios, individual and social learning mechanisms exhibit a dynamic interplay (Wu et al. 2023), working in tandem to achieve efficient generalization. Here, there are also new applications for familiar concepts from individual generalization, since social information cannot always be taken verbatim but needs to account for differences in individual preferences, abilities, and goals (Witt et al. 2024). Additionally, our ability to communicate via language in social settings offers new advantages for rule-based mechanisms, since they can be easily transmitted to one other (Wu et al. 2024). Such scenarios offer a promising avenue for investigating how humans generalize and make decisions in real-world contexts, where social information plays a vital role in shaping adaptive behavior.

5.5. Conclusions

Human generalization has long been considered a hallmark of our unique cognitive abilities, with Roger Shepard famously proclaiming that the first general law of psychology should be a law of generalization. Here, we have traced the development of theories of generalization, illustrating a continuity of formerly competing mechanisms—rules and similarity—culminating in hybrid approaches. Ultimately, the future of generalization will hold new and exciting ideas but still carry echoes of perennially reoccurring principles from history.

SUMMARY POINTS

1. Rules and similarity are foundational concepts across the entire expanse of research on how humans generalize from limited experiences to novel situations.
2. Hybrid models include elements of both rule- and similarity-based approaches, providing a unified computational framework for investigating human generalization across diverse contexts.
3. Gaussian Process function learning, coupled with uncertainty-directed exploration, provides a model of generalization and active learning in a wide range of reinforcement learning problems with large decision spaces.

4. Structure learning supports similarity-based generalization by representing latent relational structure and the temporal dynamics of the environment; conversely, similarity-based mechanisms may play a key role in learning latent structure.
5. Rule- and similarity-based representations have complementary advantages, with an exchangeability between these representations offering insights into how humans simultaneously display flexible and compositional generalization.

FUTURE ISSUES

1. The mechanisms underlying the integration of rule- and similarity-based generalization are still unknown, and the dynamic interplay between these processes should be explored in different learning contexts.
2. Combining structure induction with models of active learning is a promising direction for developing more comprehensive models of generalization that leverage the advantages of both rule- and similarity-based mechanisms.
3. Exploring the relationship between Gaussian Process regression and episodic reinforcement learning provides a foundation for investigating how cognitive constraints, like working memory load, influence generalization under bounded rationality.
4. Investigating how humans and computational models navigate and generalize in high-dimensional spaces will require new methods for identifying and prioritizing relevant features and relevant hypotheses.
5. The role of social learning has been underrepresented in theories of generalization, with a need for new research studying how social and cultural contexts influence the mechanisms of generalization.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

AUTHOR CONTRIBUTIONS

All authors contributed to the conception of the article and approved the final version. C.M.W. wrote the manuscript with contributions from B.M. and created the visualizations.

ACKNOWLEDGMENTS

C.M.W. is supported by the German Federal Ministry of Education and Research, Tübingen AI Center (FKZ: 01IS18039A), and funded by the Deutsche Forschungsgemeinschaft under Germany's Excellence Strategy (EXC2064/1–390727645). The authors thank the members of the Human and Machine Cognition Lab for their helpful feedback.

LITERATURE CITED

Allen KR, Smith KA, Tenenbaum JB. 2020. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *PNAS* 117(47):29302–10

Ashby FG, Alfonso-Reese LA, Turken AU, Waldron EM. 1998. A neuropsychological theory of multiple systems in category learning. *Psychol. Rev.* 105(3):442–81
Ashby FG, Gott RE. 1988. Decision rules in the perception and categorization of multidimensional stimuli. *J. Exp. Psychol. Learn. Mem. Cogn.* 14(1):33–53
Ashby FG, Maddox WT. 2005. Human category learning. *Annu. Rev. Psychol.* 56:149–78
Auer P. 2002. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.* 47:235–56
Austerweil JL, Gershman SJ, Tenenbaum JB, Griffiths TL. 2015. Structure and flexibility in Bayesian models of cognition. In *Oxford Handbook of Computational and Mathematical Psychology*, ed. JR Busemeyer, Z Wang, JT Townsend, A Eidels, pp. 187–208. Oxford, UK: Oxford Univ. Press
Bao X, Gjorgieva E, Shanahan LK, Howard JD, Kahnt T, Gottfried JA. 2019. Grid-like neural representations support olfactory navigation of a two-dimensional odor space. *Neuron* 102(5):1066–75.e5
Barron HC, Dolan RJ, Behrens TEJ. 2013. Online evaluation of novel choices by simultaneous representation of multiple memories. *Nat. Neurosci.* 16(10):1492–98
Behrens TEJ, Muller TH, Whittington JCR, Mark S, Baram AB, et al. 2018. What is a cognitive map? Organizing knowledge for flexible behavior. *Neuron* 100(2):490–509
Blanco NJ, Love BC, Ramscar M, Otto AR, Smayda K, Maddox WT. 2016. Exploratory decision-making as a function of lifelong experience, not cognitive decline. *J. Exp. Psychol. Gen.* 145(3):284–97
Blanco NJ, Sloutsky VM. 2021. Systematic exploration and uncertainty dominate young children's choices. *Dev. Sci.* 24(2):e13026
Bonawitz E, Denison S, Gopnik A, Griffiths TL. 2014. Win-stay, lose-sample: a simple sequential algorithm for approximating Bayesian inference. *Cogn. Psychol.* 74:35–65
Botvinick M, Ritter S, Wang JX, Kurth-Nelson Z, Blundell C, Hassabis D. 2019. Reinforcement learning, fast and slow. *Trends Cogn. Sci.* 23(5):408–22
Bowman CR, Iwashita T, Zeithamova D. 2020. Tracking prototype and exemplar representations in the brain across learning. *eLife* 9:e59360
Brehmer B. 1974. Hypotheses about relations between scaled variables in the learning of probabilistic inference tasks. *Organ. Behav. Hum. Perform.* 11(1):1–27
Brehmer B. 1976. Learning complex rules in probabilistic inference tasks. *Scand. J. Psychol.* 17(1):309–12
Breit S, Sakaki M, Murayama K, Wu CM. 2022. In search of lost memories: modeling forgetful generalization. In *Proceedings of the 15th Biannual Conference of the German Society for Cognitive Science*. Freiburg, Ger.: Albert-Ludwigs-Universität
Bruner JS, Goodnow JJ, Austin GA. 1956. *A Study of Thinking*. New York: John Wiley & Sons
Buchsbaum D, Bridgers S, Skolnick Weisberg D, Gopnik A. 2012. The power of possibility: causal learning, counterfactual reasoning, and pretend play. *Philos. Trans. R. Soc. B* 367(1599):2202–12
Busemeyer JR, Byun E, Delosh EL, McDaniel MA. 1997. Learning functional relations based on experience with input-output pairs by humans and artificial neural networks. In *Knowledge, Concepts and Categories*, ed. K Lamberts, D Shanks, pp. 408–37. Cambridge, MA: MIT Press
Carroll JD. 1963. Functional learning: the learning of continuous functional mappings relating stimulus and response continua. *ETS Res. Bull. Ser.* 1963(2):i–144
Chater N, Vitányi PMB. 2003. The generalized universal law of generalization. *J. Math. Psychol.* 47(3):346–69
Chollet F. 2019. On the measure of intelligence. arXiv:1911.01547 [cs.AI]
Cogliati Dezza I, Cleeremans A, Alexander W. 2019. Should we control? The interplay between cognitive control and information integration in the resolution of the exploration-exploitation dilemma. *J. Exp. Psychol. Gen.* 148(6):977–93
Constantinescu AO, O'Reilly JX, Behrens TEJ. 2016. Organizing conceptual knowledge in humans with a gridlike code. *Science* 352(6292):1464–68
Csibra G, Gergely G. 2009. Natural pedagogy. *Trends Cogn. Sci.* 13(4):148–53
Cushman F. 2020. Rationalization is rational. *Behav. Brain Sci.* 43:e28
Cybenko G. 1989. Approximation by superpositions of a sigmoidal function. *Math. Control Sign. Syst.* 2(4):303–14
Dasgupta I, Grant E, Griffiths T. 2022. Distinguishing rule and exemplar-based generalization in learning systems. In *Proceedings of the 39th International Conference on Machine Learning*, ed. K Chaudhuri, S Jegelka, L Song, C Szepesvari, G Niu, S Sabato, pp. 4816–30. Cambridge, MA: PMLR

Dasgupta I, Schulz E, Goodman ND, Gershman SJ. 2018. Remembrance of inferences past: amortization in human hypothesis generation. *Cognition* 178:67–81

Dasgupta I, Schulz E, Tenenbaum JB, Gershman SJ. 2020. A theory of learning to infer. *Psychol. Rev.* 127(3):412–41

Dayan P. 1993. *Improving generalization for temporal difference learning: the successor representation*. Work. Pap., Salk Inst., San Diego, CA

Dehaene S, Al Roumi F, Lakretz Y, Planton S, Sablé-Meyer M. 2022. Symbols and mental programs: a hypothesis about human singularity. *Trends Cogn. Sci.* 26(9):751–66

Deleu T, Góis A, Emezue C, Rankawat M, Lacoste-Julien S, et al. 2022. Bayesian structure learning with generative flow networks. In *Uncertainty in Artificial Intelligence*, pp. 518–28. Cambridge, MA: PMLR

DeLosh EL, Busemeyer JR, McDaniel MA. 1997. Extrapolation: the sine qua non for abstraction in function learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 23(4):968–86

Denison S, Bonawitz E, Gopnik A, Griffiths TL. 2013. Rational variability in children's causal inferences: the sampling hypothesis. *Cognition* 126(2):285–300

Doolittle WF, Bapteste E. 2007. Pattern pluralism and the tree of life hypothesis. *PNAS* 104(7):2043–49

Doucet A, Johansen AM, Others. 2011. A tutorial on particle filtering and smoothing: fifteen years later. In *The Oxford Handbook of Nonlinear Filtering*, ed. D Crisan, B Rozovskii, pp. 656–704. New York: Oxford Univ. Press

Duvenaud D, Lloyd JR, Grosse R, Tenenbaum JB, Ghahramani Z. 2013. Structure discovery in nonparametric regression through compositional kernel search. *PMLR* 28(3):1166–74

Ekman G. 1954. Dimensions of color vision. *J. Psychol.* 38(2):467–74

Ellis K, Wong L, Nye M, Sable-Meyer M, Cary L, et al. 2023. Dreamcoder: growing generalizable, interpretable knowledge with wake–sleep Bayesian program learning. *Philos. Trans. R. Soc. A* 381(2251):20220050

Epstein RA, Patai EZ, Julian JB, Spiers HJ. 2017. The cognitive map in humans: spatial navigation and beyond. *Nat. Neurosci.* 20(11):1504–13

Erickson MA, Kruschke JK. 1998. Rules and exemplars in category learning. *J. Exp. Psychol. Gen.* 127(2):107–40

Fodor JA. 1975. *The Language of Thought.* Cambridge, MA: Harvard Univ. Press

Fränken JP, Theodoropoulos NC, Bramley NR. 2022. Algorithms of adaptation in inductive inference. *Cogn. Psychol.* 137:101506

Garvert MM, Dolan RJ, Behrens TEJ. 2017. A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *eLife* 6:e17086

Garvert MM, Saanum T, Schulz E, Schuck NW, Doeller CF. 2023. Hippocampal spatio-predictive cognitive maps adaptively guide reward generalization. *Nat. Neurosci.* 26(4):615–26

Geirhos R, Medina Temme CR, Rauber J, Schütt HH, Bethge M, Wichmann FA. 2018. Generalisation in humans and deep neural networks. *Adv. Neural Inf. Process. Syst.* 31. **https://papers.nips.cc/paperiles/paper/2018/file/0937fb5864ed06ffb59ae5f9b5ed67a9-Paper.pdf**

Gershman SJ. 2018. Deconstructing the human algorithms for exploration. *Cognition* 173:34–42

Gershman SJ, Daw ND. 2017. Reinforcement learning and episodic memory in humans and animals: an integrative framework. *Annu. Rev. Psychol.* 68:101–28

Gershman SJ, Malmaud J, Tenenbaum JB. 2017. Structured representations of utility in combinatorial domains. *Decision* 4(2):67–86

Giron AP, Ciranka S, Schulz E, van den Bos W, Ruggeri A, et al. 2023. Developmental changes in exploration resemble stochastic optimization. *Nat. Hum. Behav.* 7:1955–67

Goodman N. 1972. Seven strictures on similarity. In *Problems and Projects*. Indianapolis, IN: Bobbs-Merrill

Goodman ND, Tenenbaum JB, Feldman J, Griffiths TL. 2008. A rational analysis of rule-based concept learning. *Cogn. Sci.* 32(1):108–54

Gopnik A. 2020. Childhood as a solution to explore–exploit tensions. *Philos. Trans. R. Soc. B* 375(1803):20190502

Gopnik A, Griffiths TL, Lucas CG. 2015. When younger learners can be better (or at least more open-minded) than older ones. *Curr. Dir. Psychol. Sci.* 24(2):87–92

Gopnik A, O'Grady S, Lucas CG, Griffiths TL, Wente A, et al. 2017. Changes in cognitive flexibility and hypothesis search across human life history from childhood to adolescence to adulthood. *PNAS* 114(30):7892–99
Griffiths TL, Tenenbaum JB. 2005. Structure and strength in causal induction. *Cogn. Psychol.* 51(4):334–84
Griffiths TL, Tenenbaum JB. 2009. Theory-based causal induction. *Psychol. Rev.* 116(4):661–716
Hahn U, Chater N. 1998. Similarity and rules: distinct? Exhaustive? Empirically distinguishable? *Cognition* 65(2–3):197–230
Hahn U, Ramscar M. 2001. *Similarity and Categorization*. Oxford, UK: Oxford Univ. Press
Henrich J. 2016. *The Secret of Our Success: How Culture Is Driving Human Evolution, Domesticating Our Species, and Making Us Smarter*. Princeton, NJ: Princeton Univ. Press
Heyes C. 2018. *Cognitive Gadgets: The Cultural Evolution of Thinking*. Cambridge, MA: Harvard Univ. Press
Ho MK, MacGlashan J, Littman ML, Cushman F. 2017. Social is special: a normative framework for teaching with and learning from evaluative feedback. *Cognition* 167:91–106
Huys QJ, Lally N, Faulkner P, Eshel N, Seifritz E, et al. 2015. Interplay of approximate planning strategies. *PNAS* 112(10):3098–103
Jäkel F, Schölkopf B, Wichmann FA. 2008. Similarity, kernels, and the triangle inequality. *J. Math. Psychol.* 52(5):297–303
James W. 1890. *The Principles of Psychology*. New York: Henry Holt & Co.
Jara-Ettinger J. 2019. Theory of mind as inverse reinforcement learning. *Curr. Opin. Behav. Sci.* 29:105–10
Kalish ML, Griffiths TL, Lewandowsky S. 2007. Iterated learning: Intergenerational knowledge transmission reveals inductive biases. *Psychon. Bull. Rev.* 14(2):288–94
Kalish ML, Lewandowsky S, Kruschke JK. 2004. Population of linear experts: knowledge partitioning and function learning. *Psychol. Rev.* 111(4):1072–99
Kemp C, Tenenbaum JB. 2008. The discovery of structural form. *PNAS* 105(31):10687–92
Kemp C, Tenenbaum JB. 2009. Structured statistical models of inductive reasoning. *Psychol. Rev.* 116(1):20–58
Keramati M, Smittenaar P, Dolan RJ, Dayan P. 2016. Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *PNAS* 113(45):12868–73
Kondor RI, Lafferty J. 2002. Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning*, ed. C Sammut, AG Hoffmann, pp. 3121–322. San Francisco: Morgan Kaufmann
Kool W, Gershman SJ, Cushman FA. 2018. Planning complexity registers as a cost in metacontrol. *J. Cogn. Neurosci.* 30(10):1391–404
Kotov R, Krueger RF, Watson D, Cicero DC, Conway CC, et al. 2021. The hierarchical taxonomy of psychopathology (HiTOP): a quantitative nosology based on consensus of evidence. *Annu. Rev. Clin. Psychol.* 17:83–108
Kruschke JK. 1992. ALCOVE: An exemplar-based connectionist model of category learning. *Psychol. Rev.* 99(1):22–44
Kruskal JB. 1964. Nonmetric multidimensional scaling: a numerical method. *Psychometrika* 29(2):115–29
Lake BM, Salakhutdinov R, Tenenbaum JB. 2015. Human-level concept learning through probabilistic program induction. *Science* 350(6266):1332–38
Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. 2017. Building machines that learn and think like people. *Behav. Brain Sci.* 40:e253
Lau T, Gershman SJ, Cikara M. 2020. Social structure learning in human anterior insula. *eLife* 9:e53162
Lengyel M, Dayan P. 2007. Hippocampal contributions to control: the third way. *Adv. Neural Inform. Process. Syst.* 20:889–96
Love BC, Medin DL, Gureckis TM. 2004. SUSTAIN: a network model of category learning. *Psychol. Rev.* 111(2):309–32
Lucas CG, Bridgers S, Griffiths TL, Gopnik A. 2014. When children are better (or at least more open-minded) learners than adults: developmental differences in learning the forms of causal relationships. *Cognition* 131(2):284–99
Lucas CG, Griffiths TL, Williams JJ, Kalish ML. 2015. A rational model of function learning. *Psychon. Bull. Rev.* 22(5):1193–215

Lynn CW, Bassett DS. 2020. How humans learn and represent networks. *PNAS* 117(47):29407–15
Machado MC, Rosenbaum C, Guo X, Liu M, Tesauro G, Campbell M. 2018. Eigenoption discovery through the deep successor representation. arXiv:1710.11089 [cs.LG]
Mark S, Moran R, Parr T, Kennerley SW, Behrens TEJ. 2020. Transferring structural knowledge across cognitive maps in humans and models. *Nat. Commun.* 11(1):4783
McClelland JL, Rumelhart DE, PDP Res. Group. 1986. *Parallel Distributed Processing*, Vol. 2. Cambridge, MA: MIT Press
McDaniel MA, Busemeyer JR. 2005. The conceptual basis of function learning and extrapolation: comparison of rule-based and associative-based models. *Psychon. Bull. Rev.* 12(1):24–42
Meder B, Mayrhofer R, Waldmann MR. 2014. Structure induction in diagnostic causal reasoning. *Psychol. Rev.* 121(3):277–301
Meder B, Wu CM, Schulz E, Ruggeri A. 2021. Development of directed and random exploration in children. *Dev. Sci.* 24(4):e13095
Medin DL, Goldstone RL, Gentner D. 1993. Respects for similarity. *Psychol. Rev.* 100(2):254–78
Medin DL, Schaffer MM. 1978. Context theory of classification learning. *Psychol. Rev.* 85(3):207–38
Mehlhorn K, Newell BR, Todd PM, Lee MD, Morgan K, et al. 2015. Unpacking the exploration–exploitation tradeoff: a synthesis of human and animal literatures. *Decisions* 2(3):191–215
Mercer J. 1909. Functions of positive and negative type and their connection with the theory of integral equations. *Philos. Trans. R. Soc. A* 209:441–58
Miller KJ, Botvinick MM, Brody CD. 2017. Dorsal hippocampus contributes to model-based planning. *Nat. Neurosci.* 20(9):1269–76
Moser EI, Roudi Y, Witter MP, Kentros C, Bonhoeffer T, Moser MB. 2014. Grid cells and cortical representation. *Nat. Rev. Neurosci.* 15(7):466–81
Moskvichev A, Odouard VV, Mitchell M. 2023. The ConceptARC benchmark: evaluating understanding and generalization in the arc domain. arXiv:2305.07141 [cs.LG]
Murphy GL, Medin DL. 1985. The role of theories in conceptual coherence. *Psychol. Rev.* 92(3):289–316
Neal RM. 1996. Priors for infinite networks. In *Bayesian Learning for Neural Networks*, pp. 29–53. New York: Springer
Nelson JD. 2005. Finding useful questions: on Bayesian diagnosticity, probability, impact, and information gain. *Psychol. Rev.* 112(4):979–99
Newell A, Simon HA. 1976. Computer science as empirical inquiry: symbols and search. *Commun. ACM* 19(3):113–26
Norbury A, Robbins TW, Seymour B. 2018. Value generalization in human avoidance learning. *eLife* 7:e34779
Nosofsky RM. 1986. Attention, similarity, and the identification–categorization relationship. *J. Exp. Psychol. Gen.* 115(1):39–57
Nosofsky RM, Palmeri TJ, McKinley SC. 1994. Rule-plus-exception model of classification learning. *Psychol. Rev.* 101(1):53–79
Pavlov IP. 1927. *Conditional Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex*. Oxford, UK: Oxford Univ. Press
Peer M, Brunec IK, Newcombe NS, Epstein RA. 2021. Structuring knowledge with cognitive maps and cognitive graphs. *Trends Cogn. Sci.* 25(1):37–54
Pettine WW, Raman DV, Redish AD, Murray JD. 2023. Human generalization of internal representations through prototype learning with goal-directed attention. *Nat. Hum. Behav.* 7(3):442–63
Piantadosi ST, Tenenbaum JB, Goodman ND. 2016. The logical primitives of thought: empirical foundations for compositional cognitive models. *Psychol. Rev.* 123(4):392–424
Poggio T, Bizzi E. 2004. Generalization in vision and motor control. *Nature* 431(7010):768–74
Pothos EM. 2005. The rules versus similarity distinction. *Behav. Brain Sci.* 28(1):1–14; discuss. 14–49
Pouncy T, Tsividis P, Gershman SJ. 2021. What is the model in model-based planning? *Cogn. Sci.* 45(1):e12928
Radulescu A, Shin YS, Niv Y. 2021. Human representation learning. *Annu. Rev. Neurosci.* 44:253–73
Rasmussen CE, Williams CKI. 2005. *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press
Reber AS, Lewis S. 1977. Implicit learning: an analysis of the form and structure of a body of tacit knowledge. *Cognition* 5(4):333–61

Rosch EH. 1973. Natural categories. *Cogn. Psychol.* 4(3):328–50
Rouder JN, Ratcliff R. 2006. Comparing exemplar- and rule-based theories of categorization. *Curr. Dir. Psychol. Sci.* 15(1):9–13
Rubino V, Hamidi M, Dayan P, Wu CM. 2023. Compositionality under time pressure. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, ed. M Goldwater, F Anggoro, B Hayes, D Ong, pp. 678–85. Seattle, WA: Cogn. Sci. Soc.
Rule JS, Tenenbaum JB, Piantadosi ST. 2020. The child as hacker. *Trends Cogn. Sci.* 24(11):900–15
Russek EM, Momennejad I, Botvinick MM, Gershman SJ, Daw ND. 2017. Predictive representations can link model-based reinforcement learning to model-free mechanisms. *PLOS Comput. Biol.* 13(9):e1005768
Sanborn AN, Griffiths TL, Navarro DJ. 2010. Rational approximations to rational models: alternative algorithms for category learning. *Psychol. Rev.* 117(4):1144–67
Schölkopf B, Smola AJ. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* Cambridge, MA: MIT Press
Schulz E, Speekenbrink M, Krause A. 2018. A tutorial on Gaussian process regression: modelling, exploring, and exploiting functions. *J. Math. Psychol.* 85:1–16
Schulz E, Tenenbaum JB, Duvenaud D, Speekenbrink M, Gershman SJ. 2017. Compositional inductive biases in function learning. *Cogn. Psychol.* 99:44–79
Schulz E, Wu CM, Ruggeri A, Meder B. 2019. Searching for rewards like a child means less generalization and more directed exploration. *Psychol. Sci.* 30(11):1561–72
Shepard RN. 1962. The analysis of proximities: multidimensional scaling with an unknown distance function II. *Psychometrika* 27(3):219–46
Shepard RN. 1987. Toward a universal law of generalization for psychological science. *Science* 237(4820):1317–23
Silver D, Huang A, Maddison CJ, Guez A, Sifre L, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature* 529(7587):484–89
Skinner BF. 1938. *The Behavior of Organisms: An Experimental Analysis*. New York: Appleton-Century
Smith EE, Medin DL. 1981. *Categories and Concepts*. Cambridge, MA: Harvard Univ. Press
Smith JD, Minda JP. 1998. Prototypes in the mist: the early epochs of category learning. *J. Exp. Psychol. Learn. Mem. Cogn.* 24(6):1411–36
Somerville LH, Sasse SF, Garrad MC, Drysdale AT, Abi Akar N, et al. 2017. Charting the expansion of strategic exploratory behavior during adolescence. *J. Exp. Psychol. Gen.* 146(2):155–64
Speekenbrink M. 2016. A tutorial on particle filters. *J. Math. Psychol.* 73:140–52
Stachenfeld KL, Botvinick MM, Gershman SJ. 2017. The hippocampus as a predictive map. *Nat. Neurosci.* 20(11):1643–53
Stojić H, Schulz E, Analytis P, Speekenbrink M. 2020. It's new, but is it good? How generalization and uncertainty guide the exploration of novel options. *J. Exp. Psychol. Gen.* 149(10):1878–907
Sutton RS, Barto AG. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA: MIT Press. 2nd ed.
Tavares RM, Mendelsohn A, Grossman Y, Williams CH, Shapiro M, et al. 2015. A map for social navigation in the human brain. *Neuron* 87(1):231–43
Taylor JE, Cortese A, Barron HC, Pan X, Sakagami M, Zeithamova D. 2021. How do we generalize? *Neurons Behav. Data Anal. Theory* 1. **https://doi.org/10.51628/001c.27687**
Tenenbaum JB, Griffiths TL. 2001. Generalization, similarity, and Bayesian inference. *Behav. Brain Sci.* 24(4):629–40; discuss. 652–791
Tesauro G. 1995. Temporal difference learning and TD-Gammon. *Commun. ACM* 38(3):58–68
Thorndike EL. 1911. *Animal Intelligence: Experimental Studies.* New York: Macmillan Co.
Tolman EC. 1948. Cognitive maps in rats and men. *Psychol. Rev.* 55(4):189–208
Torgerson WS. 1952. Multidimensional scaling: I. Theory and method. *Psychometrika* 17(4):401–19
Tsividis PA, Loula J, Burga J, Foss N, Campero A, et al. 2021. Human-Level reinforcement learning through theory-based modeling, exploration, and planning. arXiv:2107.12544 [cs.AI]
Tversky A. 1977. Features of similarity. *Psychol. Rev.* 84(4):327–52
Van den Bos W, Cohen MX, Kahnt T, Crone EA. 2012. Striatum–medial prefrontal cortex connectivity predicts developmental changes in reinforcement learning. *Cereb. Cortex* 22(6):1247–55

Whittington JC, McCaffary D, Bakermans JJ, Behrens TE. 2022. How to build a cognitive map. *Nat. Neurosci.* 25(10):1257–72

Whittington JCR, Muller TH, Mark S, Chen G, Barry C, et al. 2020. The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. *Cell* 183(5):1249–63.e23

Wilson RC, Geana A, White JM, Ludvig EA, Cohen JD. 2014. Humans use directed and random exploration to solve the explore-exploit dilemma. *J. Exp. Psychol. Gen.* 143(6):2074–81

Wimmer GE, Elliott Wimmer G, Daw ND, Shohamy D. 2012. Generalization of value in reinforcement learning by humans. *J. Neurosci.* 35(7):1092–104

Wingate D, Diuk C, O'Donnell T, Tenenbaum J, Gershman S. 2013. *Compositional policy priors.* Tech. Rep. MIT-CSAIL-TR-2013-007, Mass. Inst. Technol., Cambridge

Wise T, Emery K, Radulescu A. 2024. Naturalistic reinforcement learning. *Trends Cogn. Sci.* 28(2):144–58

Witt A, Toyokawa W, Lala KN, Gaissmaier W, Wu CM. 2024. Humans flexibly integrate social information despite interindividual differences in reward. *PNAS* 121(39):e2404928121

Wu CM, Dale R, Hawkins RD. 2024. Group coordination catalyzes individual and cultural intelligence. *Open Mind* 8:1037–57

Wu CM, Deffner D, Kahl B, Meder B, Ho MH, Kurvers RH. 2023. Visual-spatial dynamics drive adaptive social learning in immersive environments. bioRxiv 2023.06.28.546887

Wu CM, Schulz E, Garvert MM, Meder B, Schuck NW. 2020. Similarities and differences in spatial and non-spatial cognitive maps. *PLOS Comput. Biol.* 16(9):e1008149

Wu CM, Schulz E, Gershman SJ. 2021. Inference and search on graph-structured spaces. *Comput. Brain Behav.* 4(2):125–47

Wu CM, Schulz E, Pleskac TJ, Speekenbrink M. 2022a. Time pressure changes how people explore and respond to uncertainty. *Sci. Rep.* 12:4122

Wu CM, Schulz E, Speekenbrink M, Nelson JD, Meder B. 2018. Generalization guides human exploration in vast decision spaces. *Nat. Hum. Behav.* 2(12):915–24

Wu CM, Vélez N, Cushman FA. 2022b. Representational exchange in human social learning: balancing efficiency and flexibility. In *The Drive for Knowledge: The Science of Human Information-Seeking*, ed. IC Dezza, E Schulz, CM Wu, pp. 169–92. Cambridge, UK: Cambridge Univ. Press

Xu F, Tenenbaum JB. 2007. Word learning as Bayesian inference. *Psychol. Rev.* 114(2):245–72

Zajkowski WK, Kossut M, Wilson RC. 2017. A causal role for right frontopolar cortex in directed, but not random, exploration. *eLife* 6:e27430

Zhang C, Bengio S, Hardt M, Recht B, Vinyals O. 2016. Understanding deep learning requires rethinking generalization. *Commun. ACM* 64(3):107–15

Zhao B, Lucas CG, Bramley NR. 2024. A model of conceptual bootstrapping in human cognition. *Nat. Hum. Behav.* 8:125–36

Zhou H, Bamler R, Wu CM, Tejero-Cantero A. 2024a. Predictive, scalable and interpretable knowledge tracing on structured domains. arXiv:2403.13179 [cs.LG]

Zhou H, Nagy DG, Wu CM. 2024b. Harmonizing program induction with rate-distortion theory. In *Proceedings of the 46th Annual Conference of the Cognitive Science Society*, pp. 2511–18. Seattle, WA: Cogn. Sci. Soc.

Zhu X, Lafferty J, Ghahramani Z. 2003. *Semi-supervised learning: from Gaussian fields to Gaussian processes.* Tech. Rep. CMU-CS-03-175, Carnegie Mellon Univ., Pittsburgh, PA