Learning to Discover Structure in Animal and Human Decision Tasks

Mingyu Song

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Neuroscience
Adviser: Yael Niv

January 2022

Abstract

Learning in real life is never as simple as forming stimulus-response mappings. It involves identifying the current context (i.e., relevant information for task at hand), figuring out the transition between contexts, and learning about complex relationships and rules. In this dissertation, I study how animals and humans learn to discover such structures in decision tasks. I begin by demonstrating the importance of studying structure or representation learning. In Chapter 2, I show that rats do not form the optimal task representation in an odor-guided decision task, even after extensive training. This suggests that we cannot assume a task representation without testing it. It also raises the following questions: How is a task representation learned? What factors may affect such learning? In the rest of this dissertation, I use two tasks to study these questions with animals and humans. In Chapter 3, I propose a latent-cause inference model to explain fear extinction in rats. This model characterizes how animals make inference about the underlying causes that generate observations (e.g., shocks) and how the causes may change over time. It explains why gradually reducing shock frequency is more effective in extinguishing fear than the standard extinction procedure, by demonstrating how different procedures lead to the learning of distinct underlying task structures. In Chapter 4, I study how humans actively learn about multi-dimensional rules with probabilistic feedback. I show that people use both value-based and rule-based learning systems, and trade off them based on the instructed task complexity. This study sheds light on how humans make strategic use of cognitive resource when learning complex task structures. In Chapter 5, I propose a novel approach to study representation learning with recurrent neural networks (RNNs). I demonstrate that RNNs can be useful for developing better cognitive models and identifying cognitive differences across individuals. In the Conclusion, I summarize the findings from the above studies, and discuss common principles that underlie animal and human representation learning.

Acknowledgements

Looking back on my journey through graduate school, I am incredibly grateful to have the most amazing and supportive mentors. I would like to first thank Wei Ji Ma, who took me (then a clueless physics undergraduate) in as a research assistant, and guided me into the fascinating field of computational cognitive science. I am deeply indebted to Weiji, for all the technical skills he taught me, from designing and running experiments, to analyzing data and fitting computational models, which were pretty much all I needed for graduate school ;) More importantly, he showed me how much fun it is to do science. One of the most unforgettable memories in Ma lab was Weiji asking me: do you feel happy doing research? (and my answer was absolutely yes, and I am so glad it has not changed since!)

I would like to thank my advisor Yael Niv, who is unarguably the most supportive and caring mentor. She gave me full freedom in exploring my research interests, as well as my personal development goals, and cares deeply down for me both professionally and personally. I cannot count how many times in the past five years I felt my life was lit up after a conversation with her, or receiving an email from her. Her brightest smile was what carried me through the inevitable dark days in graduate school. I learned so much from Yael (and will continue learning more) as a brilliant female scientist, mentor, mother and activist, and I am so grateful to have her as a role model to look up to.

I would like to thank my unofficial mentors in Niv lab: Ming Bo Cai and Angela Langdon. I kept learning from them and getting inspired by the intellectually-stimulating discussions we had; these collaborations were what made research life lively and exciting, and what kept me going, during Yael's sabbatical year and the covid era. I would also like to thank my thesis committee members: Nathaniel Daw and Tom Griffiths, who provided helpful feedback on my research projects and writing of this dissertation.

I was fortunately to have amazing lab mates and colleagues who I enjoyed discussing science with and learning from: Angela Redulescu, Sam Zorowitz, Lili Cai, Qihong Lu, Val Felso, Kevin Miller, and Sashank Pisupati. One thing I will miss a lot about academia is the serendipity of meeting new people with shared interests and clicking right away over a few conversations: I remember fondly the summer at Dartmouth in MIND summer school and the wizards squad: Julie Lee, Kate Nussenbaum, Pete Hitchhock, and Sev Harootonian, as well as friends I met over conferences and continued to connect later and beyond: Dalin Guo, Qiong Zhang, and Jimmy Xia.

I am incredibly lucky to spend the past five years with my dear friends. First shout out to friends in Ma lab: Bas van Opheusden, Maija Honig, Andra Mihali, Zahy Bnaya, Zhiwei Li, and Peiyuan Zhang. You made me feel at home when I first came to the U.S., and every time I see you later I always feel the same warmth from deep down. I had the most amazing time, especially the first year of graduate school and every year before PNI retreats, with my cohort: Norbert Cruz, Rolando Masis-Obando, Nivedita Rangarajan, Ellia Miller, Anna Zhukovskaya, and Mike Morais. We made the best skits together, and the days and nights we rehearsed and filmed were the most joyful memory I had in PNI. Life at Princeton was made so much more colorful with my lovely housemates: Xiaofang Yang, Anran Li, Xiaoxuan Li, and Jin Du. Thank you for all the joy and laughter over the hot pots, BBQs, game nights, and outdoor adventures.

Last and the most importantly, my family. Nothing would have been possible without the love and support from my parents. Over the years, you have full-heartedly supported every academic and career decision I made for myself. However far I go, you are always my roots and the place I can return to and count on. My then boyfriend, now husband, Minghao Qiu, you mean so much more to me than the titles. The countless flight, train and bus rides from Princeton to Boston had been what I looked forward to the most in the first four years of graduate school. I thank the pandemic

for bringing us physically together, and I treasure every moment we spent together in our little apartment in Plainsboro.

To 大北京 and Charles River.

Contents

Chapter 1

Introduction

Learning to acquire reward and avoid punishment is a ubiquitous task for animals and human beings. There have been extensive empirical and theoretical studies of this learning process: early experimentalists characterized it as acquiring the association between a stimulus (or action) and positive or negative outcomes [1, 2]; later theoretical work formulated learning as driven by error signals comparing expectation and outcomes [3], with robust neural underpinnings in the midbrain dopaminergic system [4, 5].

1.1 Introducing "state" and state representation

In real life, however, learning is rarely as simple as evaluating how good a stimulus or action is in an absolute sense. Context is critical in learning: the outcome of actions often depends on the context, and actions may in turn change the context. In reinforcement learning theory, context is often termed a "state"; it comprises of everything in the environment that is relevant to the agent's current decision. For example, when foraging for a tasty croissant at Little Chef Pastry Shop in downtown Princeton, I would go on different routes if I leave from home (south of Princeton town), or if I just finished my morning run in the woods up north. In this foraging task, the task "state" would consist of my current location. It determines what action (moving direction) I should take, and the actions subsequently transition me into new states (locations). State information can be local (my current location) or more global (I am on my way to get a croissant), with some aspects perceptually available in the immediate surrounds, while others are more circumstantial or internal (e.g., my current goal, or the time of day – croissants are often sold out before noon, so I should avoid going later in the day). The state in reinforcement learning should ideally include all such information, as long as it is relevant for decision making.

All task states (and the transitions between them, if available) together form a "state representation". State representations serve as the foundation for learning what stimuli to approach or avoid and what actions to take. Given a state representation, we can use reinforcement learning theory [6] (e.g., Bellman Equations [7] and dynamic programming for model-based reinforcement learning, or trial-and-error learning from prediction errors for model-free reinforcement learning [8]) to derive the decision policy and make predictions about the agents' behavior. In the croissant-foraging example, this equates to having a map of the Princeton area, and figuring out the path from my starting point to the pastry shop.

It has been widely shown that animals and humans can solve decision tasks by utilizing the knowledge of states and state representations. They are able to identify the relevant information for task at hand: not only the predictors for reward, but also what information is irrelevant and what scenarios can be treated as the same despite perceptual differences [9, 10]. In tasks with multiple states and transitions, animals and humans are able to acquire complex task representations. For example, it was first proposed by Tolman [11] that animals form "cognitive maps" (mental representation of the physical maps) of their environments in navigation tasks. This idea was later supported by neural findings of place cells in hippocampus [12] and grid cells in the entorhinal cortex, providing the neural basis for mental representation. Similarly, humans have been shown to use knowledge of task structures in decision tasks that have complex transition structures [13] or require multi-step planning [14]. Neurally, task representation has been found to be encoded in the entorhinal cortex and orbitalfrontal cortex in both rodents and humans [15, 16, 13].

1.2 Representation learning: what is it and important questions to study

Despite the rich findings on the use of state representations in decision tasks, it is less known how these representations are learned by animals and humans, which we term "representation learning".

Although spatial navigation is a good example of using task representations, representation learning is not limited to forming cognitive maps – it is useful in a wide range of decision scenarios. In this dissertation, I primarily focus on the learning of states. The definition of states can sometimes be trivial, especially in tasks where each state is uniquely signalled by a specific stimulus, or a particular coordinate on a map. However, real life decision scenarios are often more complicated. In some cases, states are unobservable, and the agent needs to figure out how to group individual experiences into clusters (hidden states) to guide decisions. In other cases, there is abundant (and potentially redundant) information, and the agent needs to identify the relevant subset of factors that define the states, which can be hard due to the combinatorial explosion of the factors.

So far, the study of representation learning has often focused on individual learning scenarios. The question remains whether there exist common mechanisms for the underlying cognitive and neural processes. This is not only important for cognitive science and neuroscience, but also relevant for artificial intelligence, which currently excels at solving individual complex tasks [17, 18] but suffers from lack of generalizability (but see [19, 20]). Similarly, it is worth developing general computational approaches to studying representation learning.

1.3 Aims of dissertation and overview of chapters

The aims of this dissertation are three-fold: (a) to demonstrate the importance of studying representation learning; (b) to make progress in understanding the computational mechanisms that underlie representation learning processes; (c) to propose useful computational approaches for representation-learning studies.

In Chapter 2, I will begin by demonstrating the importance of studying representation learning. Research investigating learning and decision making often assumes that animals and humans use the correct state representation for a given task, i.e., the representation that accords with the true generative model of the task or environment. However, this assumption may not be true. For example, some roads on the Princeton campus do not allow non-university vehicles to drive through, yet a newcomer to Princeton may not have such knowledge and thus may form an incorrect map for croissant-foraging. It is therefore important to examine the actual task representation used by the decision maker. In this chapter, I study how rats perform and represent a seemingly-simple odor-guided choice task. By comparing the predictions of several reinforcement-learning models with various state representations to animals' behavioral data, I show that rats do not use the most parsimonious task representation as designed by the experimenters. I also examine how their representations change over time and show that representation learning is a very slow process in this task, potentially explaining animals' sub-optimal representation even after extensive training. These findings demonstrate the importance of carefully examining the state representation held by experimental subjects, as well as the value of studying how representation may change over time through experience.

Next, I study how animals and humans acquire state representations. Specifically, I study how they learn to group individual experiences into latent states (Chapters 3), and how they learn to identify relevant information for current task (Chapter 4).

One way to think about a state is that it is a grouping of similar (but slightly different) experiences so that they can be treated similarly in decision making (i.e., the same policy can be used in all of those situations) and learning from one situation can be applied to the other. For instance, when deciding where to have lunch on campus, after seeing certain cafes always having the same daily menus, you may be able to infer that they share the same caterer, which can help reduce the number of options and simplify decisions. These states are unobservable, and it is up to the learner to figure out how to group individual experiences into hidden states that are useful for guiding decisions. In Chapter 3, I study the learning of latent states in fear extinction experiments with rats. Gershman and colleagues [21] showed that extinction of previously acquired fear was more effective with a gradual extinction procedure, where the frequency of an aversive stimulus (there, shock) was reduced gradually over time, compared to the standard extinction procedure where no shock was delivered, or a gradual reverse procedure where shock frequency increased instead. These phenomena can be explained by a latent-cause inference theory in which animals form a rich model of the environment, inferring what underlying causes generate distinct experiences with shocks, and treating these causes as states. Building on such theory, I further demonstrate with a computational model that animals' inference process relies on their understanding of a dynamic environment: old causes become less likely to reoccur later on, and the tendency for a shock to appear may change over time. Additionally, I show that animals consider multiple possibilities during their inference on latent causes, but only keep the most likely one after long delays. This work provides a quantitative account on how animals learn latent states through inferring the unobservable structure of a task.

When there is redundant information in a decision task, the learner needs to figure out the relevant subset of factors for decision. For example, learning to pick the best coffee beans requires identifying which factors may determine its flavor: brand,

package design, country of origin, level of roast, etc. In Chapter 4, I study such learning in humans with a multi-dimensional probabilistic learning task where the underlying rule for reward depends on an unknown subset of dimensions. I show that people combine rule-based and value-based strategies in learning: they serially test hypotheses about the underlying rule, and simultaneously learn the value of individual factors that comprise rules, which helps form better hypotheses for later testing. The integration of the two learning strategies is sensitive to task complexity: when rules are known to be low-dimensional (and hence simpler), people do more hypothesis-testing; whereas when rules are more complex (high-dimensional), learning values is more efficient than sequentially testing many possibilities, and people indeed rely more on value learning. These results suggest that humans can sensibly choose between representation learning strategies based on their costs and benefits.

Then, in Chapter 5, I propose a novel computational approach for studying representation learning using recurrent neural networks (RNNs). Compared to simpler reward-learning tasks (e.g., multi-armed bandit tasks) where the cognitive mechanisms are relatively well-understood and characterized by existing computational models (e.g., reinforcement learning models), representation-learning tasks are often more complex, with little consensus on the underlying cognitive processes, making it hard to evaluate how good a model is in capturing learning behavior. Thus, it can be useful to apply RNNs (which are flexible general function approximators, although less interpretable) to fit behavior: RNNs can set targets for developing cognitive models, and help identify room for improvement. With network embedding analyses, RNNs can also reveal potential cognitive variability across individuals. In this chapter, I apply RNNs to the rule-learning task reported in Chapter 4, use them to predict participants' choice behavior, and demonstrate the utility of this approach.

Lastly, in Chapter 6 (Conclusion), I summarize all the findings and their implications, and lay out clearly the contributions of this dissertation. I end with a discussion

of common cognitive principles underlying animal and human representation learning, revealed by the above studies, as well as the computational approaches that are broadly useful in this field of research.

Chapter 2

Learners may not acquire the optimal task representation

The contents of this chapter were submitted for publication in: Mingyu Song, Yuji K. Takahashi, Amanda C. Burton, Matthew R. Roesch, Geoffrey Schoenbaum, Yael Niv, and Angela J. Langdon. Minimal cross-trial generalization in learning the representation of an odor-guided choice task.

All data and code are available at `https://github.com/mingyus/`.

2.1 Introduction

Much knowledge of the world is acquired not from instructions, but through observations and inference. For example, you might choose which campus cafeteria to visit by checking their daily menus. Eventually, you may realize that cafeterias A and B have the same menu (unbeknownst to you, they are run by the same caterer). This implicit knowledge allows you to apply whatever you learn about one dining location to the other: upon hearing that cafeteria A is serving your favorite dish, you can get it at the close-by cafeteria B.

Acquiring such knowledge can be considered as learning the structure of a task, or in reinforcement learning (RL) terminology, learning a state representation for the task [22, 23, 24]. A state representation forms the basis upon which values (expectations about future rewards) and policies (rules for action in different settings) can be learned [25]. In tasks in which different settings (e.g., cafeteria A or B) lead to the same outcome (the same dishes on the menu), the state representation for A and B can be *shared*. The benefit of this is two-fold: first, it allows compression of state representations, excluding irrelevant variation and thus reducing the complexity of the learning problem. Second, it accelerates learning as only a single experience of a tasty salad in cafeteria A is required in order to exploit that knowledge in both locations. While a range of alternative state representations can support learning in any given setting, one that matches the "true" underlying structure of a task supports efficient learning and accurate task performance.

The challenge for a learner to build an appropriate state representation is particularly acute when there is no explicit instruction on the "rules" for solving a task. This occurs by necessity in experiments on non-human animals, in which subjects are trained solely through ongoing experience. Experimenters know the ground-truth structure of a task, and often assume the subjects understand it similarly. However, even relatively simple tasks may be represented in a multitude of ways, often with

only subtle differences in overt behavioral performance. Despite rapid progress in the development of artificial learning algorithms that can extract appropriately abstract task representations from reinforcement [26, 27, 28, 29], it remains unknown how animals form a state representation solely through their experience of stimuli, rewards and the contingency of each of these on their actions. In particular, it is an open question how animals might generalize their learning about upcoming rewards across distinct features of experience, thereby building a concise state representation of a task.

2.2 Results

We directly tested the extent of generalization in learned state representations that guide choice behavior in an odor-guided decision-making task in rats [30]. Rats were trained to sample an odor at a central odor port, before responding at one of two fluid wells (Figure 2.1A). The odor stimulus provided a cue for which of two wells would be baited with a sucrose reward. Two odors signalled "forced choice" trials, one indicating reward will be available in the left well, and one indicating the right well. In either case, choosing the unrewarded well terminated the trial immediately. A third odor—"free choice"—indicated reward will be available in either well. Importantly, if a "valid" well were chosen on any trial (i.e., the rewarded well on forced-choice trials, or either well on free-choice trials), the delay to and amount of reward was determined by the side of the well, not the odor. Unsignaled to the animal, in each block of the task, one well delivered a "better" reward outcome, either at a shorter delay or a larger amount than the other well; reward contingencies changed between blocks during a session (see Figure 2.1B for details).

Because of the shared reward setting across odors, it would be beneficial for the animal to acquire a representation of the task in which learning from valid forced-choice

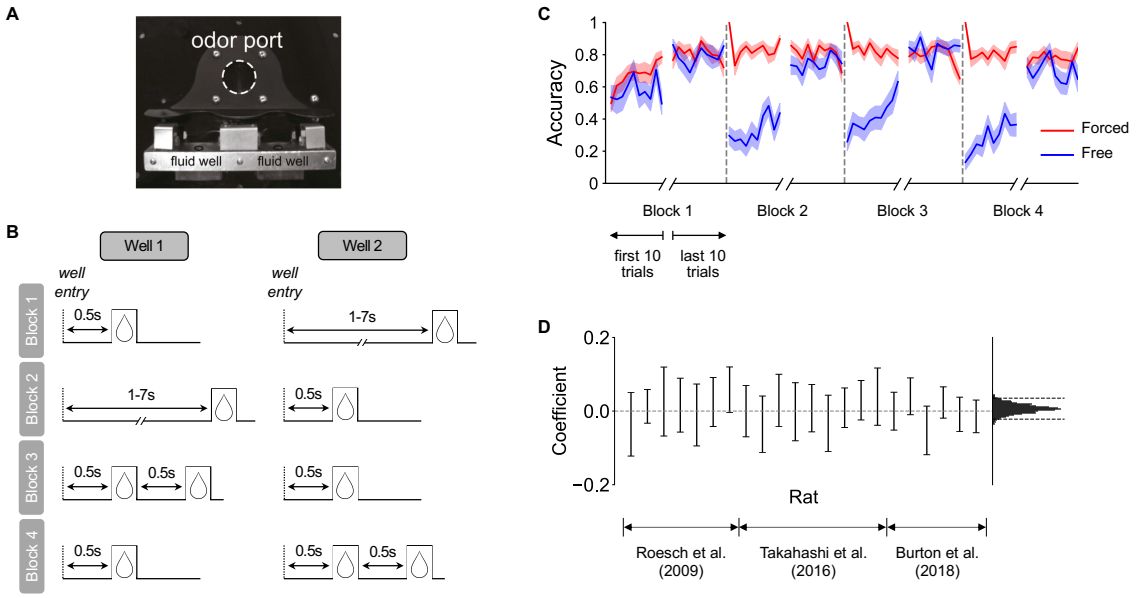

Figure 2.1: **The odor-guided choice task and animals' behavior.** **(A)** The experiment apparatus included an odor port and two fluid wells where rewards were delivered. To start each trial, the animal first poked into the odor port; after 0.5 seconds, one of three odors was delivered, signaling the current trial type. One odor signaled a left forced-choice trial, another a right forced-choice trial, and a third indicated free choice between left and right wells. After odor offset, the animal could make a choice by entering either the left or right fluid wells. Reward was delivered if they made the correct choice on a forced-choice trial and as long as they successfully made one of the choices on a free-choice trial. **(B)** Block sequence in an example session. Sessions always started with two "delay blocks" (blocks 1 and 2), followed by two "magnitude blocks" (blocks 3 and 4). In block 1, the "short" reward (delivered 0.5s after well entry) was available in one well and the "long" reward (delivered 1-7s after well entry) in the other; the reward contingency switched between the wells on block 2. In block 3, "long" reward then changed to "big" reward (two sucrose drops delivered 0.5s after well entry), while "short" reward stayed the same but is now referred to as "small" reward (one sucrose drop) in comparison to the alternative; these reward contingencies were switched again on block 4. The well that was initiated with the better (short) reward option was randomized across sessions. **(C)** Learning curves for forced-choice (red) and free-choice (blue) trials. The curves are aligned to block-switch points (gray dashed lines), with the first and last 10 trials of each block shown. Accuracy is evaluated as the percentage of trials the animal chose the better option for that trial type (forced-choice trials: the rewarded well; free-choice trials: the well with reward at shorter delay or larger amount). Shaded areas represent 1 s.e.m across animals ($N = 22$). **(D)** Coefficients of a hierarchical logistic regression predicting the accuracy of the first free-choice trial after a previous incorrect free-choice as a function of the number of intervening correct forced-choice trials. Left (error bars): coefficients for individual animals, ordered by dataset, with error bars representing 95% highest posterior density interval (HDI). Right (histogram): the posterior distribution of the group mean with dashed lines representing 95% HDI. At both individual and group levels, 95% HDI of the coefficients overlapped with zero, suggesting that there was minimal generalization of learning from correct forced-choice trials to subsequent free-choice trials.

trials generalizes to the same well location in free-choice trials. This representation aligns with the underlying generative structure of the task, and would support faster learning when reward contingencies change between blocks.

To study how rats interpret the structure of this odor-guided choice task, we collated behavior from several experiments using the same behavioral paradigm [31, 32, 33]. On average, across sessions, rats learned to choose the well with the better reward on free-choice trials within each block, while maintaining high choice accuracy on forced-choice trials throughout the session (Figure 2.1C). To determine whether rats learned to choose the better option on free-choice trials by generalizing from rewards delivered on forced-choice trials (and not from experience in free-choice trials alone), we first performed a behavioral analysis. If animals had knowledge of the shared reward setting, reward outcomes in valid forced-choice trials should provide information about the reward available in the two wells, and as a result, improve performance on subsequent free-choice trials. We therefore conducted a hierarchical logistic regression predicting the accuracy of free-choice trials as a function of how many rewarded forced-choice trials the animal had experienced since the last *incorrect* free-choice trial (i.e., the last time they chose the worse well). A positive coefficient would indicate use of forced-choice experience to inform free-choice decisions. We did not find evidence for such signature of generalization (Figure 2.1D), either at the group level (95% highest posterior density interval (HDI) of the group mean of the coefficient: [-0.022, 0.035]), or for individual animals (95% HDI of individual coefficients all include zero). Adding trial index in the block as an additional regressor (to account for the increase in accuracy over each block) did not change the results.

To examine the extent of generalization more directly, we constructed a series of reinforcement learning (RL) models with different state representations of the task (Figure 2.2), and tested how well these alternative models could predict the trial-by-trial choice behavior for each animal. In all models, animals learned the values of

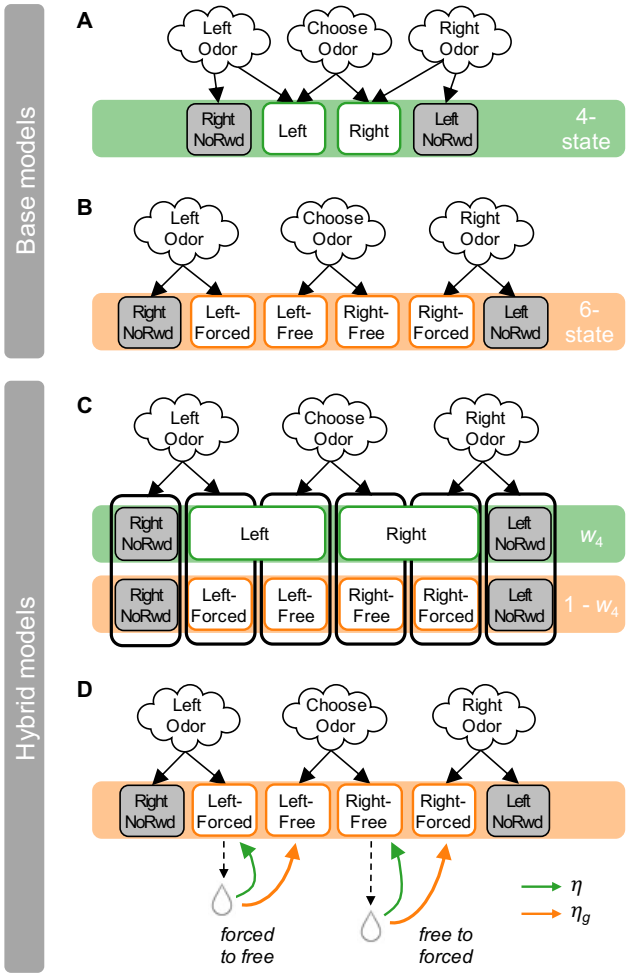

Figure 2.2: **State representations of RL models. (A) The four-state model**: free-choice trials and correct forced-choice trials share the same "Left" and "Right" states; "Right-NoRwd" and "Left-NoRwd" are the corresponding states for incorrect forced-choice trials. This is the true structure of the task as designed by the experimenters, as the same reward was available in forced-choice trials and free-choice trials if a correct choice was made. **(B) The six-state model**: each of the three odors leads to one of two states for left and right choices, with no generalization across odors. **(C) The hybrid-value model**: this model uses both the four-state and six-state representations (with a total of 10 states), with state values combined using weights w_4 and $(1-w_4)$ (illustrated as vertical boxes). **(D) The hybrid-learning model**: the same state representation and learning rule (green arrows, with learning rate η) as in the six-state model, with additional generalization (orange arrows, with generalization rate η_g) between states representing valid forced choices and free choices. For simplicity, shown here only half of the learning and generalization updates (when reward is delivered in Left-Forced and Right-Free states), each representing generalization from forced-choice states to free-choice states or vice versa; the same rules apply to Right-Forced and Left-Free states. Boxes in white and gray represent rewarded and unrewarded states, respectively.

left and right actions for each odor through trial and error, choosing actions by comparing their values, with some decision noise (see Supplementary Methods). What differed between the models was the assumed state representation, i.e., whether and how learning generalized across odors. The *four-state model* assumed full generalization between valid responses on forced-choice trials and corresponding responses on free-choice trials, with shared states between them; this model correctly reflects the generative structure of the task. The *six-state model* assumed no generalization between trial types, with separate states based on odor and action. We also considered two hybrid models to probe for partial generalization: the *hybrid-value model* combined values from the four-state and six-state representations at decision time, with a relative weight parameter w_4. Finally, the *hybrid-learning model* assumed six states, but generalized across two pairs of states ("Left-Forced" and "Left-Free"; and similarly for right choices) with a generalization rate η_g.

The free parameters of each model were fit to choice data from all animals using hierarchical Bayesian inference with Markov Chain Monte Carlo (MCMC) sampling [34, 35]. We evaluated model fits using the Watanabe–Akaike information criterion (WAIC; Figure 2.3A, lower values indicate better fits to data) [36]. Model comparison showed clear evidence for the six-state model, which out-performed the four-state model with a WAIC score that was 1211 ± 78 (mean $\pm$ standard error across samples [37]) lower. The hybrid models were only slightly better than the six-state model (WAIC difference: -134 ± 24 for the hybrid-value model, and -53 ± 17 for the hybrid-learning model), suggesting little engagement of the four-state representation. Posterior estimates of the parameter values for the hybrid models revealed the dominance of the six-state representation: in the hybrid-value model, posterior estimates showed that the weight parameter w_4 was smaller than 0.5 (equal reliance on the six- and four-state representations) both at the group level (95% HDI of group mean: [0.05, 0.28]; Figure 2.3B) and for all but one rat (Figure 2.3E). Similarly, in

the hybrid-learning model, the learning rate η was an order of magnitude higher than the generalization rate η_g (95% HDI for group mean η : $[0.22, 0.28]$, η_g : $[0, 0.01]$; Figure 2.3C). In fact, most rats had a generalization rate close to zero (Figure 2.3D). These results indicate that, at the group level, rats did not acquire knowledge of the shared reward contingencies between valid forced-choice trials and free-choice trials, consistent with the earlier behavioral analysis. Instead, most rats appeared to learn about the two trial types separately.

Interestingly, we found marked heterogeneity in model fits at the individual level. Although for most animals the hybrid models did not predict choice better than the six-state model, for a subset of rats, model comparison indicated some degree of generalization (i.e., individual ΔWAIC of hybrid compared to the six-state model was negative; Figures 2.3A and 2.4). This indicated that the extent to which the animals recruited the four-state representation varied, which was confirmed by the span of individual parameter estimates of the generalization rate η_g and the four-state weight w_4. Estimates of these two parameters were positively correlated at the individual level ($r = 0.82$, $p < .001$; Figure 2.3E), indicating the consistency of the two hybrid models. For only one animal, the four-state model fit better than the six-state model; this rat also had $w_4 > 0.5$ and the largest η_g value. Comparing model fit for the first half to the second half of the behavioral sessions per individual showed the reliance on a shared representation was slightly stronger in later sessions (assessed both through comparison between the six-state and hybrid-value models, and the magnitude of the w_4 parameter in the hybrid-value model; Figure 2.5), suggesting generalization on this task may have emerged with experience.

In principle, adopting a state representation that conforms to the true generative structure of the task should afford the most efficient learning and maximum accuracy, and thus maximize reward. To test the predicted performance of models that used different representations, we simulated choice behavior using the hybrid-value model

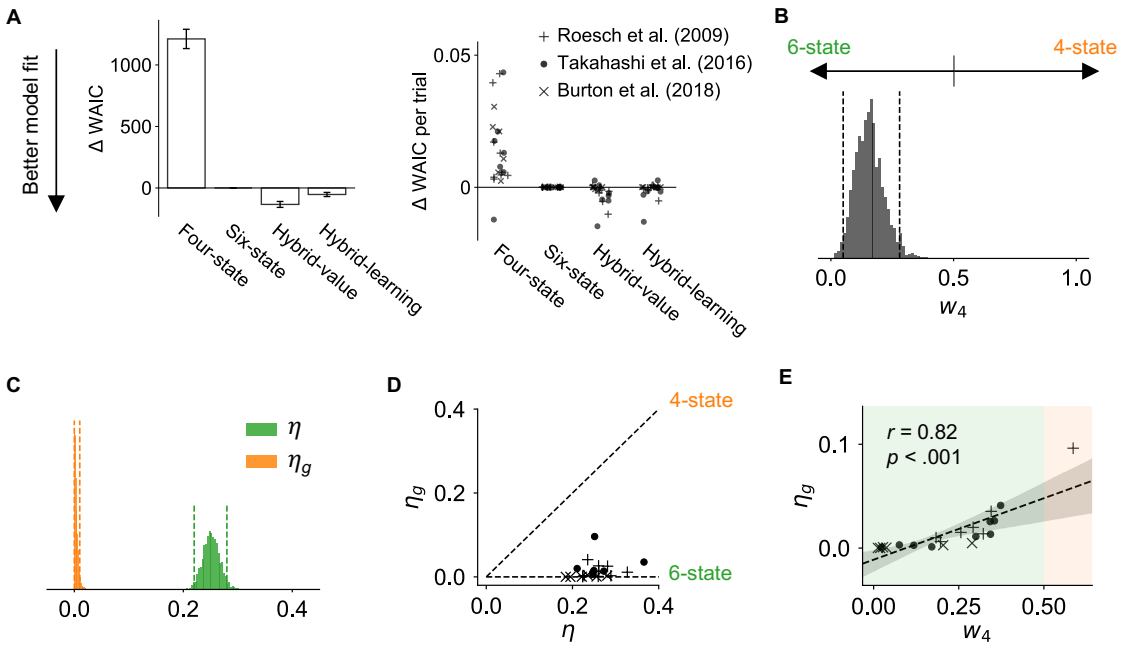

Figure 2.3: **The six-state representation explains animals' choices better than the four-state alternative.** **(A)** Model comparison results. Left: WAIC difference between all four models and the six-state model for the entire dataset (summed across all trials from all animals). Lower values indicate better model fits. Error bars represent standard errors across samples [37]. Right: individual differences in average WAIC per trial between all models and the six-state model. Each marker corresponds to an individual animal, with different markers representing different datasets (same in D and E). **(B)** Posterior distribution of the group mean of four-state weight w_4 in the hybrid-value model. Dashed lines represent 95% HDI. $w_4 = 0$ corresponds to the six-state model, and $w_4 = 1$ corresponds to the four-state model. **(C)** Posterior distribution of the group mean of learning rate η (in green) and generalization rate η_g (in orange) in the hybrid-learning model. Dashed lines represent 95% highest density interval (HDI). Generalization is almost negligible due to the low values of η_g. **(D)** Posterior mean of η and η_g for each animal. The horizontal dashed line corresponds to the six-state model; the diagonal dashed line corresponds to the four-state model. **(E)** The correlation between w_4 in the hybrid-value model and η_g in the hybrid-learning model at the individual level.

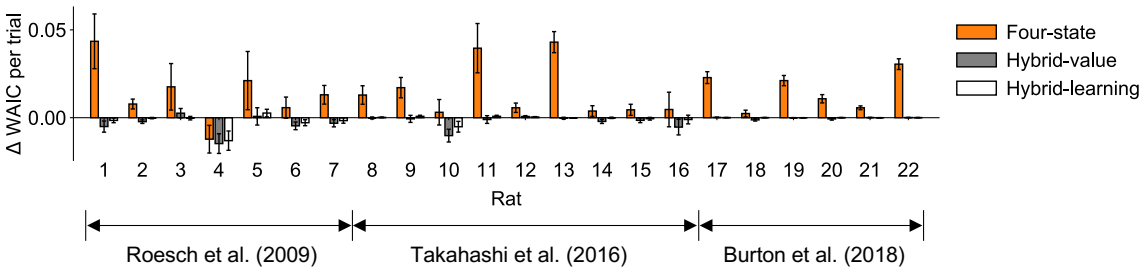

Figure 2.4: **Difference in WAIC per trial for each animal shows individual variability.** We used the six-state model as a baseline to which we compared the four-state model (in orange), the hybrid-value model (gray) and the hybrid-learning model (white). For the majority of animals, the four-state model fit much worse than the six-state. However, for a small subset, the four-state model performed equally well or even better (for rat 4) than the six-state model.

with the best-fit group-level parameters (i.e., the "average rat"; see Supplementary Methods). Changing the weight parameter w4 from 0 (equivalent to the six-state model) to 1 (equivalent to the four-state model) greatly accelerated learning in the early part of each block, as information could be appropriately generalized across trial types (Figure 2.6A). However, asymptotic accuracy at the end of a block was only slightly improved with shared reward representation, as was also reflected in the average reward yield in these simulations (Figure 2.6B). Indeed, despite the gains in learning after block changes, adopting a shared reward representation only increased reward per trial by ~ 0.05 drops across the task. Thus, in this task at least, there was not strong pressure to learn a task representation that closely matches the generative structure of the environment.

2.3 Discussion

Our results showed that most rats did not use a parsimonious state representation in the odor-guided choice task, even though, in principle, this representation could have helped them learn faster and earn more reward. Rather than exploiting a shared

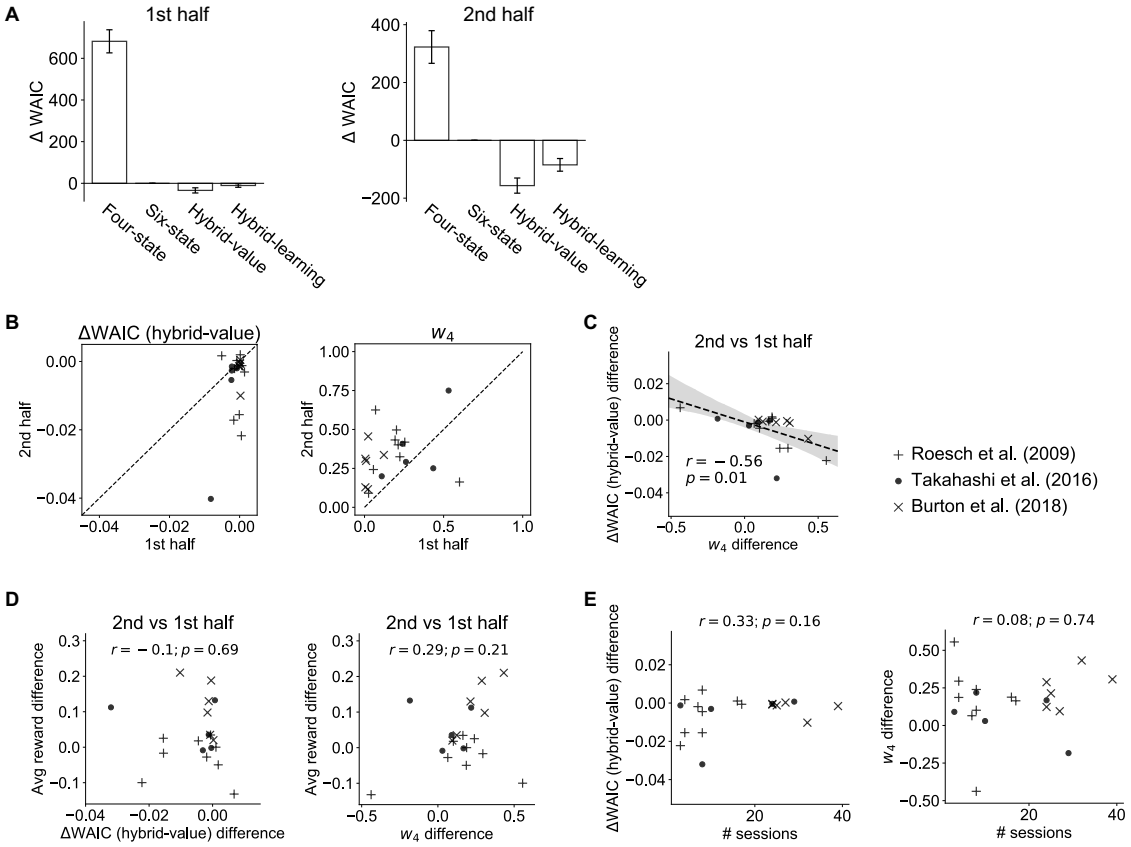

Figure 2.5: **Split-half analysis shows slow learning of the shared representation through experience. (A)** On average, the hybrid-value model provided only a modest improvement in model fit over the six-state representation in the first half of sessions per subject, however it showed a marked improvement in model fit over the six-state representation for the second half of sessions. **(B)** Most animals had very similarΔWAIC (between hybrid-value model and six-state model; same below) in the first and second halves; a small subset had a lower ΔWAIC in the second half, representing an increase in use of the shared representation. Most animals had a higher w_4 in the hybrid-value model in the second half of sessions than in the first half, also pointing towards greater generalization during the later sessions. **(C)** Difference in w_4 between the second and first half of sessions was correlated with the difference in ΔWAIC between the second and first half of sessions. **(D)** Acquisition of the shared representation did not result in more reward gains: there was no correlation between either ΔWAIC or w_4 difference (between the second and first half) with the reward amount difference ($p = .69$ and $p = .21$, respectively). **(E)** Having more task experience (more sessions performed) was not associated with a greater ΔWAIC or w_4 difference ($p = .16$ and $p = .74$, respectively). Note the animals who had the largest changes in ΔWAIC magnitude experienced very few sessions. Two animals with only one session of data were excluded from this split-half analysis. Throughout: dataset is coded by marker type.

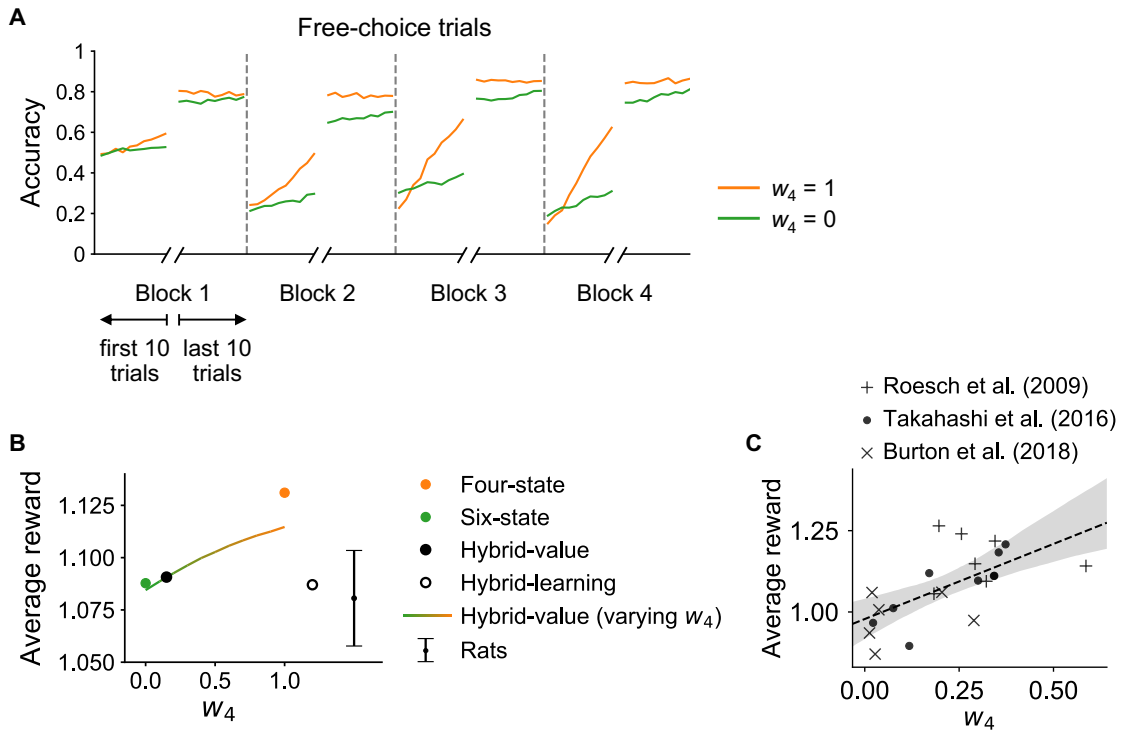

Figure 2.6: **Simulations show faster learning, but a modest increase in reward earned, under the four-state representation.** **(A)** Learning curves for free-choice trials in data simulated using the best-fit parameter values (i.e., posterior mean of the group-level parameters) of the hybrid-value model, but setting w_4 to 1 or 0 (corresponding to the four-state and six-state models, respectively). The four-state model ($w_4 = 1$) shows faster learning in each of the blocks and higher asymptotic accuracy than the six-state model ($w_4 = 0$). **(B)** Average amount of reward per trial obtained by the models (dots and curve) and by animals (error bar, mean $\pm$ 1 s.e.m.). Dots represent model-simulation results obtained with their best-fit parameter values. The colored curve represents simulation results of the hybrid-value model with its best-fit parameters but varying w_4 from 0 to 1. Average reward earned increases with w_4, but the differences are relatively small (on the order of 5%). Rats performed, on average, in line with the six-state model and markedly worse than the four-state model. **(C)** Average reward obtained by each animal is positively correlated with their mean w_4 parameter estimate ($r = 0.64$, $p = 0.0015$). Animals with a state representation more similar to four-state earned more reward on average. Dataset coded by marker type.

representation between valid forced-choice and free-choice trials, most rats learned the values of actions separately for each odor, with little to no generalization between trial types. This finding was consistent across both behavioral analyses and more detailed computational modeling approaches.

Why did most rats fail to exploit the shared reward structure of this task, even though rats have been shown to acquire quite complex task representations in other settings [38, 39, 40]? First and foremost, while the six-state representation is not as compact as the four-state and does not capture the true generative statistics of the task, it is sufficient to support good performance in this task: high accuracy for forced choice trials throughout the session and a reversal in preference between the left and right reward wells after block changes in the free choice trials. Indeed, our simulations found surprisingly little average benefit from learning the more compact four-state representation in terms of reward yield, suggesting that this task does not strongly incentivize acquiring such a representation. This might also imply that forming the more parsimonious task representation carries some cognitive cost. Indeed, the six-state representation assumes separable features for odor and location, while the four-state representation requires encoding the interaction between them. Accordingly, the prevalence of the simpler six-state representation in the choice behavior of these animals may be seen as less of a "failure" and more the result of a rational allocation of resources [41, 42].

Learned generalization in task representation has been shown in "acquired equivalence" [9], where animals respond equivalently to two stimuli that have been followed by the same consequence (e.g., food). If one stimulus is later paired with a new outcome (e.g., electric shock), the animals exhibit the same (fear) response to the other stimulus, demonstrating the shared representation. In the current task, however, odor cues only lead to the same consequence if followed by the correct action. The added complexity of instrumental contingencies perhaps limited the generalization strategies

available in purely Pavlovian settings [43, 44]. The instrumental contingencies in this task may also prompt animals to use a different learning strategy entirely, acquiring a policy over available actions in a given state directly (e.g. [45]) rather than via action value representations as we have modeled here. In policy learning, generalization between odor trial types would be limited as alternative actions are grouped together, separating forced-choice from free-choice trials through the presence of the unrewarded choice option in these trial-types [46].

We found rats to segregate learning by trial type, with little generalization between them. This may indicate that odor representations dominated other features in this task. Rats display rapid learning, excellent memory and highly discriminative responses for odors, in line with the ethological relevance of these cues [47, 48, 49]. In contrast, a preference for spatial representation might have favored the four-state model. It may be that receiving reward at the well where choice was reported interfered with learning of spatial location as a dominant state component of the task. To better understand the conditions under which generalization may be acquired, it would be interesting to investigate other choice tasks that permit different representational strategies. For example, in a similar task in which reward identity (rather than delay) is changed between blocks [50], the distinct encoding of identities may help animals group together trials with the same reward identity, potentially facilitating generalization. Further, studying the very early stages of training, including the staggered introduction of different trial types, may demonstrate critical features of early experience that favor one type of task representation over another. Future work may also benefit from examining behavioral features beyond choices (e.g. reaction times), as well as neural representations, in order to further dissect the learning of task representation. Of note, we did not test for other forms of generalization across trials in our data, for instance, whether acquired knowledge about the block structure of the task facilitates faster learning after contingency changes in subsequent sessions. This

kind of generalization has been observed in other odor-guided choice tasks as well as in numerous reversal paradigms [51, 52], and likely reflects learning of a higher-level structure of the task than we have investigated here.

Although none of the animals represented the generative task structure, few rats did acquire partial knowledge. This presents interesting directions for future studies: how was the partial knowledge acquired by these animals, and what gave rise to this individual difference in learning? Fitting our computational models to the first and second half of data separately provides some hints: overall, animals' task representation was more hybrid towards the second half, and this was driven by a small subset of animals who relied more heavily on the shared representation in later sessions (Figure 2.5A-C). However, such representation learning did not result in a higher reward gain for these animals in later sessions (Figure 2.5D). It was unclear what contributed to the differential learning effects between animals as the amount of representational change was not associated with task experience (i.e., the number of sessions performed; Figure 2.5E). Nevertheless, we can conclude from our data that the acquisition of shared representations through experience is quite slow, and longer training experience may be needed to study this learning process.

Previous theoretical and empirical studies may help shed light on the principles of representation learning that facilitate generalization, as well as individual variability in this process. For instance, models of latent-cause inference propose animals use similarity to infer whether different experiences arise from a shared latent state [53, 54, 21]. Individual differences in task representations may also arise from idiosyncrasies in immediate experience, long-term effects of development or even genetic differences [55, 56].

In designing experiments, we often choose to randomize over irrelevant features, for instance, what side a stimulus is presented on, or whether an outcome is experienced through a forced- or free-choice trial (e.g., [57]). It is tempting to assume that

our subjects also know to gloss over these nuisance task factors, however, learning to represent a task optimally is not a trivial process [58], especially when we cannot give subjects explicit instructions. Our results highlight that the factors influencing state representation in behaving animals extend beyond the experimenter-controlled generative statistics of a task, and reveal fine-grained differences in individual strategies that may be elicited in even a relatively simple reward learning task. Such discrepancies between the assumed representation and the one animals are actually using may be especially critical when interpreting neural data, but also in understanding behavioral data, and the effects of interventions. This suggests a humble approach to analysis that leans on the data—rather than an experimenter-centric view—to reveal how animals model the tasks they are performing.

2.4 Supplementary Methods

2.4.1 Subjects

The behavioral data of 22 rats (322 sessions in total) performing an odor-guided choice task (see description below) were obtained from three previous studies [31, 32, 33]. Data from 7 rats (76 sessions) were obtained from [32]: these animals had electrodes implanted in their left ventral striatum for single-unit recordings (neural data not used in this paper; same for the other two studies). Data from 9 rats (75 sessions) were obtained from the control group in [31]: recording electrodes were implanted in their left or right ventral tegmental area. Finally, data from 6 rats (171 sessions) were obtained from the sucrose control group in [33]: self-administration catheters (that were used only for the cocaine group, not the control animals analyzed here) and driveable electrodes were implanted, and a twelve-day self-administration of maximum two sucrose pellets via lever press was completed a month before the experiment. All animals received extensive prior training on the task before data acquisition.

2.4.2 The odor-guided choice task

Trial structure. Rats were trained on a well-studied odor-guided choice task [30]. The experiment apparatus is shown in Figure 2.1A. Each trial started with the illumination of a light inside the experimental box. When the light was on, a nose poke into the odor port resulted in the delivery of one of three distinct odor cues. At odor offset, the rat had 3 seconds to make a response at one of the two fluid wells located below and to the left or right of the odor port. One odor cue was reliably associated (through excessive pre-training) with reward delivery in the left well (a left forced-choice trial), a second odor was similarly associated with reward delivery in the right well (a right forced-choice trial), and a third odor was associated with reward delivery at either well (a free-choice trial). Odors were presented in a pseudorandom sequence such that 7 out of 20 trials were free choices, and the remaining were approximately equal numbers of left and right forced choices. If the rat made a correct response in a forced-choice trial, or either response in a free-choice trial, a reward was delivered, with a delay and a magnitude determined by the side of the well and the current block (see below for block structure); otherwise, the light would turn off immediately, signaling the end of the trial.

Block structure. Each session (one per day) consisted of four blocks (Figure 2.1B). All sessions started with two "delay blocks", followed by two "magnitude blocks". In "delay blocks", reward (one drop of sucrose) at one well was delivered immediately (500ms; "short"), while reward at the other well was delayed (1-7s; "long"). The timing of the delayed reward varied according to an adaptive staircase procedure to ensure a fixed proportion of "long" free choices across individual rats (see respective papers from which data were reanalyzed for details). In "magnitude blocks", the delay of reward was held constant (500ms), but the magnitude was one drop ("small") at one well, and two drops in succession ("big"; drops 500ms apart) at the other well.

Each drop of reward was a 0.05 ml bolus of 10% sucrose solution. For the first block of each session, the reward contingencies were assigned randomly to the two wells; they were then switched in the second block. In the third block, reward delivery at the "short" reward well remained the same (now called "small") while delivery at the previously "long" well became "big"; these contingencies were switched again in the fourth block. All block switches were unsignaled. Blocks were on average 70 trials long, with varying lengths (standard deviation: 14 trials).

2.4.3 Reinforcement-learning models with different state representations

We characterized the pattern of choices across a session using a series of reinforcement-learning (RL) models. We assumed the Rescorla-Wagner update rule [3], with reward discounted according to the delay between well entry and reward delivery, d (in units of seconds):

$$V_{t+1}(s) = V_t(s) + \eta \left(\gamma^d r_t - V_t(s)\right),$$

where $V_t(s)$ is the value of state s on trial t and r_t is the amount of reward (0, 1 or 2) delivered on the same trial. Learning rate η and discount rate γ were free parameters bounded in the range [0,1].

We denoted the possible choices on each trial by $a \in$ [left,right]. The decision variables governing the likelihood of left and right choices, DV(left) and DV(right), were determined by combining the value of the predicted state following that action (denoted by s_a) with a side bias term b and a perseveration term p:

$$DV(\mathrm{a}) = V_t(s_\mathrm{a}) + b \cdot I_{a,\mathrm{right}} + p \cdot I_{a,a_{t-1}},$$

where $I_{i,j}$ is 1 for $i = j$ and 0 otherwise. Thus, $b < 0$ indicates a general bias towards choosing the left side, and $b > 0$ indicates a bias towards right; $p > 0$ indicates a

tendency to repeat the same choice as the previous trial (regardless of the odor cue), and $p < 0$ indicates a tendency to avoid the preceding choice and choose the alternate action.

Decision variables for the left and right choices were compared using a softmax (logistic) function to determine the probability of each choice, with a free parameter β controlling the randomness of choices (the slope of the logistic function). Finally, we also assumed a lapse rate of λ, where lapses involved a completely random choice.

$$P(\text{left}) = (1-\lambda) \cdot \frac{1}{1+e^{-\beta(DV(\text{left})-DV(\text{right}))}} + \frac{\lambda}{2}.$$

Alternative models. Core to all RL models is the state representation of the task with which an agent is engaged. For the current task, we considered two distinct representations: four-state and six-state. We built four alternative learning models: one each of the four-state and six-state representations, and two hybrid models with mixed state representations (Figure 2.2E).

- The *four-state model* assumed full knowledge of the shared reward representation. Thus, there were four subsequent states upon choice, with free-choice trials and correct forced-choice trials sharing the same subsequent states "Left" and "Right". Reward outcomes in both trial types were used to update the value of these shared states. Incorrect forced choices led to two non-rewarding subsequent states "Left-NoRwd" and "Right-NoRwd".

- The *six-state model* assumed no knowledge of the shared reward representation. Each odor led to a separate pair of subsequent left and right states (six states in total). Reward outcomes (including no reward upon incorrect choices) were used to update the value of the subsequent state determined by the current odor and choice.

- The *hybrid-value model* assumed both four-state and six-state representations, with two sets of state values updated in parallel following each outcome. When predicting choices, the hybrid model calculated the values for left and right choices using a weighted sum of the values under each representation:

$$V(s) = w_4 V_4(s) + (1 - w_4) V_6(s),$$

where V_4 and V_6 were the state values under four- and six-state representations, respectively, and w_4 controlled the balance between the two representations. For $w_4 = 1$, the hybrid-value model was equivalent to the four-state model, whereas for $w_4 = 0$, it was equivalent to the six-state model, interpolating smoothly between the two models for the range of values of $w_4 \in [0, 1]$.

- The *hybrid-learning model* assumed a mixed representation. It had six subsequent states whose values were updated in the same way as in the six-state model, with learning rate η. In addition, generalization between free-choice trials and correct forced-choice trials occurred by using outcomes on those trials to update values of the other subsequent state with the same choice, with generalization rate η_g. For $\eta_g = 0$, the hybrid-learning model was equivalent to the six-state model, whereas for $\eta_g = \eta$, it was equivalent to the four-state model, interpolating smoothly between the two models for the range of values of $\eta_g \in [0, \eta]$.

All four models had the following free parameters: η, γ, β, b, and p. In addition, the hybrid-value model had a free parameter w_4, and the hybrid-learning model had a free parameter η_g.

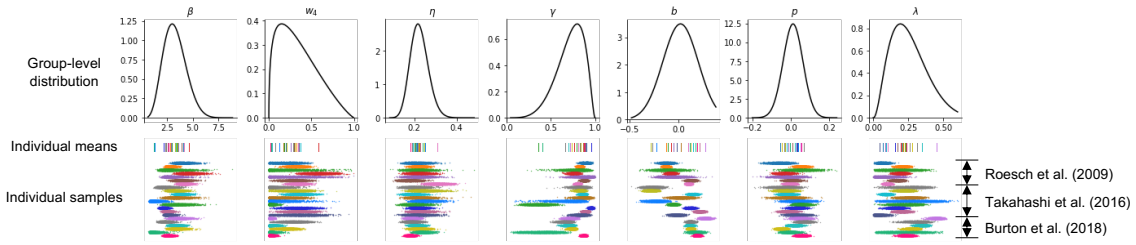

Figure 2.7: **Posterior estimates of parameter values in the hybrid-value model.** From top to bottom: the group-level posterior distributions; the posterior means of individual parameters for each animal; MCMC samples of individual parameters (sampled from a distribution with the above mean and individual variances). Different colors indicate different animals, ordered by dataset.

2.4.4 Hierarchical model fitting using Stan

In order to test whether and to what extent rats acquired and took advantage of the shared reward structure of the task, we fit the above four RL models to their choice data. Hierarchical model fitting was performed with PyStan [59], where the parameters of individual animals are assumed to be drawn from a group-level distribution. For each model, we ran 4 chains of Hamiltonian Monte Carlo with 2000 iterations each (among which 1500 were warm-up samples). Model performance was evaluated using the Watanabe-Akaike information criterion (WAIC) [36], with a lower WAIC value indicating a better fit to the data. Results from this hierarchical fitting procedure were compared to those obtained by fitting each animal individually, and the parameter estimates and model comparison results were found to be consistent.

2.4.5 Model simulation

Through hierarchical model fitting, we obtained posterior estimates of model parameters (both the group-level distribution, and individual parameters for each animal; Figure 2.7). We then simulated the model to perform the task using these parameter values. The reward contingencies in the simulation matched the original experiment, including the block sequences, total number of trials per session, the proportion of

forced-choice and free-choice trials, and the titration of the reward delay. For each model, we simulated a single agent governed by the group mean parameters (i.e., the "average rat"), which we used to calculate and compare the average amount of reward obtained (see Figure 2.6).

Chapter 3

Rats learn about underlying task structure in fear extinction through latent-cause inference

The contents of this chapter were published in: Mingyu Song, Carolyn Jones, Marie-H. Monfils, and Yael Niv. Explaining the effectiveness of fear extinction through latent-cause inference, Oct 2021. psyarxiv.com/2fhr7.

All data and code are available at

https://github.com/mingyus/fear-extinction-latent-cause-inference.

3.1 Introduction

Fear memories are notoriously hard to erase. After an association has been formed between an originally neutral cue (e.g., a tone) and some aversive outcome (e.g., a foot shock), animals cannot simply unlearn this association through standard extinction procedures (i.e., being presented with the cue repeatedly in the absence of the aversive outcome) [61]. During extinction, the animals' fear response towards the cue gradually reduces, but it usually returns if tested after a long delay (spontaneous recovery)[1, 62], or if the animal is reminded of the aversive outcome (reinstatement)[1, 63]. Associative learning theory explains these phenomena by postulating that extinction involves learning a *new* association rather than updating the original fear association [64]. However, this theory does not delineate the particular circumstances under which a new association is initiated, and how this can be avoided so that the original association may be modified through new experience.

Recent theoretical work [65] suggested a formalization of the way animals decide when to learn a new association and when to update old associations using a *latent-cause inference* framework. In this framework, all observations (e.g., cues and reinforcers) are assumed to be generated by latent (unobservable) causes that are each active for a certain (unknown) amount of time. Each latent cause has a certain tendency to generate observations, characterized by its "generative strength"[1] of each observation. While interacting with the environment, animals infer what latent cause is currently active based on their prior experience and current observations, and behave accordingly. At the same time, they learn and update their estimates of the generative strength of the currently active latent cause.

[1] In associative learning, the tendency for two stimuli to co-occur is characterized by an "associative strength". Inspired by this, here we term the tendency for a latent cause to generate observations "generative strength" to emphasize the causal/generative relationship. This concept is also similar to the "emission probability" in Hidden Markov Models.

According to this latent-cause inference framework, in fear extinction, animals infer that there are two distinct latent causes, based on their distinct tendency to generate shocks: one dangerous latent cause (active during conditioning, with a high probability of generating shocks), and one safe latent cause (active during extinction, with low or no probability of generating shocks). The reduction of fear response during extinction is a result of the animal's increasing belief that the second cause is active as more no-shock observations accumulate. Presentation of a shock in a reinstatement procedure is taken to indicate that the original dangerous cause is likely to be active again, causing the return of fear. Similarly, after some passage of time, both causes (dangerous and safe) are equally likely to be active again, leading to the spontaneous recovery of fear response as compared to that measured shortly after extinction.

In addition to explaining the return of fear, the latent-cause inference framework prescribes a solution for effective extinction of the original fear association: instead of abruptly "cutting off" the pairing between tone and shock, which encourages the animals to infer a new latent cause, gradually decreasing their co-occurrence will make animals more likely to infer that the old cause is still active during extinction, but with decreasing tendency to generate shocks. To demonstrate this, Gershman and colleagues [21] conducted two *gradual extinction* experiments (see Figure 3.1 for experimental design), and indeed they observed reduced fear responses in both a spontaneous recovery test and a reinstatement test. In both experiments, they contrasted this condition with both a standard extinction condition and a *gradual reverse* condition. In the latter, instead of gradually decreasing the frequency of the shock, they gradually increased it, while keeping the total number of shocks the same. Despite the surface similarity between the two gradual conditions (only two trials were different, though this difference markedly changed the trend of shock frequency from decreasing to increasing; Figure 3.2), extinction was far less successful in preventing

the return of fear in the gradual reverse condition, supporting the idea that abrupt changes encourage the inference of new latent causes. Since then, the effectiveness of gradual extinction has also been replicated in human participants [66], as well as in similar studies with occasional reinforcements during extinction [67, 68].

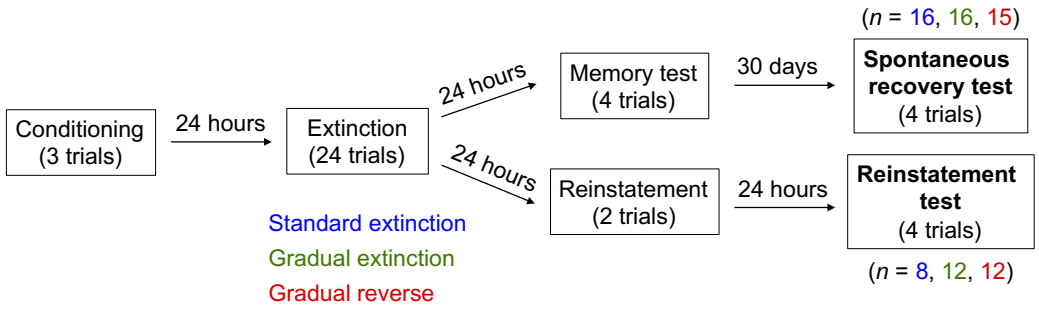

Figure 3.1: **Experimental design in Gershman et al. [21].** Across two experiments (spontaneous recovery and reinstatement), rats were assigned to three different extinction conditions: standard extinction, gradual extinction and gradual reverse. All animals first underwent a conditioning session (3 trials of tone-shock pairing). This was followed by an extinction session (procedures differed based on the extinction condition) 24 hours later. In the spontaneous recovery experiment, animals were first tested on their memory of extinction after another 24 hours (termed a "long-term memory test" in the original paper), and then tested for spontaneous recovery of fear response 30 days later. In the reinstatement experiment, animals underwent 2 reinstatement trials (shocks presented alone) 24 hours after extinction, and then were tested for fear response after another 24 hours. The number of animals participating in each experimental condition is noted for each experiment, color-coded based on the extinction procedure: standard extinction in blue, gradual extinction in green, and gradual reverse in red.

Conceptually, these effects can be explained by the latent-cause inference framework [65]; in fact, the experiments done by Gershman and colleagues [21] were inspired by this theoretical framework and aimed to test its predictions. However, it turned out that the original latent-cause inference model (presented in detail, e.g., in [69]), was not able to explain the empirical observations, in particular the difference between gradual extinction and gradual reverse conditions (S. Gershman, personal communication). The lack of a satisfactory computational account limited the con-

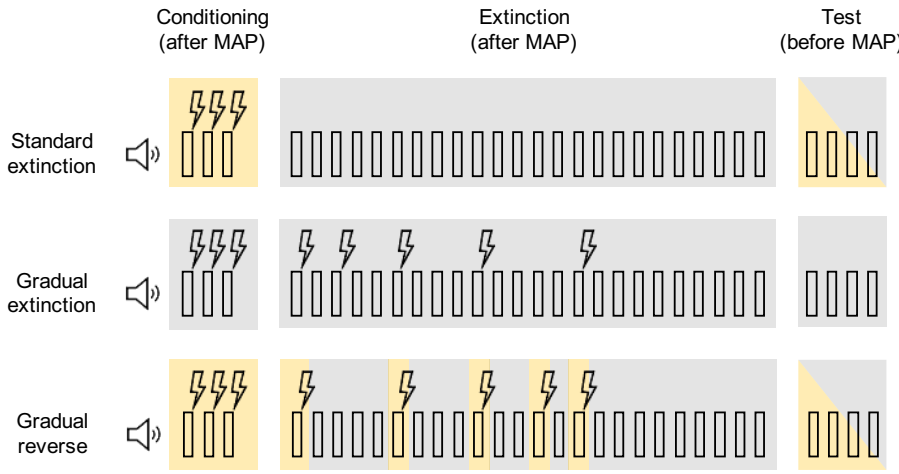

Figure 3.2: **Trial sequences and model predictions for latent-cause assignments under the three extinction conditions.** Trial sequence in conditioning, extinction and test sessions: each rectangle represents the tone presentation in one trial; trials in which the tone co-terminated with a shock are marked by the lightening signs. Background shading colors indicate latent-cause assignments predicted by our model: a two-cause sequence for standard extinction and gradual reverse, and a one-cause sequence for gradual extinction. Splitting color at test indicates that both causes are likely. The possibility of a third, new, latent cause at test is not illustrated, as it is largely consistent across conditions and contributes little to freezing behavior. For conditioning and extinction sessions, the latent-cause assignments shown are those after each session, after collapsing to the mode of the posterior (MAP estimation; see text). For the test session, we show the probabilistic assignments before the MAP estimation as those generated the behavior measured. That is, all predicted assignments shown (both for conditioning and extinction sessions, and for the test session) are presumably what animals had in mind at test time, and therefore what governed freezing behavior during test.

clusions that could be drawn from the original work, and its potential to inspire future investigation.

To address this gap, here we describe a latent-cause model with additional assumptions that captures the pattern of the original empirical results. We show the model's predictions through simulations, and demonstrate the necessity of each model assumption by comparing with alternative models. Indeed, it is not trivial for the model to generate predictions that match animals' behavior across all three extinction conditions, and several important modeling assumptions are needed. These include a specific form for the prior belief over latent causes, how generative strengths are

learned, a reduction of uncertainty in the inference process, and accounting for animals' habitual behavior. We note that even with these assumptions, our model can only predict the comparative behavioral patterns between conditions, but not the exact freezing rates in each condition. We discuss potential reasons for this discrepancy. Although the original empirical results we model here may benefit from additional examination to show its generality, we propose that understanding the potential computational underpinnings of these findings can advance our understanding of extinction processes and how they can be made more effective. More broadly, this work provides insights into the potential mechanisms that support animal learning and inference, and generates new predictions to be tested in future experiments.

3.2 Methods

3.2.1 The latent cause inference model

We use a latent-cause inference model to explain the effectiveness (in terms of degree of return-of-fear) of standard extinction, gradual extinction and gradual reverse procedures, in both spontaneous recovery and reinstatement experiments. The model describes how animals infer the active latent cause on each trial by combining a prior belief with current observations, and how animals learn about the statistics of each latent cause (i.e., its generative strength of observations). Specifically, as detailed below, the model we found to account for the experimental results uses the distance-dependent Chinese restaurant process as the prior, uses Rescorla-Wagner learning to update generative strength, and assumes that animals approximate the (intractable) Bayesian inference process by collapsing their belief to the posterior mode between sessions. Additionally, to compare model predictions to empirical measurements, we make assumptions about how the prediction of shock maps to freezing behavior. We now describe each part of the model.

Prior: distance-dependent Chinese restaurant process

We assume that the animals' prior belief over latent causes takes the form of a distance-dependent Chinese restaurant process [70], a variant of the Chinese restaurant process (CRP) infinite mixture-model prior [71].

The standard CRP describes a categorization process with an *a priori* unlimited number of categories, whereby a new trial is more likely to be generated by a latent cause (category) that has generated more trials in the past. Specifically, the prior probability of an old cause generating the current trial is proportional to the number of trials this cause has already generated; the probability of the next trial being generated by a completely new latent cause is proportional to a fixed concentration parameter α. Denoting the active latent cause on trial i by c_i, the prior probability distribution over c_i is thus:

$$P(c_i = c|c_{1:i-1}) = \begin{cases} \frac{1}{i-1+\alpha} \sum_{j<i} \delta(c_j, c) & (c \text{ is an old cause}) \\ \frac{1}{i-1+\alpha}\ \alpha & (c \text{ is a new cause}) \end{cases}$$

where $\delta(x, y)$ is the Kronecker delta function: $\delta(x, y) = 1$ if $x = y$; otherwise, $\delta(x, y) = 0$. Thus, $\delta(c_j, c)$ denotes whether the current cause c is the same one that generated trial j. $\frac{1}{i-1+\alpha}$ is the normalization constant for this distribution. This distribution is exchangeable, meaning that it results in the same prior distribution over latent causes regardless of the order of trials.

Because trial order is important in the task we model, we used the the distance-dependent Chinese restaurant process (ddCRP)[70], in which more distant experience contributes less to current inference through a decay function applied to the trial count. Specifically, we compute distance over time, using an exponential function with slope k. Since exponential decay can be arbitrarily close to zero (that is, an old latent cause not experienced for a long time can have a close-to-zero prior probability),

departing from the classic ddCRP, we also add a baseline probability b to all old latent causes. This corresponds to an assumption that any old latent cause has a non-negligible prior probability of becoming active again. The equations then become:

$$P(c_i = c|c_{1:i-1}) \propto \begin{cases} \sum_{j<i} e^{-k(t_i - t_j)} \delta(c_j, c) + b & (c \text{ is an old cause}) \\ \alpha & (c \text{ is a new cause}) \end{cases} \tag{3.1}$$

where t_i is the time of trial i. The normalization constant for this distribution can be calculated by summing over the probability of all old causes and the new cause.

Likelihood: Rescorla-Wagner learning

We denote the generative strength of each observation x for latent cause c by $V(x|c)$, which describes the tendency of latent cause c to generate x. Using the simplest and most widely-used learning rule in animal conditioning, i.e., the Rescorla-Wagner learning rule [3], we assume that V is updated for the currently active latent cause c_i based the observation of x:

$$\Delta V(x|c_i) = \eta(x_i - V(x|c_i))$$

where η is the learning rate, and x_i is the observation of x on trial i ($x_i = 1$ or 0 means x is present or absent, respectively, on that trial).

To model fear extinction, we consider two types of observations: tone and shock. Given the different *a priori* prevalence of such stimuli in the animals' natural environment, we assume different initial values for the generative strengths for new latent causes: $V_0(\text{tone}) = 0.5$ and $V_0(\text{shock}) = 0.05$. We also assume a higher learning rate η_{shock} for shocks (when shocks are present) considering their high motivational valence. These assumptions are important for the pattern of results, but the specific numeric values were not chosen through formal optimization or model-fitting.

We use the generative strength V as the proxy for the animal's estimated probability of observing the tone or shock, i.e., the likelihood of the corresponding observation on that trial given a latent cause:

$$P(x|c_i) = V(x|c_i). \tag{3.2}$$

Update: Exact inference, with collapse of belief distribution between sessions

We assume that animals perform Bayesian inference during each session. That is, they combine the prior probability and likelihood of both observations (assumed to be independently generated given the latent cause) to calculate the posterior belief distribution over the active latent cause:

$$P(c_i|\mathbf{x}, c_{1:i-1}) \propto P(c_i|c_{1:i-1}) \prod_{x \in \mathbf{x}} P(x|c_i).$$

Here, the first term on the right-hand side is the ddCRP prior from above (Equation 3.1) and the second term is the likelihood of the current observations (Equation 3.2 above). $\mathbf{x}$ denotes the combination of both observations (tone and shock). In this way, the animal maintains a probability distribution over each of the latent causes being active on each trial, updating this distribution as trials unfold.

Between experimental sessions, however, we assume that animals collapse their posterior belief distribution to its mode, i.e., a maximum *a posteriori* (MAP) estimation. In other words, we assume that animals do not maintain uncertainty over what latent cause was responsible for what observation in the previous session; instead, they "pick" the most likely sequence of latent causes for the past session, moving forward to the next session with only this deterministic assignment of trials to latent causes as the prior. Note that this is not a technical choice for faster model simu-

lation, but an important modeling assumption for predicting the difference between gradual extinction and gradual reverse conditions (see 3.3.4).

Mapping prediction of shock to freezing behavior

On each trial, as animals hear the tone and before the observation of a shock (on reinforced trials), we assume that animals use their current estimate of latent-cause assignment to predict how likely a shock is to occur on that trial, and thus decide whether or not to freeze in anticipation. The estimated shock probability is calculated by marginalizing over all possible latent causes:

$$P(\text{shock}|\text{tone}) = \sum_{c_i} P(\text{shock}|c_i)P(c_i|\text{tone}, c_{1:i-1}).$$

Here, the last term is calculated using Bayes rule:

$$P(c_i|\text{tone}, c_{1:i-1}) \propto P(c_i|c_{1:i-1})P(\text{tone}|c_i).$$

In mapping the animal's prediction of shock probability to freezing behavior, we make two more assumptions. First, we assume a non-zero baseline freezing rate (denoted by p_0): animals do not freeze in their natural environments; if, however, the animal becomes aware of shocks through experimental manipulations and anticipates them, empirical findings suggest that animals will show some baseline level of freezing behavior [72], regardless of how unlikely the shock is. In addition, for simplicity, we assume that freezing probability is proportional to the predicted shock probability:

$$P(\text{freezing}) = (1 - p_0) * P(\text{shock}|\text{tone}) + p_0.$$

We also consider the locally perseverative nature of animals' behavior [73, 74], and assume that there is a chance of p_r that the animal will exhibit the same behavior as in last trial, regardless of its current prediction for the presence or absence of shock.

3.2.2 Model simulations

We used the following parameter values for simulating the model: $\alpha = 0.2, k = 0.1, b = 0.1, \eta = 0.2, \eta_{\text{shock}} = 0.4, V_0(\text{tone}) = 0.5, V_0(\text{shock}) = 0.05, p_0 = 0.2, p_r = 0.7$. We note that simulation results were consistent across a range of parameter values. Because it is computationally intractable to compute the full posterior distribution analytically, we used the particle filter algorithm as in [65] to approximate the posterior distribution with 10,000 particles.

3.3 Results

In the following, we describe behavioral predictions of the model with all the assumptions described above (distance-dependent prior on cause assignment, Rescorla-Wagner rule for learning the generative strength of stimuli, MAP estimation between sessions, direct mapping from shock prediction to freezing behavior), and compare them to behavioral results in Gershman et al. [21]. We then turn to evaluating the necessity of each assumption by comparing with alternative models.

3.3.1 Experimental measures and modeling goals

We simulated the model and obtained its prediction for the three extinction conditions: standard extinction, gradual extinction, and gradual reverse. We considered two types of test (as in [21]): spontaneous recovery and reinstatement. Here, we first describe the experimental conditions in brief (see [21] for details on experimental design and subjects), as well as the goals of our modeling.

In the experiment, each rat completed one of the experimental conditions (Figure 3.1). All experiments started with a conditioning session (3 trials of a 20s tone, co-terminated with a 0.5s foot-shock). 24 hours later, animals underwent an extinction session (24 trials of the same tone) in one of three ways (Figure 3.2): in standard extinction, the tone was presented alone in all trials; in gradual extinction, the tone co-terminated with a shock on trials 1, 3, 6, 10 and 15, and no shocks on other trials; in gradual reverse, the tone co-terminated with a shock on trials 1, 6, 10, 13 and 15, and no shocks on other trials. In this way, in gradual extinction, shocks gradually became less frequent, whereas in gradual reverse, shocks were made to be more frequent as the extinction session proceeded. In all cases, the last 9 trials of the extinction session did not involve shocks, to allow comparable extinction of freezing behavior before the subsequent test.

Animals were then tested for their fear response to the tone, measured by the percentage of time they spent freezing during the tone. In the spontaneous recovery experiment, 24 hours after the extinction session, the animals first underwent a so-called long-term memory test to test the extinction memory (with 4 trials of the tone alone); 30 days later, they were tested again for spontaneous recovery (4 trials of the tone alone). In the reinstatement experiment, 24 hours after the extinction session, animals experienced 2 unsignaled shocks without the tone. After another 24 hours, they were tested with 4 trials of the tone alone.

Our simulations focused on model predictions for (1) latent-cause assignment throughout the experiment; (2) animals' freezing rate in the last 4 test trials. Our goals were to demonstrate the latent-cause assignments that support animals' behavior under each extinction procedure, and more importantly, explain the qualitative differences between the three procedures in terms of their effectiveness in preventing the return-of-fear: gradual extinction being the most effective (i.e., lowest freezing rate at test, compared to the end of extinction), in comparison to standard extinction

and gradual reverse. We note here, and return to this point in the Discussion, that while we strove to predict behavior throughout the experiment, due to large variability between individual rats, the behavioral pattern during extinction was hard to discern. We therefore did not attempt to quantitatively predict trial-by-trial behavior, or to fit the behavior of individual animals by using different parameter values for each animal. Instead, we focused on the average qualitative pattern of results at test, which meaningfully separated the different extinction procedures in terms of effectiveness.

3.3.2 Latent-cause assignment

Figure 3.2 illustrates the model's assignment of trials to latent causes for the three extinction conditions. In this schematic, assignments in conditioning and extinction sessions are the deterministic results after the post-session MAP estimation (see Figure 3.3A for the probabilistic assignments during the extinction session); these are the assignments that influence behavior at the test session, which is the focus of our interest. We also combine here the two types of test (spontaneous recovery and reinstatement) because their latent-cause assignments are similar (see Figure 3.3B for latent-cause probability in each test). In standard extinction, two different latent causes are inferred for the conditioning session (all shock trials) and the extinction session (all no-shock trials); at test, both causes are likely (due to either reminder shocks or passage of time, both elevating the probability of the initial conditioning-session latent cause). In gradual extinction, because of the gradual reduction of shock probability, conditioning and extinction sessions are assigned to the same latent cause, as are the test trials. In gradual reverse, in contrast, all shock trials throughout conditioning and extinction are assigned to one latent cause, whereas all no-shock trials are assigned to a different latent cause; then, at test, both causes are likely.

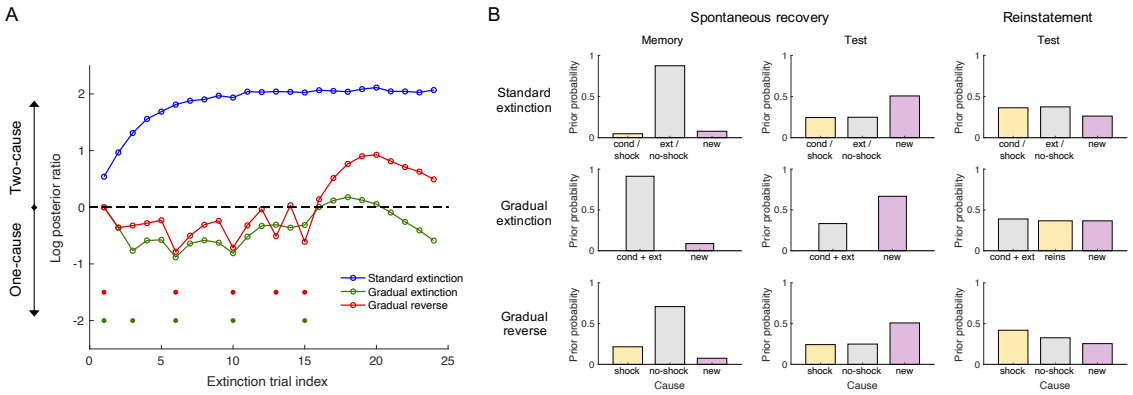

Figure 3.3: **Model prediction for latent-cause probabilities. (A) Comparison between a two-cause sequence and one-cause sequence during extinction**. Log posterior ratio greater than (less than) zero indicates the dominance of the two-cause (one-cause) sequence in inference; the dashed line at zero represents the two types of cause sequence being equally likely. At the end of the extinction session, the two-cause sequence is more likely than the one-cause sequence in standard extinction and gradual reverse; in contrast, the one-cause sequence is more likely in gradual extinction. Here, "two-cause" sequence corresponds to the assignment of all shock trials to one cause, and all non-shock trials to another; "one-cause" sequence corresponds to all conditioning and extinction trials being generated by the same cause. **(B) Prior probability of latent causes in the first trial of the long-term memory and test sessions**. Left and middle columns: long-term memory and test sessions in the spontaneous recovery experiment; right column: test session in the reinstatement experiment. Top, middle and bottom rows: standard extinction, gradual extinction, and gradual reverse conditions, respectively. Causes are labeled based on what types of past trials they have generated: conditioning (cond), extinction (ext), reinstatement (reins), shock or no-shock. They are color-coded as in Figure 3.2: yellow indicates a "dangerous cause", grey indicates a "safe cause", and purple indicates a new cause (with minimal prediction of shock, $V_0(\text{shock}) = 0.05$).

3.3.3 Prediction of freezing behavior

Figure 3.4 shows model prediction for freezing rate across experiment sessions. For all three extinction conditions, freezing rate increases during conditioning, decreases during extinction (more so in standard extinction, with no shocks in the extinction session), and continues to decrease in the long-term memory test (24 hours after extinction in the spontaneous recovery experiment). However, in the test session (both spontaneous recovery and reinstatement experiments), the predictions for the three conditions diverge, due to the distinct latent-cause assignments: in standard extinction and gradual reverse conditions, freezing rate increases compared to the end of extinction, showing the return of fear; in the gradual extinction condition, it continues to decrease both in the spontaneous recovery and the reinstatement tests.

Figure 3.5A and 3.5B summarize model predictions on the difference in freezing rate between the four test trials and the last four extinction trials, for comparison with the qualitative pattern in the empirical results (Figure 3.5C and 3.5D). According to the model, for both spontaneous recovery and reinstatement tests, fear response reduces the most in gradual extinction, followed by gradual reverse; fear response at test increases in standard extinction. We note here (and discuss in more detail below) the clear discrepancies between our simulation results and the empirical results, where in simulation gradual reverse does not result in overall increase in fear at test as compared to the end of extinction. Nevertheless, the model predictions are qualitatively consistent with the empirical findings, illustrating the success of the current model in explaining the relative pattern seen in the experimental results.

3.3.4 Necessity of model assumptions

The model described above was tailored to explain the behavioral patterns in the data by adding assumptions as needed where the original CRP model [65] did not suffice. We now turn to discussing these model assumptions and demonstrating their

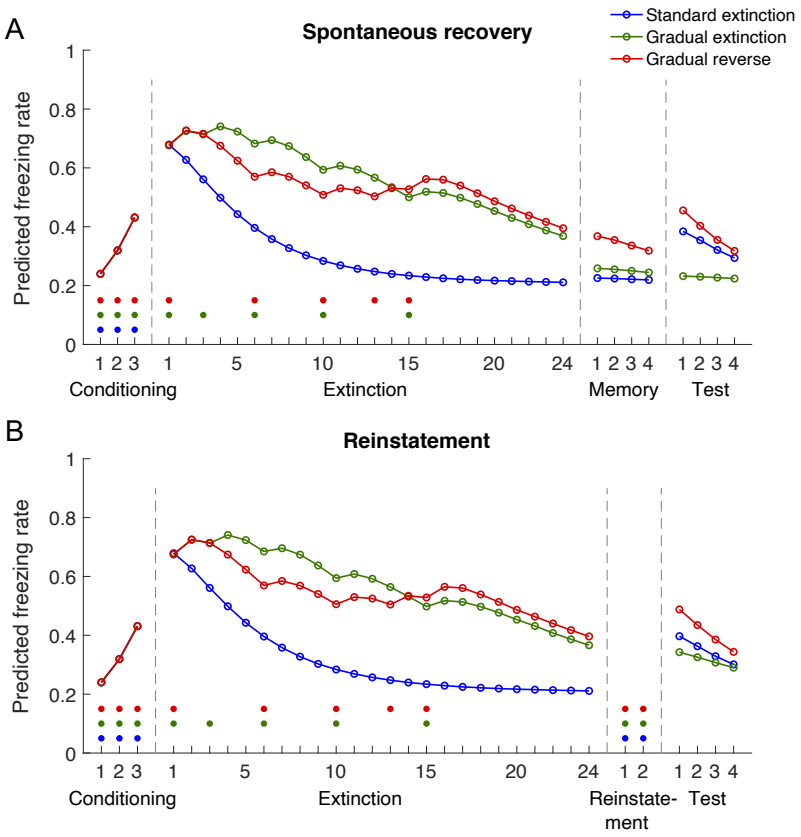

Figure 3.4: **Model prediction of freezing rate for (A) spontaneous recovery and (B) reinstatement experiments.** Under all three extinction conditions, freezing rate increases during conditioning, and decreases during extinction. In the beginning of test session, however, freezing rate jumps back up for standard extinction and gradual reverse conditions, but remains low in gradual extinction. Note that the model predicts freezing rate upon tone presentation, before the actual delivery (or absence) of shock. Dots at the bottom of the plots indicate shocks in the corresponding trials, color-coded based on the extinction conditions. Dashed gray vertical lines indicate session boundaries (in practice: at least 24h gap).

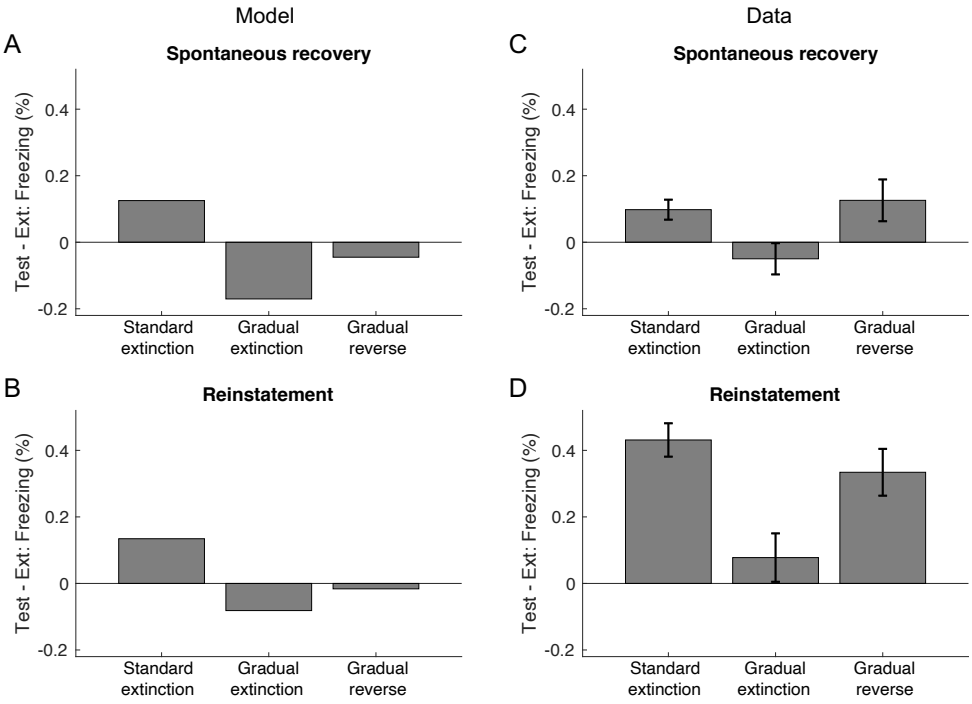

Figure 3.5: **Impact of different forms of extinction on return of fear: model predictions (A,B) are qualitatively consistent with empirical results (C,D) in both experiments.** Model simulations correctly predict that standard extinction leads to the greatest return of fear, whereas gradual extinction is the most effective in permanently reducing fear, across both spontaneous recovery and reinstatement tests. The effectiveness of extinction is calculated as the difference in freezing rate between the four test trials and the last four extinction trials (all no-shock trials). Panels C and D are reproduced from [21].

necessity. We do so by comparing the predictions of the current model (referred to as "main" model below) to reduced models that do not include these assumptions. Since we did not fit the parameters of the main model to the data using statistical techniques, we do not present formal model comparisons, but rather focus on the qualitative patterns in the data.

Distance-dependent CRP: spontaneous recovery depends on test delay

Experiments suggest that spontaneous recovery of fear response increases with the delay between extinction and test [1, 62]. For example, the standard extinction group

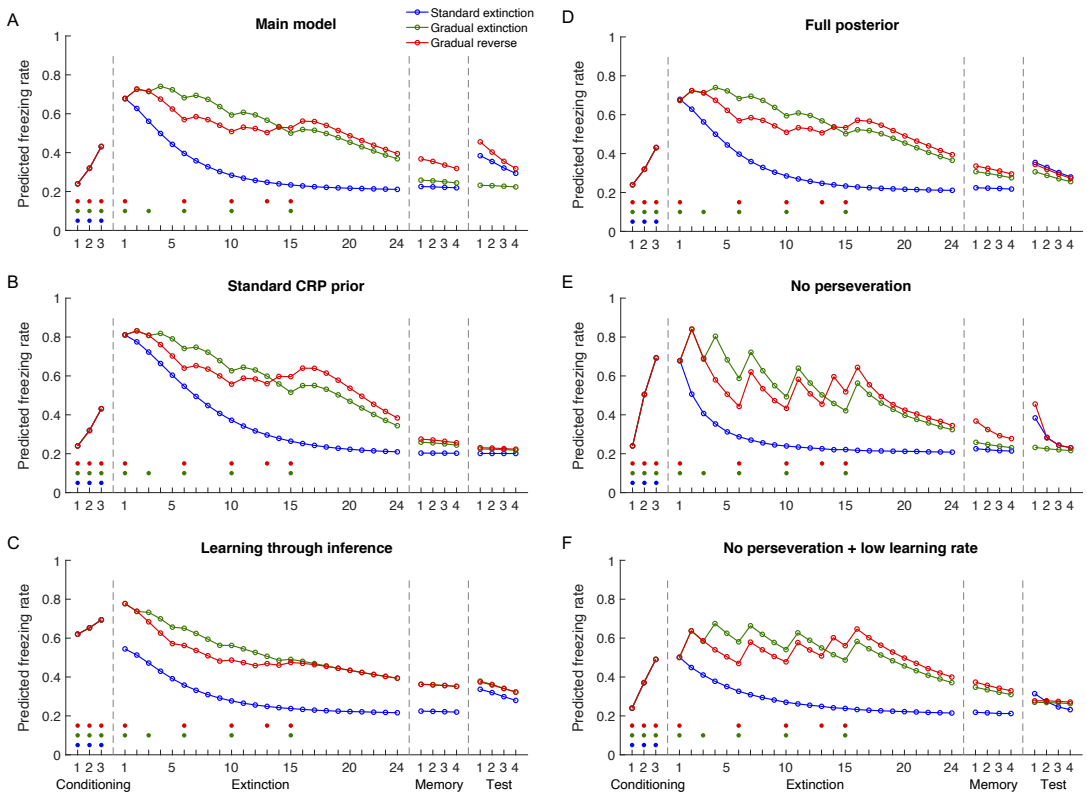

Figure 3.6: **Simulation of alternative models demonstrates the necessity of assumptions in the main model.** Each alternative model differs from the main model in only one assumption; model simulations were conducted with the same parameter values, except for the alternative assumptions, as noted below. Shown are simulations of the spontaneous recovery experiment; results are consistent for the reinstatement experiment. **(A)** The main model, same as Figure 3.4A. **(B)** Standard CRP prior (without distance dependence; $k = 0$ in the ddCRP) assigns trials in the memory test (24 hours later) and spontaneous recovery test (30 days later) to the same latent causes, and thus shows no spontaneous return-of-fear after 30 days for any of the conditions. **(C)** Learning generative strengths through inference (i.e., Bayesian inference on Bernoulli probabilities for tone and shock, instead of Rescorla-Wagner learning) predicts the same latent-cause assignments and expected shock probability for gradual extinction and gradual reverse conditions, and thus cannot explain their difference in fear responses during test. In this alternative model, we replaced the Rescorla-Wagner learning parameters $\eta, \eta_{\text{shock}}, V_0(\text{tone}), V_0(\text{shock})$ with parameters for pseudo-counts: $N(\text{tone}) = 0.5, N(\text{no-tone}) = 0.5, N(\text{shock}) = 0.95, N(\text{no-shock}) = 0.05$. These pseudo-counts determine the prior probabilities of tone and shock (each denoted by x): $p(x) = N(\text{x}) / \left(N(\text{x}) + N(\text{no-x})\right)$, reflecting a similar asymmetry in the *a priori* prevalence of tone and shock as in the main model. **(D)** Keeping the full posterior distribution over latent causes across sessions (no MAP estimation) predicts minimal return-of-fear in the gradual reverse condition. **(E)** No perseveration ($p_r = 0$) leads to over-sensitivity to shock/no-shock experience during extinction which was not seen in the empirical data, as well as rapid reduction in return-of-fear during test. **(F)** Replacing the perseveration assumption with lower learning rates to stabilize responding across trials ($p_r = 0, \eta = 0.1, \eta_{\text{shock}} = 0.2$) fails to predict the return-of-fear effect in the gradual reverse condition.

in Gershman et al. [21] showed no significant return of fear in the memory test 24 hours after extinction (paired sample t-test: $t(15) = -1.80, p = .09$, in comparison to the last four extinction trials). In contrast, the rats showed significantly more freezing at test 30 days later, compared to the last four trials of extinction (paired sample t-test: $t(15) = 3.26, p < .01$).

Our model captures the dependency of spontaneous recovery on delay duration using the distance-dependent CPR prior, similar to what was used in [75]. According to this prior, the probability of an old latent cause being active depends on the time between its previous instances and the current trial. The closer they are in time, the more likely the old cause will be active again. As a result, the model predicts that the latent cause inferred for the extinction session is very likely to continue to be active in the memory test (Figure 3.3B left column), as the memory test is much closer to the extinction session (24 hours apart) than to the conditioning session (48 hours). After 30 days, however, both conditioning and extinction sessions are similarly distant, so the model predicts that both the conditioning and extinction latent causes are equally likely to be active (Figure 3.3B middle column; of course, a completely new latent cause is more likely in this case), resulting in an increased fear response in standard extinction and gradual reverse. In gradual extinction, since there has been only one latent cause throughout, this cause (with reduced generative strength of shock) is inferred to be active again during test, resulting in no return-of-fear above and beyond what was observed at the end of extinction.

An alternative model that uses the standard CRP prior with no distance dependence (and keeps all other assumptions and parameter values the same as the main model) produced latent-cause assignments that are independent of test time. It thus made similar predictions for the memory test and the spontaneous recovery test, failing to predict the spontaneous recovery of fear in standard extinction and gradual reverse conditions (Figure 3.6B).

Note that we are not committed here to the specific form of the distance dependence and have not exhaustively tested other forms of recency-weighted or perseverative CRP priors (e.g., [76]), or different decay functions. Our claim is only that an order-agnostic exchangeable prior distribution, as is commonly used in machine learning applications to categorization, would not accord with the empirical data.

Rescorla-Wagner learning (recency-weighted estimates): difference between gradual extinction and gradual reverse

In the latent-cause inference framework, it is common to assume that the generative strength of observations is fixed (though unknown) for each latent cause [65, 69]. Under this assumption, the optimal estimator of the shock probability for each latent cause is the proportion of trials in which shocks appeared under that cause. This assumption is at odds with the behavioral differences between gradual extinction and gradual reverse, as both conditions had the same number of shocks and would therefore predict identical test behavior (Figure 3.6C).

To explain the behavioral difference between gradual extinction and gradual reverse, we thus found it necessary to assume a recency-weighted estimate of generative strength, such as given by the Rescorla-Wagner (RW) learning rule. Similar learning rules have been used in latent-cause inference models of associative learning and memory modification [77, 75]. The RW rule learns a dynamic shock probability through an error-correcting process that adjusts estimates proportionally to the error in predicting the current observation. A fixed learning rate results in over-weighing of recent experiences, effectively estimating the generative strength as an exponentially-decreasing average of previous experiences. This is normative if animals assume that the environment may change over time and track such changes. In our case, this allows them to treat differently a decrease in shock probability in gradual extinction versus an increase in gradual reverse. As a result, they make different latent cause

assignments for gradual extinction and gradual reverse, and in turn, show distinct levels of fear responses at test.

Collapsing uncertainty (MAP) between sessions: return-of-fear in gradual reverse

Latent-cause inference involves evaluating the likelihood of all possible cause sequences for the past trials – from all trials being generated by one cause, to each trial being generated by its own unique cause, through any combination in between. Probabilistic inference means that the model does not commit to any of these assignment sequences; instead, all are likely during inference, and each is associated with some non-zero probability. In particular, in all three extinction conditions, both a one-cause sequence and a two-cause sequence are likely (Figure 3.3A). While the two-cause sequence dominates in standard extinction, the assignments for gradual extinction and gradual reverse are more uncertain, starting with similar probabilities in both conditions and diverging through extinction. At the end of the extinction session, the one-cause sequence establishes some advantage over the two-cause sequence for gradual extinction, and the opposite is true for gradual reverse. The MAP assumption collapses this uncertainty over latent-cause sequences at the end of the extinction session, obtaining the deterministic (and different) assignments shown in Figure 3.2. Specifically, the model commits to the one-cause assignment for gradual extinction, and the two-cause assignment for gradual reverse, accentuating the small differences between these conditions at the end of the extinction session. Similar collapsing of uncertainty has been applied in domains like perceptual decision-making (e.g. [78]) to facilitate inference and decisions where estimating the full posterior distribution is computationally intractable.

We found the MAP assumption to be necessary for predicting the return-of-fear effect in gradual reverse. An alternative model that keeps the full distribution over

latent causes throughout the experiment failed to predict the increase in freezing rate during test for gradual reverse (Figure 3.6D). This is because the two-cause sequence was deemed only somewhat more likely during test, which resulted in minimal return of fear.

Perseveration: gradual change in behavior during extinction, and persistence of return-of-fear during test

In the main model, we postulate that animals tend to repeat what they have been doing in past trials with probability p_r. Such perseverative behavior has been widely observed in perceptual and value-based choice tasks, for both animals and humans [79, 80, 81, 82]. In the current task, perseveration is important for predicting the persistence of return-of-fear during the test session. Without this assumption, fear responses will decrease rapidly in both standard extinction and gradual reverse conditions, as early as the second test trial (Figure 3.6E). This is due to animals' inference of latent causes: experiencing one trial without shock is sufficient to infer that the "safe cause" is active in the test session, and due to the distance-dependent CRP prior, this inference comes to dominate latent-cause assignments at test. However, such rapid reduction in freezing behavior was not seen behaviorally.

Simulations without perseveration also predicted freezing behavior that was overly sensitive to shock and no-shock experiences during extinction (Figure 3.6E): in gradual extinction and gradual reverse, simulated freezing rate jumped up following shocks and dropped following no-shocks. Examining animals' freezing rate during extinction (by re-scoring of the original videos using a convolutional neural network [83], as the middle trials during extinction session were not scored or reported in the original paper) suggested that behavior did not reflect trial-by-trial shock delivery or absence; on average, freezing rate noisily but gradually decreased through extinction session in all three conditions (not shown). An alternative explanation for the more gradual

behavioral change in extinction can be a lower learning rate in updating the generative strengths of latent causes. However, reducing the learning rate interfered with latent-cause inference as it rendered animals less sensitive to shocks and thus biased them towards the one-cause assignment. An alternative model where we halved the learning rates predicted no return-of-fear in gradual reverse (Figure 3.6F).

We therefore used the perseveration assumption to account for both rapid learning in fear conditioning and extinction, and gradual change in behavior. Additionally, we note that due to local perseveration, the main model predicted an increase in freezing rate at the beginning of the extinction session (compared to the end of conditioning session; Figure 3.6A), potentially related to the well-documented "extinction burst" phenomenon [84, 85].

3.4 Discussion

In this work, we use a latent-cause inference model to explain the differential effectiveness of gradual versus abrupt (standard) fear-extinction procedures. The model explains the return-of-fear effect commonly observed in standard extinction by presuming that animals infer a new state of the world during extinction and thus form a new association between tone and shock, as opposed to unlearning the original association. Similar ideas have been proposed both under similar statistical inference framework [53, 65] and through reinforcement-learning models with a state-classification mechanism [54]. This explanation also aligns with decades-old suggestions that extinction results in learning of a new safe association that competes with the original threat association, though does not override it [64]. The novelty of our work lies in formulating the inference and learning processes, demonstrating with quantitative simulation the differences between three extinction procedures, and verifying the necessity of various model assumptions.

We show through model simulations that animals make distinct inferences of latent cause assignments under different extinction procedures: gradual extinction is the most effective at extinguishing the original fear association because both conditioning and extinction sessions are assigned to the same latent cause, and the gradual reduction in shock helps animals acquire a decreasing estimate of shock probability, leading to minimal return-of-fear during test. In contrast, the abrupt change in shock frequency in the other two procedures (complete absence of shocks in standard extinction; abrupt reduction in shock appearance, followed by increasing shock frequency in gradual reverse) results in the creation of a new "safe" latent cause to which all subsequent no-shock trials are assigned. This new latent cause protects the old "dangerous" latent cause from being updated by the no-shock experiences. As a result, the original fear association remains intact and can resurface at test.

3.4.1 Additional model assumptions

We found that several additional model assumptions were needed to predict the behavioral findings: (1) a distance-dependent prior on latent-cause assignments that used the passage of time to determine distance; (2) learning the dynamics of the environment through a recency-weighted rule such as the Rescorla-Wagner learning rule; (3) reduction of uncertainty over the posterior distribution through MAP estimation between sessions; (4) behavioral perseveration. Without each of these assumptions, the model failed to replicate the higher freezing rates at test in standard extinction and gradual reverse conditions in comparison to gradual extinction. Most of these assumptions build on past models that have successfully explained a wide range of phenomena in animal (and human) learning and decision-making, as we discuss below. Moreover, the necessity of each assumption also suggests principles of animal learning mechanisms in Pavlovian tasks and beyond.

The distance-dependent CRP prior we used highlights the important role of time in latent-cause inference, especially when behavior is examined at different time intervals, or even on different time scales. We are not the first to propose such time-dependency. Gershman and colleagues [75] used a similar distance-dependent CRP prior (with a power-law temporal kernel) to explain how spontaneous recovery depends on extinction-test interval, as well as why the effect of post-retrieval memory modification is sensitive to memory age. Additional empirical evidence for time-dependent inference process comes from work on extinction delay. Myers and colleagues [86] found that varying the delay between acquisition and extinction affected the amount of return-of-fear: animals that experienced extinction trials 10 minutes or 1 hour after acquisition showed little or no subsequent spontaneous recovery, reinstatement or renewal effects; whereas those who had extinction trials 24 or 72 hours after acquisition showed strong return of fear effects. This effect of extinction delay cannot be explained by a latent-cause inference model without temporal dependency. The distance-dependent CRP prior can explain these findings as a result of the decreasing probability of the extinction trials being generated by the same latent cause as the acquisition trials when the delay between extinction and acquisition increases. Thus, with a shorter delay (10 min or 1h), the animal is more likely to classify the extinction trials into the same latent cause as the acquisition trials, which helps the successful unlearning of the shock probability, and thus prevents the return of fear.

The importance of time is also reflected in the recency-weighted learning rule we used. The Rescorla-Wagner learning rule has been widely used to model classical conditioning. It has also been used in latent-cause inference models to explain compound generalization in associative and causal learning [77] and memory modification [75].

Together, the distance-dependent CRP prior and the Rescorla-Wagner learning rule imply specific beliefs that animals may have about the environment. Both modeling assumptions reflect the animal's inner model of the environment, i.e., whether

they consider it as static or changing over time. According to the distance-dependent CRP prior, latent causes that were last active a long time ago are less likely to be active again, suggesting that the passage of time can lead to changes in the environment, making older causes less likely. Similarly, the Rescorla-Wagner learning rule over-weighs more recent experience, inherently capturing how the statistics of observations within a latent cause may change over time. Alternative assumptions (standard CRP prior and learning through exact Bayesian inference) are better suited to static environments. The advantage of the current model over alternatives provides evidence that animals are capable of acquiring rich dynamics in their learning environment, and that their inner model of the environment is in accord with the actual changing nature of naturalistic environments.

We can further consider a more general approach to modeling learning in a changing environment: deriving the normative learning rule based on the generative model. For example, the Kalman filter model [87] has been suggested as a model for estimating the mean and standard deviation of a Gaussian reward distribution under the assumption that it evolves over time[2]. Another option, normative for environments in which the amount of reward changes at a constant rate, is to add a "momentum" term (calculated as the running average of recent prediction errors) to the Rescorla-Wagner learning rule [89].[3] Future work can derive normative learning rules for the three extinction procedures in the current experiment, and test whether they account for behavior better than the Rescorla-Wagner learning rule we used.

Compared to the above two assumptions, the MAP assumption (collapse of the posterior to its mode between sessions) is found less often in the animal-learning literature. This assumption, nevertheless, also reflects the effect of the passage of time on inference, and can be construed as a result of memory consolidation (e.g., via replay

[2]In fact, Rescorla-Wagner learning can be seen as a special case of the Kalman filter with a fixed Kalman gain and without tracking uncertainty [88]

[3]We also tested an alternative model with momentum; the results were largely consistent with the RW learning rule. Thus, for simplicity, we used the more basic RW rule in the main model.

of past events during sleep [90]). Through consolidation, animals may revisit their learning experience in the previous session, and continue to update their belief over latent causes accordingly. Such update can reinforce the most probable latent-cause assignment, and eventually make it the only possibility, equivalent to a MAP estimate. Similar consolidation mechanisms have been proposed during long inter-trial intervals. For example, to explain memory modification, Gershman and colleagues [75] introduced a "rumination" process taking place between the re-exposure of previous memory and the attempt to modify/extinguish it. Such rumination reinforces the dominant belief of returning to the past context, and facilitates memory modification. There is also extensive evidence on the superiority of spaced learning (with longer intervals between training examples, including overnight session boundaries) over massed learning, facilitated by memory consolidation [91], in both animals and humans [92, 93, 94, 95].

From a normative perspective, MAP estimation may be justified to allow inference in complex scenarios where the full distribution over latent causes is computationally intractable. Because of the combinatorial explosion of possible latent-cause sequences as more trials are experienced, any experience that unfolds over a sequence of events quickly becomes intractable. It is thus reasonable to assume that the brain collapses uncertainty over previous inference periodically (perhaps facilitated by large gaps in experience, replay of experiences, and memory consolidation), to be able to continue building on past knowledge without necessarily maintaining all of it. Indeed, Stocker and colleagues [78, 96] have shown that biases in human decision making can be explained by postulating that people collapse parts of their posterior distribution between sequential decisions, essentially discarding beliefs that are inconsistent with actions they have already made. Our MAP assumption is a similar form of commitment to the most likely past beliefs over others.

The MAP assumption provides novel theoretical predictions that can be tested in future experiments. A direct test of the memory consolidation account of the MAP estimation would be to reduce the time between extinction and test in the reinstatement experiment. For example, conducting extinction, reinstatement and test all within one day or even one session. If memory consolidation indeed facilitates the collapse of the posterior distribution and strengthening of one interpretation of past events, animals with no or less time between extinction and test should maintain the probabilistic belief at test, and thus show less freezing in the gradual reverse condition. Our model does not differentiate between an abrupt collapse of the posterior distribution and a gradual reduction of posterior uncertainty. To further examine the process of uncertainty reduction, future work can manipulate the time interval between extinction and test, and examine how freezing rate at test changes as a function of such time interval.

The reduction of uncertainty due to the MAP estimation may seem at odds with the increase in uncertainty over time that results from the dynamic generative model as discussed earlier. However, it is worth noting that the MAP reduction of uncertainty affects categorization of past experiences, whereas the increase in uncertainty pertain to predicting future experience. On the one hand, the collapse of the posterior may suggest the limitation of animals' representation of the world; perhaps animals are able to learn a rich representation, but fail to maintain uncertainty about this representation over long periods of time. On the other hand, MAP estimation can also be seen as a method by which animals (and humans) build concise models of the world.

Last but not least, the perseveration assumption suggests a value-free habitual system that exists alongside a model-based system corresponding to the latent cause inference mechanism in the current model. Such a dichotomy has been widely observed in animal learning and decision-making [73]. The fact that animals' behavior

does not fully reflect their learned world model or even stimulus values underscores the importance of accounting for common habitual behavior (e.g., perseveration, side-bias, etc.) in modeling so that the habitual part of behavior will not mask the rich learning mechanisms.

3.4.2 Related empirical results

In this work, we provided a quantitative and theoretical account explaining why gradual extinction is more effective in permanently reducing fear than standard extinction and gradual reverse procedures. We focused specifically on predicting the empirical findings in Gershman et al. [21]; however, it is worth noting that there is other evidence on the effectiveness of gradual extinction.

In a fear-extinction experiment with human participants [66], gradual extinction was shown to prevent the return of fear better than standard extinction, as measured by startle response (although there were no effects for contingency rating or skin-conductance response). Additional evidence comes from occasional reinforcement experiments: having occasional reinforced trials during extinction (effectively reducing the shock probability more gradually compared to standard extinction procedure) has been shown to eliminate spontaneous recovery in both rodents and humans [67, 68]. Similarly, in appetitive conditioning experiments, occasional presentations of reinforcement in extinction slowed the re-acquisition of conditioned responses [97, 98], suggesting unlearning during extinction rather than learning of a competing association that would allow rapid relearning by activating the original association.

We note that seemingly opposing evidence was observed in a reinstatement experiment by Rescorla [63], where gradual extinction was less effective than standard extinction in preventing the reinstatement effect. However, the reminder shocks used in this experiment were very weak (compared to other reinstatement experiments reported in the same study: 0.5mA vs 3mA), and therefore may not have been per-

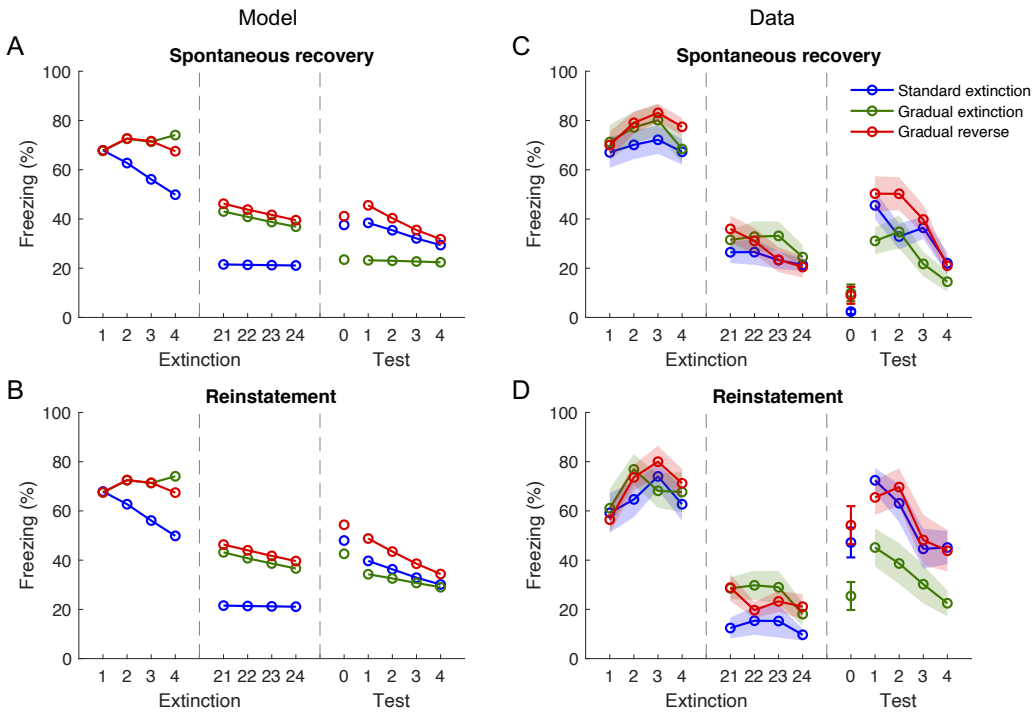

Figure 3.7: **Freezing rate comparison between model predictions (A,B) and empirical results (C,D).** Shown are the first and last four trials of extinction, the beginning of the test session (before the first test trial; marked as trial 0), and the four test trials. Panels C and D are reproduced from [21].

ceived as aversive stimuli by the animals [99]. Without valid reminder shocks, the test trials would simply reflect the effect of extinction after a short delay (similar to the long-term memory test in the spontaneous recovery experiment in Gershman et al. [21]). Indeed, the gradual extinction group showed a higher conditioned response than the standard extinction group in the memory test in Gershman et al. [21] ($t(30) = 3.05, p < .005$), consistent with the findings by Rescorla [63].

3.4.3 Limitations: differences between model predictions and empirical results

Although the current model captures key aspects of the empirical findings, its predictions deviate from the experimental data quantitatively. First, the model correctly

predicts the comparative differences between the three extinction conditions, but fails to predict the absolute return-of-fear effects. It over-predicts the reduction of fear for gradual extinction and gradual reverse, as compared to the empirical results (Figure 3.5). In fact, the model predicts lower fear responses in gradual extinction at test, and no change in gradual reverse when compared to the end of extinction, whereas animals showed minimal change and an increase in fear in these two conditions, respectively. Similarly, the predicted freezing rate during extinction deviates from the empirical results quantitatively (Figure 3.7).

However, it is worth noting that behavioral variability was profound in these experiments. For instance, despite having the exact same procedures in conditioning and extinction sessions, the extinction effects differed in spontaneous recovery and reinstatement experiments (Figure 3.7). Specifically, during the last four trials of extinction, freezing rate was significantly different between the two experiments under the same standard extinction procedure (one-way ANOVA: $F(1, 22) = 4.38, p < .05$; there were no significant difference for gradual extinction or gradual reverse procedures). These potentially different freezing rates at the end of extinction, nevertheless, served as the baselines for comparing test behavior in Figure 3.5, and provided the "ground truth" for comparison with our simulation results. Given this variability across animals and between experiments in the empirical findings, we decided to forego matching the empirical results quantitatively, and instead focused on correctly predicting the comparative difference between extinction procedures: gradual extinction being the most effective in reducing the return of fear, compared to either standard extinction or gradual reverse.

Quantitative deviations between the model's predictions and the empirical results may also be due to non-linear mapping between estimated shock probability and animals' freezing behavior, which we did not model. Because the form of this mapping was not the focus of the current work, we assumed a linear function for simplicity, but

this is likely incorrect. Future work can design targeted experiments to investigate the underlying mechanism of this mapping.

Finally, the model captures behavior at the group level but does not make predictions regarding individual differences, which were marked in the empirical results (also observed in similar studies in humans [100]). To make individual predictions with the current model, we can use different parameter values for each animal to capture their distinct learning, inference and behavioral processes. For example, animals often demonstrate abrupt switching (rather than gradual changes) in behavior that only seem gradual at the group level because different animals switch behavior at different times [101]. This between-animal variability in change points may be the result of different mapping functions from shock prediction to freezing rate. Another possibility is that individual animals do not perform full Bayesian inference but only take a small number of noisy samples from the probability distribution; in this case, aggregating over a group of animals will result in average behavior that resembles full inference, all the while each individual shows what seems like an idiosyncratic pattern of behavior [102]. Under this explanation, the current model is only adequate for explaining group-level behavior, and we need an additional sampling model to capture individual behavior. Both accounts for individual differences are likely in the task we modeled. We leave for future work to examine these sources of variability by fitting models to individual animals' behavior and comparing their predictions.

3.4.4 Conclusion

In sum, our work explains the effectiveness, or lack thereof, of different behavioral manipulations aimed at reducing maladaptive fear responses. Our results suggest that in acquiring and extinguishing fear responses, animals form a dynamic model of the environment, using inference of latent causes to predict future events. When representing and memorizing past experiences, however, our model suggests that animals

summarize previous inference by collapsing distributions over potential latent-cause explanations and preserving the most likely explanation. Our findings suggest that, even in simple Pavlovian tasks such as fear extinction, animals' behavior can reveal a rich array of mechanisms of learning and representation.

Chapter 4

Humans learn about complex rules through value-based serial hypothesis testing

The contents of this chapter were submitted for publication in: Mingyu Song, Persis A. Baah, Ming Bo Cai, and Yael Niv. Humans combine value learning and hypothesis testing strategically in multi-dimensional probabilistic reward learning.

All data and code are available at `https://github.com/mingyus/`.

4.1 Introduction

Learning in a complex environment, with numerous potentially relevant factors and noisy outcomes, can be quite challenging. For example, when learning to make bread, a collection of decisions needs to be made, including the amount of yeast to use, the flour-to-water ratio, the proof time, and the baking temperature. An inexperienced baker can be clueless when facing these decisions, especially when the results are variable even if following the same procedure: the ambient temperature may affect rising, the oven temperature may not be as accurate as its marks, etc., making feedback unreliable.

Learning scenarios like this are quite common in real life, but have not been studied systematically. In controlled, laboratory conditions, each of the key components of such learning has traditionally been investigated separately. Decisions based on combining multiple factors (features) are common in category learning tasks [103, 104] where multi-dimensional rules determine the category judgements, although feedback is often deterministic in these tasks. In contrast, the need to integrate and learn from stochastic feedback has been widely studied in probabilistic learning tasks [105, 106, 107], but often with a simplistic rule that involves only one relevant feature dimension. Finally, the freedom to determine learning examples and try out different possibilities (rather than select among a few available options) is at the core of active learning tasks. As very few tasks have combined all these components (but see [108, 109]), it remains unclear how people learn actively in an environment with complex rules (with multiple and potentially an unknown number of relevant dimensions) and probabilistic feedback. To study this, we developed a novel decision task: participants were asked to configure three-dimensional stimuli by choosing what features to use in each dimension, earning rewards that were probabilistically determined by features in a subset or all of these dimensions. To earn as much reward as possible, participants

needed to figure out which dimensions were important through trial-and-error, and learn what specific features yielded rewarding outcomes in those dimensions.

Despite the computational challenge and combinatorial explosion of possible solutions, human beings are remarkably good at solving such complex tasks. Usually, after a few successful or unsuccessful attempts, the amateur baker will gradually figure out the rules for bread-making. Similarly, participants in above task improved their performance over time, and learned to correctly identify rewarding features through experience. To understand how they achieved this, we turned to the extensive literature regarding two systems that support representation learning [110, 111]: a rule-based system that explicitly represents and evaluates hypotheses, and a value-based reinforcement-learning system that incrementally learns the value of stimuli based on trial-and-error feedback. In previous studies, the two mechanisms were often examined separately, as which of them is used often depends on the specific task. For instance, in probabilistic reward learning tasks, people have been shown to learn through trial-and-error to identify relevant dimensions, and gradually focus their attention onto the rewarding features in those dimensions [105, 106, 107]; in contrast, in category learning, people seem to evaluate the probability of all possible rules via Bayesian inference, with a prior belief favoring simpler rules [104]. However, the two learning systems are likely simultaneously engaged in most tasks [112], and contribute to different extents depending on how efficient they are in each specific setting. Direct hypothesis-testing can be more efficient when fewer hypotheses are likely and when feedback is relatively deterministic, whereas incremental learning may be more beneficial with numerous possible combinations and stochastic outcomes.

Here, we systematically examined the integration of the two learning systems and how it depends on task condition. Specifically, we varied task complexity by setting the rules such that one, two, or all three dimensions of the stimuli were relevant for obtaining reward; in addition, we manipulated whether such information (i.e.,

rule dimensionality) was explicitly provided to participants. We fit computational models that represent each learning system (and their hybridization) to participants' responses, and compared how well they predicted participants' choices. We found evidence that people used both learning systems when solving our task, across all task conditions. Furthermore, when participants were informed of the task complexity, they used this information to set the balance between the two systems, relying more on serial hypothesis testing when the task was simpler with fewer candidate rules, and more on reinforcement learning when more rules were possible. Our findings shed light on how the rule-based and value-based systems cooperate to support representation learning in complex and stochastic scenarios, and suggest that humans can evaluate their effectiveness based on task complexity and make strategic arbitration between them.

4.2 Experiment and behavior results

4.2.1 The "build icon" task

In our task, stimuli were characterized by features in three dimensions: color (red, green, blue), shape (square, circle, triangle) and texture (plaid, dots, waves). In each of a series of games, a subset of the three dimensions was relevant for reward, meaning that one feature in each of these relevant dimensions would render stimuli more rewarding (henceforth the "rewarding feature").

To earn rewards and figure out the underlying rule, participants were asked to configure stimuli ("icons") by selecting features for any (zero to all) of the dimensions (Figure 4.1); for dimensions in which they did not make a selection, the computer would randomly select a feature. The resulting stimulus was then shown on the screen, and the participant would receive probabilistic reward feedback (one or zero points) based on the stimulus: the more rewarding features included in the stimulus,

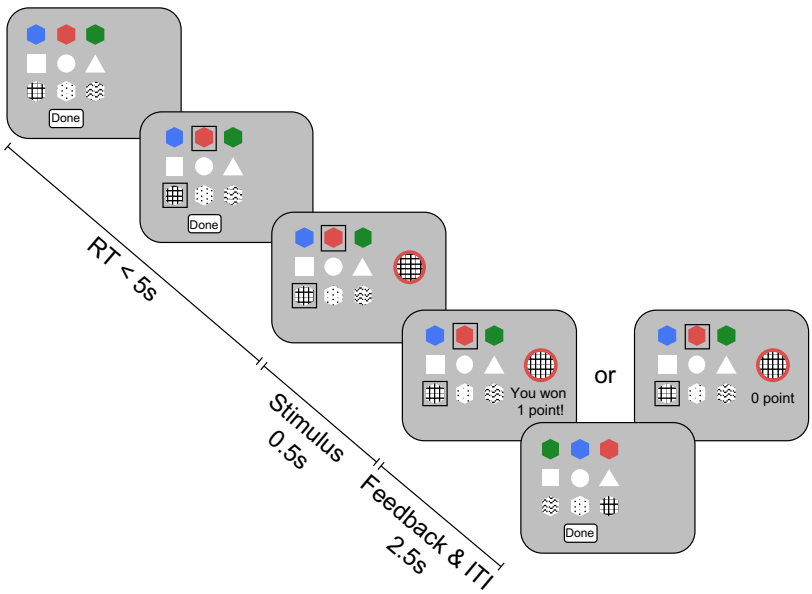

Figure 4.1: **The build-icon task.** Participants built stimuli by selecting a feature in any (zero to three) of the three dimensions (marked by black squares). After hitting "Done", the stimulus showed up on the screen, with features randomly determined for any dimension in which participant did not make a selection (in this example, circle was randomly determined). Reward feedback was then shown.

Table 4.1: The reward probability of a stimulus in each game type (1D, 2D, and 3D-relevant games) was determined by the number of rewarding features in the stimulus.

Game type	Number of rewarding features			
	0	1	2	3
1D-relevant	20 %	80%	–	–
2D-relevant	20 %	50%	80%	–
3D-relevant	20 %	40%	60%	80%

the higher the reward probability, with the lowest reward probability being $p = 0.2$ and the highest $p = 0.8$ (see Table 4.2.1). The participant's goal was to earn as many reward points as possible.

Each game had one, two, or three relevant dimensions (henceforth 1D-, 2D-, and 3D-relevant conditions). This information was provided to participants in half of the games ("known" condition) and the other half was designated as "unknown" games.

This resulted in six game types in total. Each participant played three games of each type for a total of 18 games, in a randomized order. Each game was comprised of 30 trials. The relevant dimensions and rewarding features changed between games.

4.2.2 Participants and procedure

Participants were recruited online from Amazon Mechanical Turk. They received a base payment of $12 for completing the task, with a performance-based bonus of $0.15 per reward point earned in three randomly-chosen games (one for each task complexity).

Participants went through a comprehensive instruction phase before starting the real experiment. During the instruction, they were first introduced to the "icons", and asked to build a few examples. They were then explained the general rules of the experiment, including the complexity levels and their respective reward probabilities (as in Table 4.2.1). They were tested about these rules and probabilities with a set of multiple-choice questions. For each task complexity, they were given an example rule, and asked about the reward probability of a few stimuli to test their understanding. In addition, they did a practice game per each complexity with the rules informed. For the understanding tests, participants had to answer all questions correctly, within a few attempts (three times for most questions, with the exception of five times for questions on the general rules), in order to proceed to the real games. During the real games, participants were required to respond within 5 seconds on each trial. Those who missed five trials consecutively were stopped from continuing the experiment.

After the instruction phase, the main experiment commenced. In "known" games, the number of relevant dimensions was instructed before the start of the game in the form of a "hint"; participants were, however, never told which dimensions were relevant or which features were more rewarding. The start of "unknown" games was also signaled; however, no hint was provided in these games. At the end of each game,

participants were asked to explicitly report, to their best knowledge, the rewarding feature for each dimension, or indicate that this dimensions is irrelevant to reward, as well as their confidence level (0-100) in these judgements. After the experiment, participants received a performance bonus proportional to the points they earned in three randomly-selected games.

106 participants completed the entire experiment, out of which 4 were excluded from our analyses due to poor performance: an overall reward probability of less than 0.468 (two standard deviation lower than the group average).

4.2.3 Learning performance and choice behavior

Across all six game types, participants' performance improved over the course of a game, with faster learning in less complex games (games with fewer relevant dimensions) (Figure 4.2A; a three-way repeated measures ANOVA on reward probability found significant effects including main effects of trial index: $F_{29,5858} = 329.5, p < .001$, and task complexity: $F_{2,5858} = 206.111, p < .001$, the three-way interaction: $F_{58,5858} = 1.946, p < .001$, and two two-way interactions, between trial index and task complexity: $F_{58,5858} = 20.6, p < .001$, and between task complexity and task instruction knowledge: $F_{2,5858} = 6.99, p = .00127$). The overall worse performance in more complex games was not necessarily a failure of learning, but rather the result of limited experience (only 30 trials per game), as participants' average reward rate across all games was 90.2% of that of an (approximate) optimal agent[1] playing this same task. Between the "known" and "unknown" games, participants' performance was better when informed of the game complexity in 3D-relevant games (a significant main effect of task instruction in a two-way repeated measures ANOVA on reward probability for 3D-relevant games only: $F_{1,101} = 11.3, p = .001$, uncorrected, same for

[1]It is computationally intractable to solve the optimal policy for this task. Therefore we trained a DQN network [17] on the task to approximate the optimal solution, and compared participants' performance with this well-trained DQN agent.

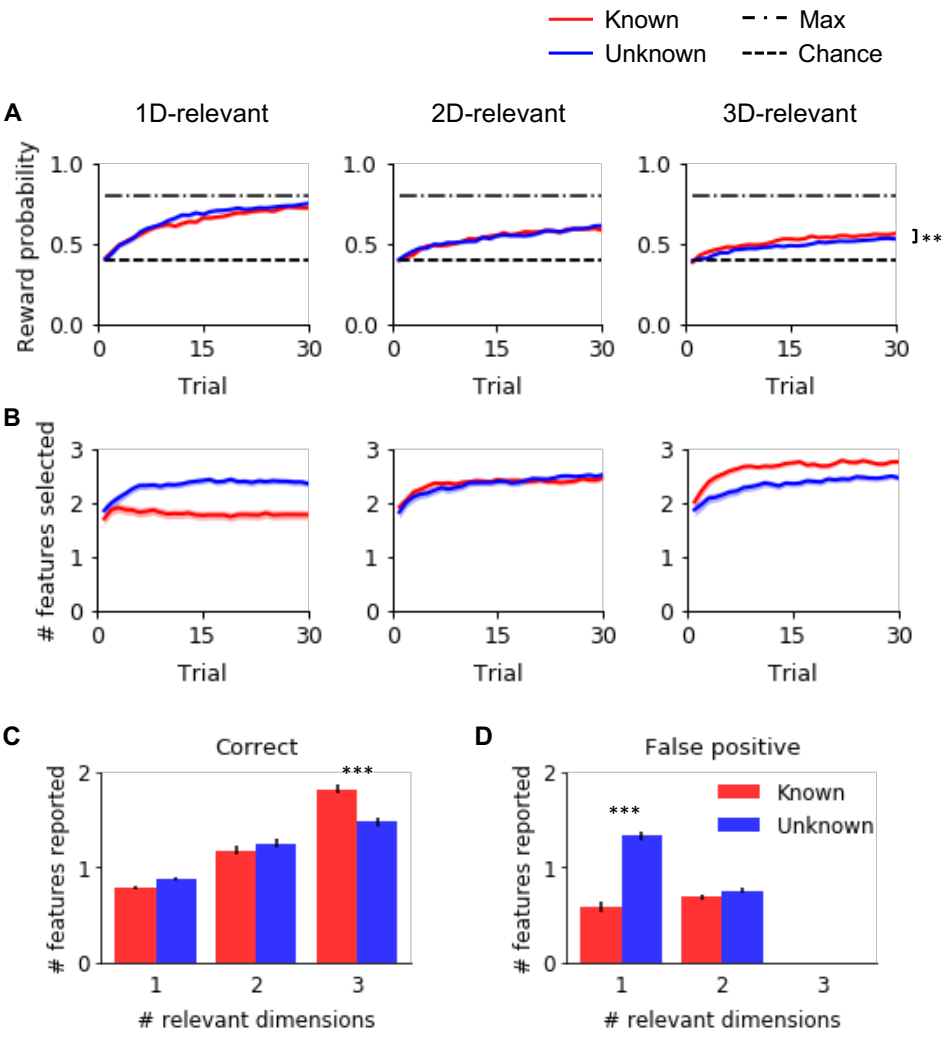

Figure 4.2: **Participants' behavior in the "build-icon" task. (A, B): Performance and choices over the course of a game, by game type.** **(A)** Participants' average reward probability (calculated based on the number of rewarding features in their configured stimulus), over the course of 1D-, 2D- and 3D-relevant games (left, middle and right columns). Red and blue curves represent "known" and "unknown" conditions, respectively. Shading represents 1 s.e.m. across participants. Dash-dotted lines represent the maximum reward probability in principle ($p = 0.8$); dashed lines represent the chance level. **(B)** Same as in (A), but for the number of features selected. **(C, D): Responses to post-game questions regarding the rewarding features in each game condition.** **(C)** Average number of correctly-identified rewarding features; **(D)** Average number of false positive responses, i.e., falsely identifying irrelevant dimension as relevant. *** $p < .001$.

tests below); there was no effect of task instruction on performance in simpler games (1D-relevant: $F_{1,101} = 3.28, p = .073$; 2D-relevant: $F_{1,101} = 0.0007, p = .98$).

Participants also showed distinct choice behavior in different game types (Figure 4.2B): a three-way repeated measures ANOVA on the number of features selected found significant main effects of trial index ($F_{29,5858} = 26.9, p < .001$), task complexity ($F_{2,5858} = 95.2, p < .001$)), and task instruction knowledge ($F_{1,5858} = 11.8, p = .004$)), a significant three-way interaction effect ($F_{58,5858} = 5.32, p < .001$), and two-way interaction effects for all pairs of variables (all $p < .001$). In "known" games, the more dimensions participants were informed to be relevant, the more features they chose on each trial on average (mixed-effects linear regression slope: 0.29 ± 0.03, $p < .001$); in "unknown" games, unsurprisingly, the number of selected features was not different between game types ($p = .47$).

We then analyzed participants' responses to the post-game questions (Figure 4.2C,D; see full results in Figure 4.7). A two-way repeated measures ANOVA on correct responses (i.e., correctly-identified rewarding features) found a significant main effect of task complexity ($F_{2,202} = 273.7, p < .001$), and a significant interaction effect ($F_{2,202} = 21.3, p < .001$). A similar ANOVA on false positive responses (i.e., the number of irrelevant dimensions falsely identified as relevant) found significant main effects of both task complexity ($F_{1,101} = 32.0, p < .001$) and task instruction knowledge ($F_{1,101} = 93.3, p < .001$), and a significant interaction effect ($F_{1,101} = 90.8, p < .001$). Specifically, in 1D-relevant games, participants correctly identified a similar number of rewarding features between "known" and "unknown" conditions (Figure 4.2C; post hoc Tukey test: $t_{101} = 1.81, p = .46$), consistent with the choice behavior in Figure 4.2A. They were, however, more likely to falsely identify an irrelevant feature as relevant (4.2D; $t_{101} = -6.27, p < .001$), indicating that not knowing the dimensionality of the underlying rule led the participants to incorrectly attribute rewards to irrelevant features, which might be the reason of a higher number of features selected in the

"unknown" condition (Figure 4.2B). In contrast, in 3D-relevant games, participants reported to have identified more correct features in the "known" condition than in "unknown" condition (Figure 4.2C; $t_{101} = 13.53, p < .001$), consistent with the better learning performance in "known" 3D-relevant games observed in Figure 4.2A.

4.3 Computational modeling

4.3.1 Two learning systems

To characterize participants' learning strategy and explain the behavioral differences between game conditions, we considered two candidate learning systems [111, 112]: an incremental value-based system that learns the value of stimuli based on trial-and-error feedback, and a rule-based system that explicitly represents possible rules and evaluates them. We tested computational models representing each of these two systems, as well as a hybrid combination, by fitting each model to participants' trial-by-trial choices and comparing how well they predict task behavior.

We first briefly describe each model in this section, and provide the detailed equations in the next section.

The value-based system was captured by a feature-based reinforcement learning model [105]. Reinforcement learning is commonly used to model behavior in probabilistic reward-learning tasks, where participants need to accumulate evidence across multiple trials to estimate the value of each choice. In particular, we used the **feature RL with decay model** from prior work [105] with a task similar to ours. This model assumes that participants learn the values of the nine features using the Rescorla-Wagner update rule: feature values in the current stimulus are updated proportional to the reward prediction error (the difference between the outcome and the expected value). In addition, values of features not present on the current stimulus decay towards zero. The expected reward for each choice (i.e., combination of fea-

tures selected) is calculated as the sum of its feature values. At decision time, the probability of each choice is determined by comparing the expected reward for all choices using a softmax function.

In contrast to the value-based strategy, the rule-based strategy directly evaluates hypotheses regarding what combinations of features are relevant for obtaining more rewards (the set of rewarding features) in a game, which we refer to as "rules". In "known" games, there are 9, 27 and 27 possible rules for 1D-, 2D- and 3D-relevant games, respectively; in "unknown" games, all 63 rules are possible.

There are multiple possibilities for how people learn the correct rules. One is to use the Bayesian principle to evaluate the probability that each rule is the correct one; we term this a **Bayesian rule learning model**. After each outcome, this model optimally utilizes feedback information to calculate the likelihood of each candidate rules, and combines this with the prior belief of the probability that each rule is correct (initially assumed to be uniform across all rules that accord with the "hint") to obtain the posterior probabilities of each rule. The expected reward for a choice is then calculated by marginalizing over the posterior belief of all possible rules, and the final choice probability was determined by a softmax function over the expected reward from each choice (as in the previous model). In a multi-dimensional category learning task, a similar Bayesian rule learning model has been shown to characterize how people learn categories better than reinforcement learning models [104].

Bayesian inference is computationally expensive and memory-intensive. A simpler alternative for the rule-base strategy is serial hypothesis testing, which assumes that people only test one rule at a time: if the evidence supports their hypothesis, they will continue with it; otherwise, they switch to a different rule, until the correct one is found. The idea of serial hypothesis testing has long roots in category learning literature [113, 114]. Recently, it has also been applied in probabilistic reward learning tasks [115] and shown to be a better account of human behavior than the

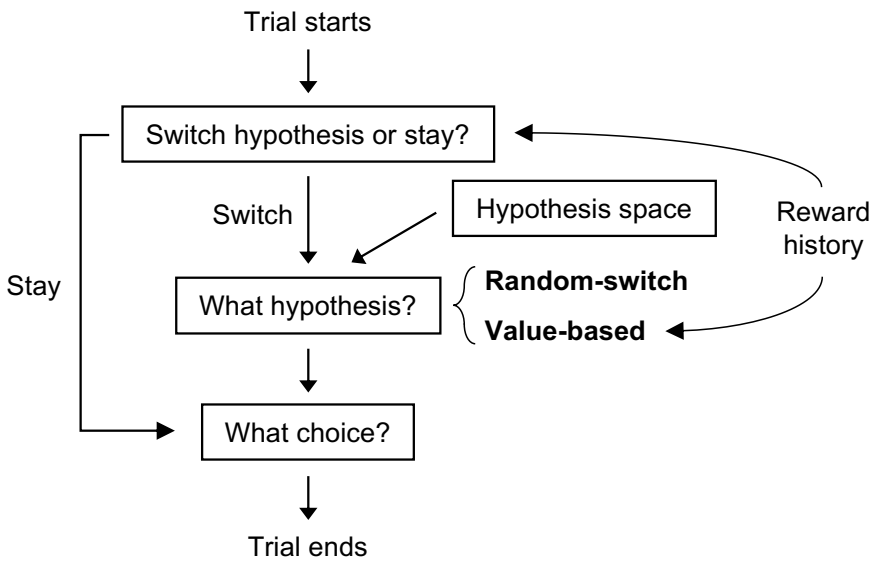

Figure 4.3: **A diagram of the serial hypothesis testing models.**

Bayesian model. Following [115], we consider a **random-switch serial hypothesis-testing model** (random-switch SHT model; Figure 4.3): it assumes that people test hypotheses about the underlying rule one at a time. When testing a hypothesis, they estimate its reward probability by counting how often they get rewarded when making choices accordingly. The lower this estimate, the more likely they will abandon the current hypothesis and switch to testing a random different one. We assume that people's choices are often consistent with their hypotheses, but with a small ($p = \lambda$) probability, they lapse and make random choices.

The SHT and RL systems are not necessarily mutually exclusive. We thus also considered a hybrid model by incorporating RL-acquired feature values into the choice of a new hypothesis in the serial hypothesis testing model. In particular, when switching hypotheses, the hybrid model favors hypotheses that contain recently rewarded features. We term this model **value-based serial hypothesis testing model** (value-based SHT model; Figure 4.3).

4.3.2 Computational models

Feature-based reinforcement learning with decay model

The feature RL with decay model learns the values of nine features (denoted by $f_{i,j}$; i and j are indices for dimensions and features respectively) using the Rescorla-Wagner update rule, with separate learning rates for features that were selected by the participant ($\eta = \eta_s$) and those that were randomly determined ($\eta = \eta_r$). Values for features not in the current stimulus s_t are decayed towards zero with a factor $d \in [0, 1]$. η_s, η_r and d are free parameters.

$$V_t(f_{i,j}) = \begin{cases} V_{t-1}(f_{i,j}) + \eta(r_t - ER(c_t)), \text{ if } j = s_t^i \\ d \cdot V_{t-1}(f_{i,j}), \text{ if } j \neq s_t^i \end{cases} \tag{4.1}$$

where r_t is the reward outcome (0 or 1) on trial t, and s_t^i indicates the feature on dimension i of s_t.

At decision time, the expected reward (ER) for each choice c is calculated as the sum of its feature values, with c^i denoting the feature on dimension i of choice c:

$$ER(c) = \sum_i V(f_{i,c^i}), \tag{4.2}$$

The average value of all three features is used for dimensions with no selected features.

The choice probability is then determined based on $ER(c)$ using a softmax function, with β as a free parameter:

$$P(c) = \frac{e^{\beta \cdot ER(c)}}{\sum_{c'} e^{\beta \cdot ER(c')}}. \tag{4.3}$$

Bayesian rule learning model

The Bayesian rule learning model maintains a probabilistic belief distribution over all possible hypotheses (denoted by h). After each trial, the belief distribution is updated according to Bayes' rule:

$$P(h|c_{1:t}, r_{1:t}) \propto P(r_t|h, c_t)P(h|c_{1:t-1}, r_{1:t-1}).$$

At decision time, the expected reward for each choice is calculated by marginalizing over the belief distribution:

$$ER(c) = \sum_h P(h)P(r|h, c).$$

The expected reward is then used to determine the choice probability as in Equation 4.3.

We note that this model is not strictly optimal, even with no decision noise, as it maximizes the reward on the current trial, but not the total reward over a game.

Random-switch serial hypothesis testing (SHT) model

The random-switch SHT model assumes the participant tests one hypothesis at any given time. We do not directly observe what hypothesis the participant is testing, and need to infer that from their choices. We do so by using the change point detection model in [115]. The basic idea is to infer the current hypothesis (denoted by h_t) from all the choices the participant has made and the reward outcomes they received so far in the current game (together denoted by $D_{1:t-1}$); see Supplementary Methods for implementation details. Once we obtain the posterior probability distribution over

the current hypothesis $(h_t|D_{1:t-1})$, we can then use it to predict choice:

$$P(c_t|D_{1:t-1}) = \sum_{h_t} P(c_t|h_t)P(h_t|D_{1:t-1})$$

In order to calculate $P(h_t|D_{1:t-1})$, we consider the generative model of participant's choices. First, we determine the participant's hypothesis space: In "known" games, participants were informed about the number of relevant dimensions, which limits the set of possible hypotheses in these game. The way people interpret and follow instructions, however, may vary. Thus, we parameterize the hypothesis space (i.e., people's prior over all possible hypotheses) with two weight parameters w_l and w_h (before normalization):

$$P(h) \propto \begin{cases} w_l & \text{if } D(h) < D \\ 1 & \text{if } D(h) = D \\ w_h & \text{if } D(h) > D \end{cases} \tag{4.4}$$

Here, $D(h)$ is the dimensionality of hypothesis h (how many rewarding features are in h), and D is the number of relevant dimensions of the current game. If a participant strictly follows the instruction, $w_l = w_h = 0$, i.e., only hypothesis with the same dimensionality as the instruction is considered to be possible; if they do not use the instruction information at all, $w_l = w_h = 1$, i.e., all 63 hypotheses are considered to be equally likely. For "unknown" games, the average $P(h)$ of 1D, 2D and 3D "known" games is used.

The generative model of participant's choice behavior contains three parts: the hypothesis-testing policy (whether to switch hypotheses or stay), the hypothesis-switch policies (what the next hypothesis is), and the choice policy. The first two policies together determine the transition from the hypothesis on the last trial to

the current one, and the choice policy determines the mapping between the current hypothesis and choice.

Following [115], we consider the following hypothesis testing policy: on each trial, the participant estimates the reward probability of the current hypothesis. Using a uniform Dirichlet prior, this is equivalent to counting how many times they have been rewarded since they started testing this hypothesis. The estimated reward probability is then compared to a soft threshold θ to determine whether to stay with this hypothesis or to switch to a different one:

$$Pr(\text{stay}) = \frac{1}{1 + e^{-\beta_{\text{stay}}\left(\hat{P}_{\text{reward}} - \theta\right)}}, \tag{4.5}$$

where $\hat{P}_{\text{reward}} = \frac{\text{reward count } +1}{\text{trial count}+2}$ is the estimated reward probability, and β_{stay} and θ are free parameters. If the participant decides to switch, they randomly switching to any other hypothesis according to the prior over hypotheses specified in Equation 4.4 (i.e. the random hypothesis-switch policy):

$$P(h_t) = \begin{cases} Pr(\text{stay}), \text{ if } h_t = h_{t-1} \\ (1 - Pr(\text{stay})) \frac{P(h_t)}{\sum_{h \neq h_{t-1}} P(h)}, \text{ if } h_t \neq h_{t-1} \end{cases} \tag{4.6}$$

We use epsilon-greedy as the choice policy: participants' choices are assumed to be aligned with their hypotheses most of the time, with a free-parameter lapse rate of λ.

Value-based serial hypothesis testing model

The value-based SHT model is the same as random-switch SHT model, except for using a value-based hypothesis-switch policy. It maintains a set of feature values updated according to feature RL with decay model as in Equation 4.1 (but with a single learning rate), and calculates the expected reward for each alternative hypothesis by adding up its feature values, similar to Equation 4.2 but for h instead of c. The

probability of switching to $h_t \neq h_{t-1}$ is:

$$P(h_t) = (1 - Pr(\text{stay})) \frac{e^{\beta_{\text{switch}} \cdot ER(h_t)}}{\sum_{h' \neq h_{t-1}} e^{\beta_{\text{switch}} \cdot ER(h')}}, \tag{4.7}$$

where β_{switch} is a free parameter.

4.3.3 Model fitting and model comparison

We fitted the models to each participant's data using maximum likelihood estimation. We used the minimize function (L-BFGS-B algorithm) in Python package scipy.optimize as the optimizer; each optimization was repeated for 10 times with random starting points. Models were evaluated with leave-one-game-out cross-validation: the likelihood of each game was calculated using the parameters obtained from fitting the other 17 games; the average likelihood per trial across all games was reported.

4.3.4 Evidence for both learning systems

The model comparison results are shown in Figure 4.4A. Among all four models, the Bayesian rule learning model, even though optimal in utilizing the feedback information, showed the worst fit to participants' choices (likelihood per trial: 0.045 ± 0.003; mean $\pm$ s.e.m.). This was potentially because the large hypothesis space (up to 63 hypotheses) made it implausible for participants to perform exact Bayesian inference. Both the feature RL with decay model and the random-switch SHT model showed better fits (likelihood per trial: 0.118 ± 0.008 and 0.160 ± 0.009, respectively). Compared to the Bayesian model, both models require lower computation and memory load: the RL model learns nine feature values individually and later combines them; the random-switch SHT model limits the consideration of hypotheses to one at a time. The hybrid value-based SHT model fit the data best (better than either component

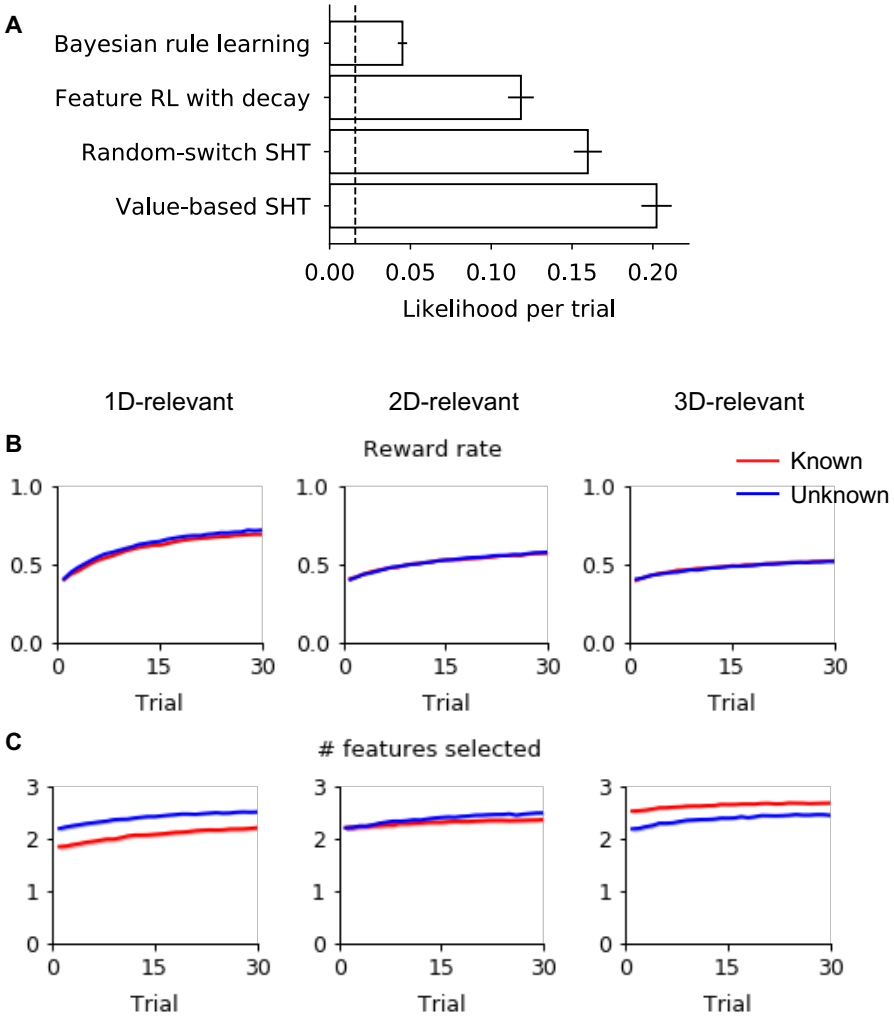

Figure 4.4: **Model comparison supports both reinforcement learning and serial hypothesis testing strategies. (A)** Geometric average likelihood per trial for each model (i.e., average total log likelihood divided by number of trials and exponentiated). Higher values indicate better model fits. Dashed lines indicate the chance level. **(B, C)** Simulation of the best-fitting value-based SHT model. The same learning curves as in Figure 4.2 but for model simulation.

model; likelihood per trial: 0.202 ± 0.009), suggesting that participants used both learning strategies when solving this task.

There was additional evidence for the involvement of both learning systems in participants' behavior. The rule-based system was evident from the influence of task instructions: both the numbers of features selected (Figure 4.2B) and the reported rewarding features in the post-game questions (Figure 4.2C,D) differed between "known" and "unknown" conditions. There is no direct way to incorporate such influences in a reinforcement learning model, but a rule-learning model can easily do so, for instance, by constraining the hypothesis spaces according to the instructions. On the other hand, the influence of value-based learning was evident in the order in which participants clicked on features to make selections. In most cases, participants followed the spatial order in which dimensions appeared on the screen, either top-to-bottom or the reverse. When the clicks violated the spatial orders, however, they followed the order of learned feature values, starting from the most valuable feature[2], at a frequency significantly above chance ($t_{101} = 7.63, p < .001$). Such behavior of following the order of learned feature values instead of the spatial order was more frequent in trials when participants switched hypotheses than when they continued testing the same hypothesis ($t_{101} = 5.71, p < .001$; in this analysis, switch trials were identified based on changes in choice, for simplicity), further supporting the value-based SHT model.

In sum, participants' strategies in this task could not be explained by either reinforcement learning or serial hypothesis testing strategies alone. The combined hybrid model explained participants' behavior best, also capturing the dependence of performance on task complexity (Figure 4.4B) and the qualitative differences between choice curves in "known" and "unknown" conditions (Figure 4.4C), which neither component model could capture (Figure 4.6).

[2]This result held regardless of which model we used to calculate value, the feature RL with decay model or the hybrid model.

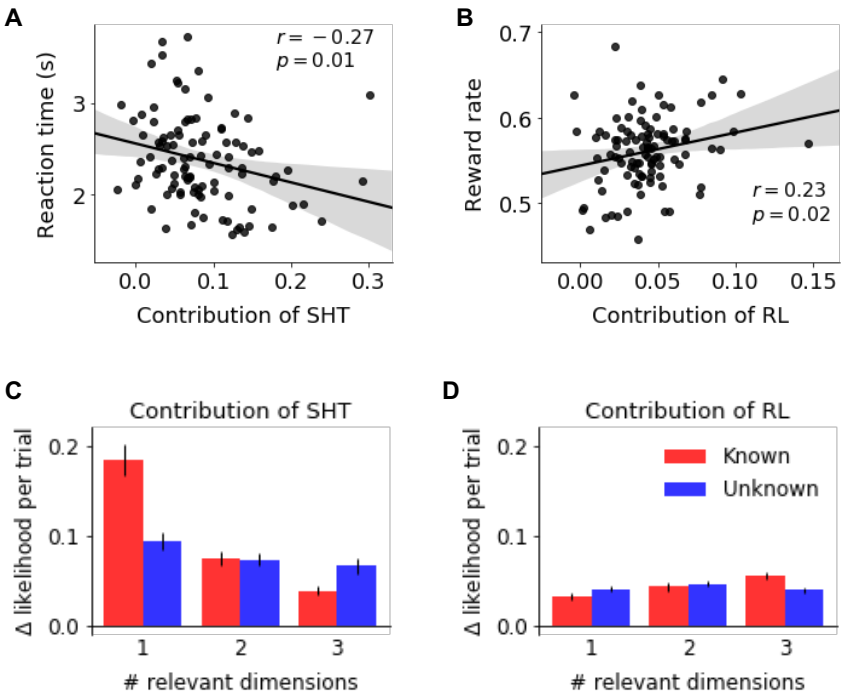

Figure 4.5: **Strategic balance of two learning systems. (A)** The contribution of serial hypothesis testing (SHT) was inversely correlated with reaction time such that participants who responded faster used SHT to a greater extent. **(B)** The contribution of reinforcement learning (RL) was correlated with average reward rate – participants for whom adding RL in the hybrid model more greatly improved their model fits earned more rewards on the task, on average. Each dot represents one participant. **(C, D)** Contribution of RL and SHT for each game type. Error bars represent 1 s.e.m. across participants. The contribution of each component is measured as the difference in likelihood per trial between the hybrid value-based SHT model and the other component model (SHT: the feature RL with decay model; RL: the random-switch SHT model).

4.3.5 The contribution of the two systems depends on task complexity

Given evidence that participants used both learning strategies in this task, the next question is to what extent each system contributed to decision making. We addressed this question by comparing the hybrid model with the two component models: the difference in likelihood per trial between the hybrid model and each component model

was taken as a proxy for the contribution of the mechanism not included in the component model.

Across all participants, a higher contribution of SHT was associated with a faster reaction time (Figure 4.5A; $r = -0.27, p = .01$), and a higher contribution of RL was associated with a higher reward rate (Figure 4.5B; $r = 0.23, p = .02$); the other two correlations (between reaction time and RL, and between reward rate and SHT) were not significant (both $p > .1$). These results suggest that, comparatively, serial hypothesis testing was an overall faster and less effortful strategy; although reinforcement learning may be slower, augmenting hypothesis testing with values yielded more reward.

To optimize for reward and reduce mental effort costs, it is advantageous to rely on the serial hypothesis testing strategy when the task is simpler, for instance, in lower-dimensional games with smaller hypothesis spaces. Indeed, when tested separately, the correlation between reward rate and contribution of RL was only significant for 2D- and 3D-relevant games (1D: $r = -0.03, p = .75$; 2D: $r = 0.27, p < .01$; 3D: $r = 0.32, p < .01$; uncorrected). In contrast, with a larger hypothesis space, serial hypothesis testing is less efficient, and there should be a higher incentive to use the value learning strategy.

We indeed observed such a strategic trade-off between the two learning systems: in "known" games, the contribution of hypothesis testing decreased as the dimensionality of the task increased (Figure 4.5C; estimated slope in a mixed-effect linear regression: -0.0631 ± 0.0051, $p < .001$), whereas the contribution of value learning increased with task complexity (Figure 4.5D; estimated slope: 0.0178 ± 0.0013, $p < .001$). In contrast, in "unknown" games, in which task complexity information was unavailable to participants, the contribution of the two mechanisms was more stable across game conditions (estimated slopes: -0.0144 ± 0.0042 for SHT, $p < .001$; -0.0011 ± 0.0012 for RL, $p = .389$; a significant three-way interaction effect

in a repeated measures ANOVA on likelihood difference per trial in Figure 4.5C,D: $F(2, 202) = 47.9, p < .001$). Taken together, these results suggest that participants took advantage of information regarding task complexity to strategically balance use of two complementary learning mechanisms.

4.4 Discussion

Using a novel "build-icon" task, we studied learning of multi-dimensional rules with probabilistic feedback as a proxy for real-world learning in situations where it is unknown *a priori* what aspects of the task are relevant to solving it, and where learners have agency to intervene on the environment and test hypotheses. In our task, participants created stimuli and tried to earn more rewards by identifying the predetermined rewarding features. Participants performed this task at various known or unknown complexity levels (i.e., rewarding features on one, two or three stimulus dimensions). They demonstrated learning in all conditions, with their performance and strategies influenced by task condition. Through behavioral analyses and computational modeling, we investigated the use of two distinct but complementary learning mechanisms: serial hypothesis testing that evaluates one possible rule at a time and is therefore simple and fast in response, but results in slow learning when many rules are possible and must be tested sequentially; reinforcement learning that learns about all features in parallel and is more accurate in the long run, but requires maintaining and updating more information. We found that a hybrid model that incorporated the advantages of both mechanisms explained participants' behavior the best. In addition, we showed that human participants used knowledge of task complexity to gauge which mechanism is more suitable, demonstrating a strategic balance between the two. Specifically, they tended to use the simpler and faster serial hypothesis testing strategy when they knew that fewer dimensions matter in the decision, but

relied more on incrementally learning feature values when multiple dimensions were important.

The current study connects large bodies of work on reward learning and category learning in multi-dimensional environments. Previous studies have extensively investigated how humans learn about complex but deterministic categorization rules [111, 103, 104], as well as how they learn through trial-and-error to identify a single relevant dimension [105, 116, 117, 118]. The former type of tasks are hard to learn because of the unknown form of the underlying rules, while the latter tasks focus on how humans integrate information in stochastic environments. Both are common challenges for human decision-making, and they often co-occur in daily tasks – in new situations, we often do not know *a priori* what aspects of the task are relevant to its correct solution, and feedback may be stochastic, not only due to task properties, but also caused by the decision maker's poor control over task factors in the beginning. Therefore, we imposed both challenges in the current task to investigate human learning strategies under these complex scenarios. Our work also helps unite the various findings on value-based or rule-based strategies in previous studies. We show that learning in complex and stochastic environments engages both systems. In fact, participants' strategy lies on a spectrum, with flexible arbitration between the two systems based on which is more efficient under current task condition. This can potentially explain why value-based strategies are often observed in probabilistic learning tasks [105, 106, 107], and rule-based strategies often in category learning tasks [104].

A few studies have pursued a similar path. For example, [108] studied a similar probabilistic reward learning task with multiple relevant dimensions. They tested hypothesis-testing strategies based on values learned with naïve RL models. Through model comparison, they showed that values learned alongside hypothesis testing were carried over when hypotheses switched, consistent with our value-based SHT model.

The novelty of our work is in systematically manipulating the complexity of the environment and people's knowledge about it, to help provided a comprehensive understanding on how people's learning strategy adapts to different situations.

Still, we considered only a simple linear combination of multiple dimensions to determine reward: each relevant dimension contributed equally to reward probability, in an additive manner. In real life, the composition can be more complex, with unequal weights for different dimensions [118, 119], and potential interactions between dimensions. We postulate that similar hybrid strategies will be adopted regardless. However, it can be hard to model the hypothesis-testing strategy in such scenarios, due to the much larger and potentially ambiguous hypothesis space. An important question is how do people construct their hypothesis space, and how likely do they deem each hypothesis *a priori*. People may favor simpler hypotheses [120]; they may not have a fixed hypothesis space to begin with, but construct new hypotheses only when the existing ones can no longer account for observations [52], or they may modify their existing hypotheses on the go with small changes [121].

It is worth noting the unique free-configuration design of the current task. In most representation-learning tasks, stimuli (i.e., the combination of features) are predetermined, and participants are asked to select between several available options, or make categorization judgements. These tasks are easy to perform, but it is hard to isolate participants' preference over single features. Our task enabled us to directly probe people's preference (or lack thereof) in each of the three dimensions. In addition, we were able to hold baseline reward probability constant across different game types (participants responding randomly would always earn reward with $p = 0.4$) while varying the complexity of underlying rules, which avoided providing additional information on rule complexity in "unknown" games due to the baseline reward rate. This would have been hard to achieve with the more commonly used alternative-choice design. The free-configuration task also resembles many real-life

decisions where choices across multiple dimensions have to be made voluntarily, from ordering a pizza takeout, to planning a weekend getaway trip.

Along with these advantages, the active-learning free-configuration design may also alter the strategy people use, compared to a passive learning scenario. On the one hand, free-choice may encourage hypothesis testing, making this strategy more efficient by allowing participants to seek direct evidence on their hypotheses. On the other hand, learning may be hindered due to confirmation bias, commonly observed in self-directed rule-learning tasks (aka "positive test strategy" [122]). Indeed, participants over-estimated the number of rewarding features in 1D "unknown" games as compared to "known games" (Figure 4.2D), suggesting that they failed to prune their hypotheses when the underlying rule was simpler. To fully understand the impact of free choice, future work can compare active and passive settings with a "yoked" design. This can help understand whether the findings reported here can be generalized to passive-learning tasks, and what may be unique to the active-learning setting.

To model the integration of the two learning strategies, we introduced the hybrid value-based SHT model. The assumptions in this model are relatively minimal, which can be a reason why the hybrid model failed to quantitatively predict the number of features selected by participants (Figure 4.4C). We explored several alternatives for the model assumptions (Figure 4.8; see Methods for details): (1) not always testing a hypothesis: if none of the hypotheses is high in value, the participant can decide not to test a hypothesis, and let the computer configure a completely random stimulus instead; (2) flexible threshold for determining whether to switch hypothesis or not, based on reward probability of the corresponding game condition (Table 4.2.1); (3) favoring choices that are supersets of the current hypothesis: rather than designing stimuli consistent with the current hypothesis (with a lapse rate), participants may tend to select more features than what their hypothesis contains (that is, they tend to select features on dimensions not specified by the current hypothesis). The first

and third alternative assumptions improved model fits, but the second did not. We then considered a "full" model that used the better alternative for each assumption. This more complex model improved average likelihood per trial on holdout games by 0.033 ± 0.006, which is a significant improvement. In terms of capturing the dependency of choices and performance on game condition, however, this model behaved similarly to the original hybrid model (Figure 4.6): both models under-predicted the differences in the number of selected features between "known" and "unknown" conditions, compared to the empirical data. For simplicity, we therefore reported the original hybrid model in the Results.

The flexibility of the value-based SHT model opens up the space for exploring more complex hypothesis-testing strategies. For instance, hypotheses may be formed in a hierarchical manner when the rule complexity is unknown, i.e., participants may first reason about the dimensionality of the game, and then the exact rule. Currently, the hypothesis-switching policy depends only on values, but the complexity of the rule may also play a role: participants may start from simpler rules, and switch to more complex rules, as suggested in the SUSTAIN model [123]. Another promising direction is to test multiple hypotheses in parallel. In the current model, only one hypothesis is tested at a time, yet participants may consider multiple possibilities simultaneously, adding and removing hypotheses flexibly. Last, the current model assumes that learning of feature values happens in parallel to and independently of hypothesis-testing; however, value learning may also be affected by hypothesis testing, for example, the amount of value update can be gated by the current hypothesis [124, 112]. The current modeling framework (and openly accessible data) can be used in future work to systematically examine these and other alternative models.

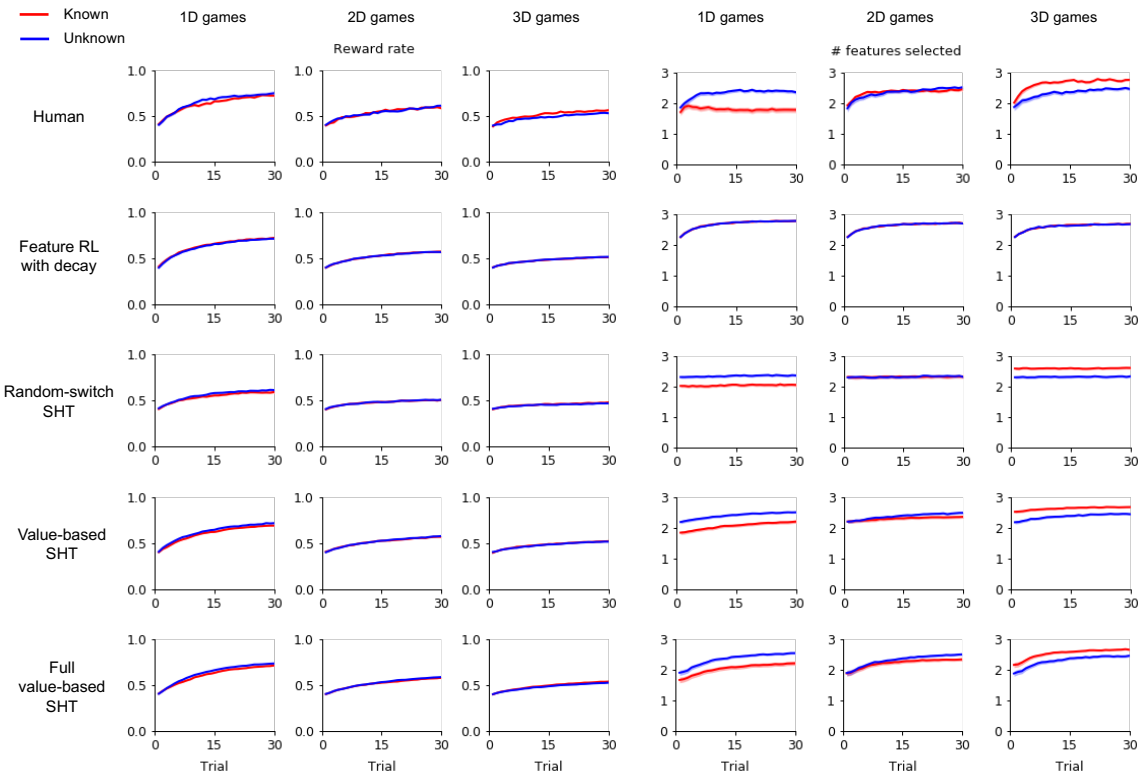

Figure 4.6: **Learning curves for data and model simulations.** The top row and the fourth row are the same as Figures 4.2A,B and 4.4B,C, respectively.

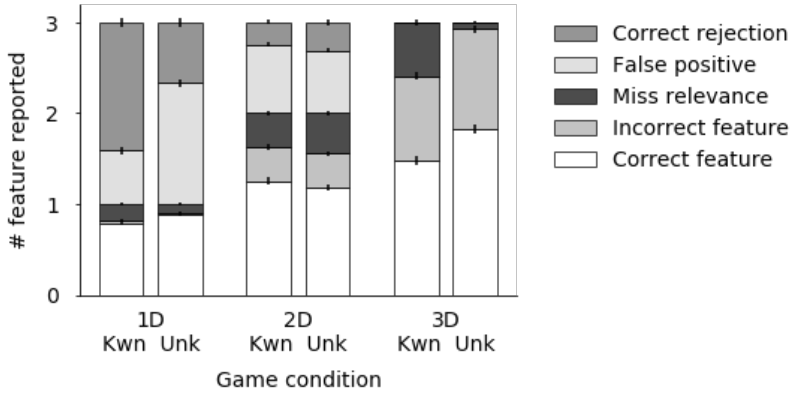

Figure 4.7: Full results of post-game responses to questions about the rewarding features in each game condition.

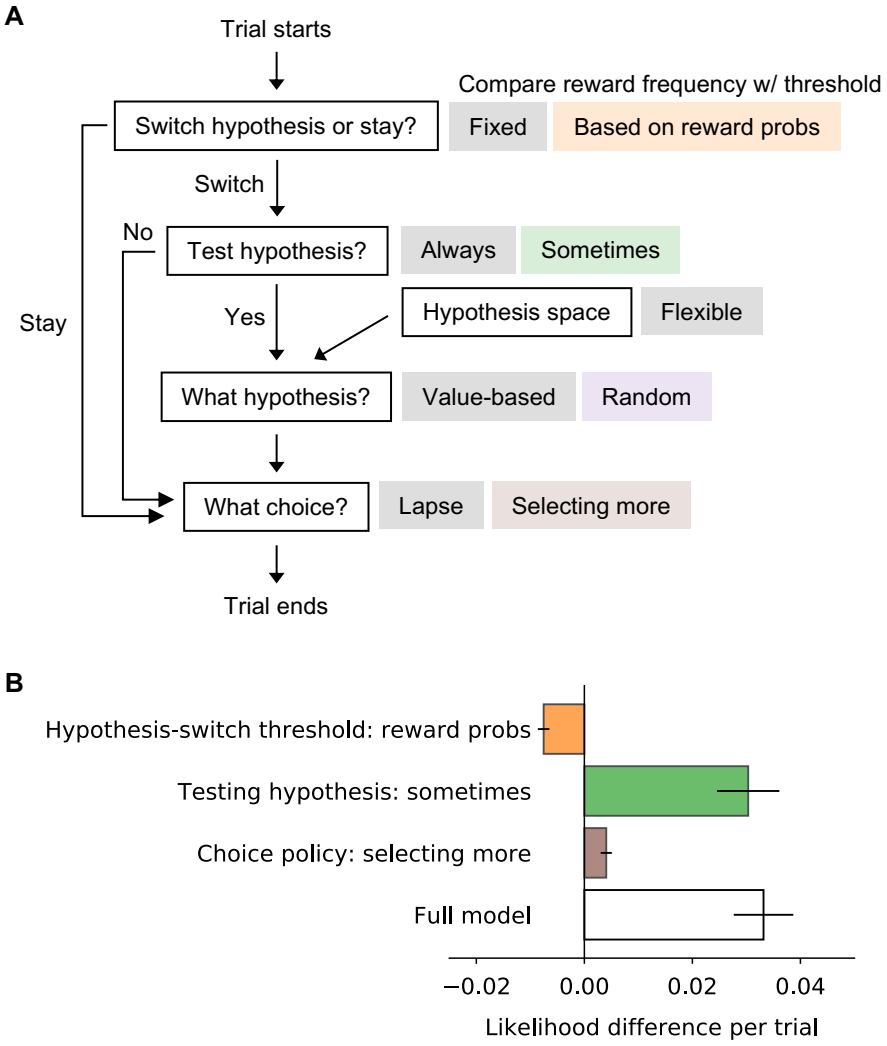

Figure 4.8: **Variants of the serial hypothesis testing model. (A)** A diagram of the serial hypothesis testing models. Behavioral and model variables are presented in circles, and model assumptions are presented in white boxes. Different variants on each model assumption are presented in colored boxes: in light gray are the assumptions adopted by the baseline model, and in other colors are those used in the model variants. **(B)** Difference in average likelihood per trial between variants of the SHT models and the baseline model (the value-based SHT model). All models except the full model are only different from the baseline model by one assumption as noted in the label; the full model adopts the better alternative in every assumption. Bar colors correspond to those in panel A, except for the full model (in white).

4.5 Supplementary Methods

4.5.1 Variants of the value-based SHT model

We consider a few variants of the value-based SHT model, by modifying the hypothesis-testing policy and the choice policy of the baseline value-based SHT model described above.

Not always testing hypothesis

In the experiment, the participant could choose not to select any feature, and let the computer configure a random stimulus. In fact, many participants did so, especially in the beginning of each game, which was potentially due to not having a good candidate hypothesis in mind. To model this inability to come up with hypotheses, we add a soft threshold on hypothesis testing: if the expected reward of the best candidate hypothesis is even below a threshold θ_{test}, participants will be unlikely to test any hypothesis:

$$Pr(\text{test}) = \frac{1}{1 + e^{-\beta_{\text{test}}(\max_h(ER(h)) - \theta_{\text{test}})}}$$

β_{test} and θ_{test} are free parameters. This mechanism is applied to the first trial of each game and at hypothesis switch points.

Alternative hypothesis-testing policy: using reward probability information

In the experiment, participants were informed of the reward probabilities for all game conditions (Table 4.2.1), which is not used by the baseline model. One way to use such information is to calculate a target reward probability $RP_{\text{target}}(h|D, D(h))$: the highest possible reward probability for the current hypothesis (if all features in the current hypothesis are rewarding features, while not exceeding the instructed number of relevant dimensions D). In "known" games, we assume that participants set their

threshold according to this target reward probability, with a free-parameter offset δ:

$$\theta = RP_{\text{target}}(h|D, D(h)) + \delta$$

The intuition is that the participant should expect a higher reward probability, for example, when testing the same one-dimensional hypothesis in a 1D game ($RP_{\text{target}} = 0.8$) compared to in a 3D game ($RP_{\text{target}} = 0.4$). The average RP_{target} of 1D, 2D and 3D games is used for "unknown" games.

Alternative choice policy: selecting more features than hypothesis

In the baseline model, participants' choices are assumed to be aligned with their current hypothesis, unless they have a lapse. In the experiment, however, we observed an overall tendency to select more features than instructed (Figure 4.2B). This was not surprising as there was no cost for selecting more features. In fact, it is strictly optimal to always make selections on all dimensions, as there is always a best feature within each dimension (at least equally good as the other two) according to the participant's mental model. Thus, we assume in this alternative model that participants may select more features than their current hypothesis h_t. The probability for choices that are supersets of h_t is determined by how similar it is to h_t (number of dimensions that differ), with a decay rate k as a free parameter:

$$P(c_t|h_t) \propto e^{k(D(c_t)-D(h_t))}$$

Participants may still lapse: all choices that are not supersets of h_t are equally likely, with probabilities sum to λ.

4.5.2 Inference in the serial hypothesis testing models

The random-switch and value-based SHT models assume that participants serially test hypotheses. As experimenters, however, we do not observe what hypotheses they are testing. In order to predict their choice c_t on trial t, we need to marginalize over all possible hypotheses h_t:

$$P(c_t|D_{1:t-1}) = \sum_{h_t} P(c_t|h_t)P(h_t|D_{1:t-1})$$

The first term $P(c_t|h_t)$ is given by the choice policy, as discussed before. In this section, we describe how to calculate the second term, i.e., infer the hypothesis that the participant is currently testing based on their choice and reward history ($D_{1:t-1}$). We do so using the change point detection model in [115].

We first introduce the run-length of hypothesis on trial t as l_t, i.e., how long the participant has been testing the current hypothesis. The probability of the current hypothesis can then be written as the marginalization over run-length of the current and last trials (on the first trial, all hypotheses are equally likely, corresponding to a uniform prior on h_1):

$$\begin{aligned}
P(h_t|D_{1:t-1}) &= \sum_{l_t,l_{t-1}} P(h_t|l_t, l_{t-1}, D_{1:t-1})P(l_t, l_{t-1}|D_{1:t-1}) \\
&= \sum_{l_t,l_{t-1}} P(h_t|l_t, l_{t-1}, D_{1:t-1})P(l_t|l_{t-1}, D_{1:t-1})P(l_{t-1}|D_{1:t-1}) \\
&= \sum_{l_t,l_{t-1}} \sum_{h_{t-1}} \left(P(h_t|l_t, h_{t-1}, l_{t-1}, D_{1:t-1})P(h_{t-1}|l_{t-1}, D_{1:t-1})\right) P(l_t|l_{t-1}, D_{1:t-1})P(l_{t-1}|D_{1:t-1}
\end{aligned}$$

For the rest of this section, we describe how to calculate each term (color-coded) receptively.

The second term $P(h_{t-1}|l_{t-1}, D_{1:t-1})$: recursive calculation

The second term can be calculated recursively using Bayes' rule (normalization is needed):

$$P(h_{t-1}|l_{t-1}, D_{1:t-1}) \propto P(c_{t-1}|h_{t-1}, l_{t-1}, D_{1:t-2})P(h_{t-1}|l_{t-1}, D_{1:t-2})$$

On the second trial (special case with $t = 2$):

$$P(h_1|l_1, D_1) \propto P(c_1|h_1)P(h_1|l_1),$$

where $P(h_1|l_1)$ is set to the prior belief distribution (uniform distribution).

Starting the third trial:

$$\begin{aligned} P(h_{t-1}|l_{t-1}, D_{1:t-1}) &\propto P(c_{t-1}|h_{t-1})P(h_{t-1}|l_{t-1}, D_{1:t-2}) \\ &= P(c_{t-1}|h_{t-1}) \sum_{l_{t-2}} P(h_{t-1}|l_{t-1}, l_{t-2}, D_{1:t-2})P(l_{t-2}|D_{1:t-2}) \\ &= P(c_{t-1}|h_{t-1}) \sum_{l_{t-2}} \sum_{h_{t-2}} \left(P(h_{t-1}|l_{t-1}, h_{t-2}, l_{t-2}, D_{1:t-2})P(h_{t-2}|l_{t-2}, D_{1:t-2})\right) P(l_{t-2}| \end{aligned}$$

where $P(c_{t-1}|h_{t-1})$ is given by the choice policy.

The fourth term $P(l_{t-1}|D_{1:t-1})$: recursive calculation

The fourth term is initialized as an array of a single element 1 on the second trial, and can be calculated recursively using Baye's rule starting the third trial:

$$
\begin{aligned}
P(l_{t-1}|D_{1:t-1}) \propto & P(c_{t-1}|l_{t-1}, D_{1:t-2})P(l_{t-1}|D_{1:t-2}) \\
= & \sum_{h_{t-1}} P(c_{t-1}|h_{t-1}, l_{t-1}, D_{1:t-2})P(h_{t-1}|l_{t-1}, D_{1:t-2}) \sum_{l_{t-2}} P(l_{t-1}|l_{t-2}, D_{1:t-2})P(l_{t-2}|D_{1:t-2}) \\
= & \sum_{h_{t-1}} P(c_{t-1}|h_{t-1}) \sum_{l_{t-2}} P(h_{t-1}|l_{t-1}, l_{t-2}, D_{1:t-2})P(l_{t-2}|D_{1:t-2}) \sum_{l_{t-2}} P(l_{t-1}|l_{t-2}, D_{1:t-2})P(\\
= & \sum_{h_{t-1}} P(c_{t-1}|h_{t-1}) \sum_{l_{t-2}} \sum_{h_{t-2}} \left(P(h_{t-1}|l_{t-1}, h_{t-2}, l_{t-2}, D_{1:t-2})P(h_{t-2}|l_{t-2}, D_{1:t-2})\right) P(l_{t-2}|D \\
& \sum_{l_{t-2}} P(l_{t-1}|l_{t-2}, D_{1:t-2})P(l_{t-2}|D_{1:t-2})
\end{aligned}
$$

The third term $P(l_t|l_{t-1}, D_{1:t-1})$: hypothesis-testing policy

We calculate the third term $P(l_t|l_{t-1}, D_{1:t-1})$ by marginalizing over the hypothesis from last trial:

$$P(l_t|l_{t-1}, D_{1:t-1}) = \sum_{h_{t-1}} P(l_t|h_{t-1}, l_{t-1}, D_{1:t-1})P(h_{t-1}|l_{t-1}, D_{1:t-1})$$

The serial hypothesis testing assumption implies that l_t can only take on two possible values: $l_{t-1} + 1$ if the participant stay with the hypothesis from last trial, and 0 if they switch hypothesis.

$$
\begin{aligned}
P(l_t = l_{t-1} + 1|h_{t-1}, l_{t-1}, D_{1:t-1}) &= P_{\text{stay}} \\
P(l_t = 0|h_{t-1}, l_{t-1}, D_{1:t-1}) &= 1 - P_{\text{stay}}
\end{aligned}
$$

P_{stay} is a function of $h_{t-1}, l_{t-1}, D_{1:t-1}$, and is determined by the participant's hypothesis-testing policy (Equation 4.5).

The first term $P(h_t|l_t, h_{t-1}, l_{t-1}, D_{1:t-1})$: hypothesis-switch policy

The first term $P(h_t|l_t, h_{t-1}, l_{t-1}, D_{1:t-1})$ is given by the switch policy. As noted before, only certain combinations of l_t and l_t values are possible: either $l_t = l_{t-1}+1$ (stay), or $l_t = 0$ (switch). Specifically, for $l_t = l_{t-1}+1$ (stay), $P(h_t = h_{t-1}|h_{t-1}, l_t, l_{t-1}, D_{1:t-1}) = 1$, otherwise, 0; for $l_t = 0$ (switch), $P(h_t \neq h_{t-1}|h_{t-1}, l_t, l_{t-1}, D_{1:t-1})$ is determined as in Equations 4.6 and 4.7. Other l_t, l_{t-1} combinations all have probability zero.

Chapter 5

Using recurrent neural networks to study representation learning

The contents of this chapter were published in: Mingyu Song, Yael Niv, and Ming Bo Cai. Using recurrent neural networks to understand human reward learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.

All data and code are available at `https://github.com/mingyus/RNN-cogsci2021`.

5.1 Introduction

Computational models of human cognition have greatly helped us in understanding the way people learn and make decisions. For example, reinforcement learning (RL) models reveal how people learn to acquire reward from trial-and-error experience; Bayesian inference models demonstrate how people combine prior knowledge and observations to form beliefs about the world. However, despite the great power of these models and what they have taught us about human mind, much variance in the data is often left unexplained even with the best-predicting cognitive models at hand. For example, in the multi-dimensional probabilistic learning task studied in Chapter 4, an average likelihood of $p = 0.22$ per trial was achieved using the best cognitive model. It is unclear whether this seemingly low model likelihood is purely due to the stochastic nature of human behavior, or whether it indicates room for improvement and potentially a better model. This is particularly relevant for complex sequential learning tasks, where decisions depend not only on the current stimulus, but also the history of experience; and is further complicated when there is a large number of candidate choices (e.g., 64 different choices in the above experiment), making it hard to precisely predict participants' behavior.

Understandably, most cognitive modeling work focus on relative model comparison, which identifies the best-fit model out of a number of alternatives under consideration. Studies do not, however, commonly evaluate the absolute goodness of fit of models, or estimate how far they are from the best possible model. In theory, there exists an upper bound for model log likelihood (i.e., the negative entropy of the data [126]). However, estimating this bound is practically impossible in sequential tasks, because the exact same experience is never repeated (it may be possible, however, in simpler perceptual decision-making tasks where the same choice can be tested repeatedly [127]). In this work, we propose an alternative empirical approach: to use recurrent neural networks (RNNs) to predict human choices.

RNNs are neural networks with recurrent connections that can pass information from one time point to the next. They are widely used to model temporal dynamics, making them particularly suitable for capturing the sequential dependence of human behavior. Compared to cognitive models that have carefully-crafted assumptions based on knowledge about human cognition, RNNs are agnostic to cognitive processes. They usually have thousands of free parameters (as compared to cognitive models with around 10). The flexibility means that RNNs can potentially capture more variance than do cognitive models, given a sufficient amount of training data (and at the expense of a clearly understood process model of how the participant made their decision). Despite the fact that RNNs have been widely used to solve cognitive tasks [29, 128, 129], only a handful of works have used RNNs to directly model the way *people* solve these tasks by predicting their behavior on a trial-by-trial basis [130, 131, 132]. In these works, RNNs have been applied to bandit problems with very sparse reward [130] or no clearly better options [132], and were able to predict either counter-intuitive win-shift-lose-stay behavior, or stereotyped alternation of options, both of which common cognitive models (e.g., reinforcement learning models) failed on. In fact, due to the uninformativeness of reward signals in those specific tasks, heuristics or stereotyped behavior were more likely, which may explain the success of RNNs as compared to reinforcement learning models.

In the current work, we sought to apply this approach to more standard reward-learning scenarios, in a task with a large enough choice space that will ensure sufficient variance to be explained. We thus considered the probabilistic multi-dimensional reward learning task studied in Chapter 4. In this task, participants tried to learn about multi-dimensional rules through actively configuring three-dimensional stimuli and receiving probabilistic feedback for their configuration. Prior work found that participants' learning strategy consisted of a combination of reinforcement learning and hypothesis testing with rich individual differences. The best available cogni-

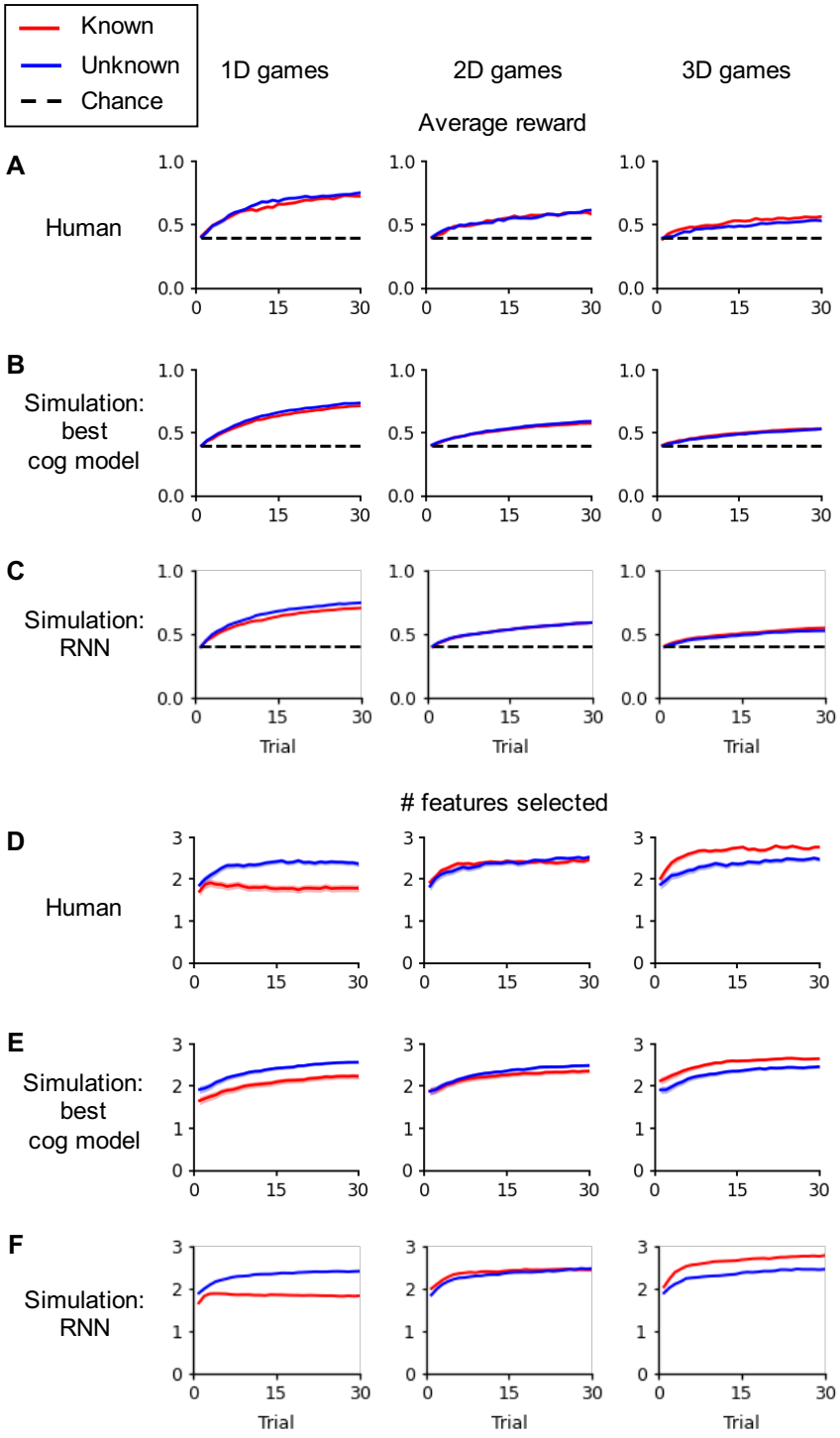

Figure 5.1: **Performance and choice behavior in the multi-dimensional reward learning task. (A-C)** The average reward, and **(D-F)** the number of features selected per trial over the course of 1D, 2D and 3D games (left, middle and right columns); red and blue curves represent the "known" and "unknown" conditions, respectively. Shaded areas: 1 s.e.m. across participants. Dashed lines: chance level for that type of game. (A,D): participants' behavior; (B,E): simulation of the best cognitive model; (A-D) are the same as Figure 4.2A,B and Figure 4.4B,C; (C,F): simulation of the RNN model.

tive model, the "value-based serial hypothesis testing model", out-performed many commonly-used cognitive models, and was able to account for individual differences in belief space and choice policy. However, it still deviated from the data in some important aspects. Specifically, it failed to precisely predict the number of features participants selected on every trial (Figure 5.1E), as mentioned in Chapter 4.

We applied RNNs to fit participants' behavior in this task, and found that RNNs predicted choices better than the best cognitive model (average likelihood increased by about 0.04 per trial, from the original $p = 0.22$), suggesting room for improvement for cognitive models. We investigated the underlying reasons for RNNs' superior performance, and found at least two: RNNs were more accurate at capturing the dependency between consecutive choices, and RNNs worked well with the large choice space by identifying the subspace which participants used more accurately than the best cognitive model. We further considered the rich individual difference observed in this task, and incorporated it into the RNN by adding an embedding of individual participants. Participant embedding helped model fits, especially for the first few trials of each independent "game". The embedding space encoded meaningful cognitive variables whose variance across participants was not captured by an RNN without embedding.

5.2 Apply RNNs to fit behavior

In this work, we used recurrent neural networks (RNNs) to fit behavior in the aforementioned task (see Chapter 4 for task design, participants' behavior and comparison of cognitive models). The RNN model used here consists of an input layer, a recurrent layer, and an output layer (Figure 5.2A). On each trial, the input variables consist of a game-start indicator (1 if it is the first trial, 0 otherwise), the game type, the participant's choice (c_{t-1}), the configured stimulus (s_{t-1}), and the outcome (r_{t-1}), with

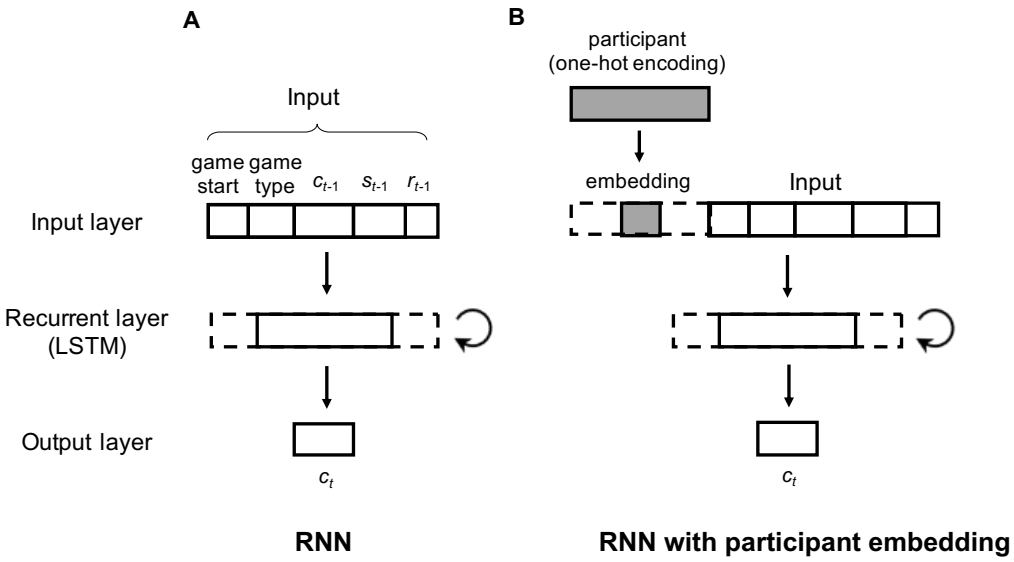

Figure 5.2: **RNN model structure. (A)** The RNN model. **(B)** The RNN with participant embedding model. Circled arrows indicate recurrent connections. Dashed rectangulars indicate varying layer sizes (set by hyper-parameters).

the last three variables taken from the previous trial and are zero on the first trial of a game. Each input variable is one-hot encoded, forming a concatenated binary input vector. The recurrent layer is a long short-term memory (LSTM) layer, followed by a linear feed-forward connection to the output layer, which is then transformed by a Softmax function (with inverse temperature fixed at 1) to determine the probability for choices on the current trial (c_t). We used the cross-entropy loss, i.e., log likelihood of participant's choice, as the cost function. We optimized the network using the Adam optimizer in PyTorch. Hyper-parameters of the models included learning rate, batch size, and the size of the recurrent layer.

We split data into training, validation and test sets. The training set consisted of 16 games from each participant (augmented 1296 times through shuffling the dimensions and features; see Discussion for details), and the validation and test sets each consisted of 1 game per participant. The game type and game index were balanced (to the extent possible) in each set to reduce potential bias or order effect. The weights of the networks were trained on the training set, the hyper-parameter values

were selected based on the validation set (the selected values were as follows: learning rate of 0.001, batch size of 10000, recurrent layer size of 50, and early stopping at epoch 28), and all results reported were evaluated on the test set using the best fit network.

5.3 Compare RNN with the best cognitive model

The RNN performed better than the best cognitive model. The RNN was able to quantitatively capture the number of features selected by participants throughout a game, which the best cognitive model failed on (Figure 5.1F). The advantage of RNN over the best cognitive model persisted through the course of a game (Figure 5.3A). Because the same network was used to predict all participants, the RNN described above could not be personalized for individual participants; this is in contrast to cognitive models that fit a separate set of parameters for each participant. This can explain the worse fit of the RNN compared to the best cognitive model on the first trial of each game.

In the following, we try to identify the reasons for the RNN's better performance, and how they might help us understand what is lacking in the cognitive models. First, we divided all trials into two types: trials where participants repeated the previous choice ("stay trials") and trials where they did not ("switch trials"); this analysis ignored the first trial of each game, which did not fall into either category. The advantage of the RNN over the cognitive model is primarily due to the switch trials (Figure 5.3B). This result was not due to the RNN being biased towards predicting switches. In fact, when examining how accurate the models are at predicting whether a trial would be a stay or switch trial, the RNN assigned a higher probability for both choice types than did the cognitive model (Figure 5.3C)[1], suggesting that it was

[1]Note that (1) predicting "switch" means correctly identifying the "switch" choice type, but not necessarily correctly predicting what the participant switched to; (2) stay trials were relatively easy

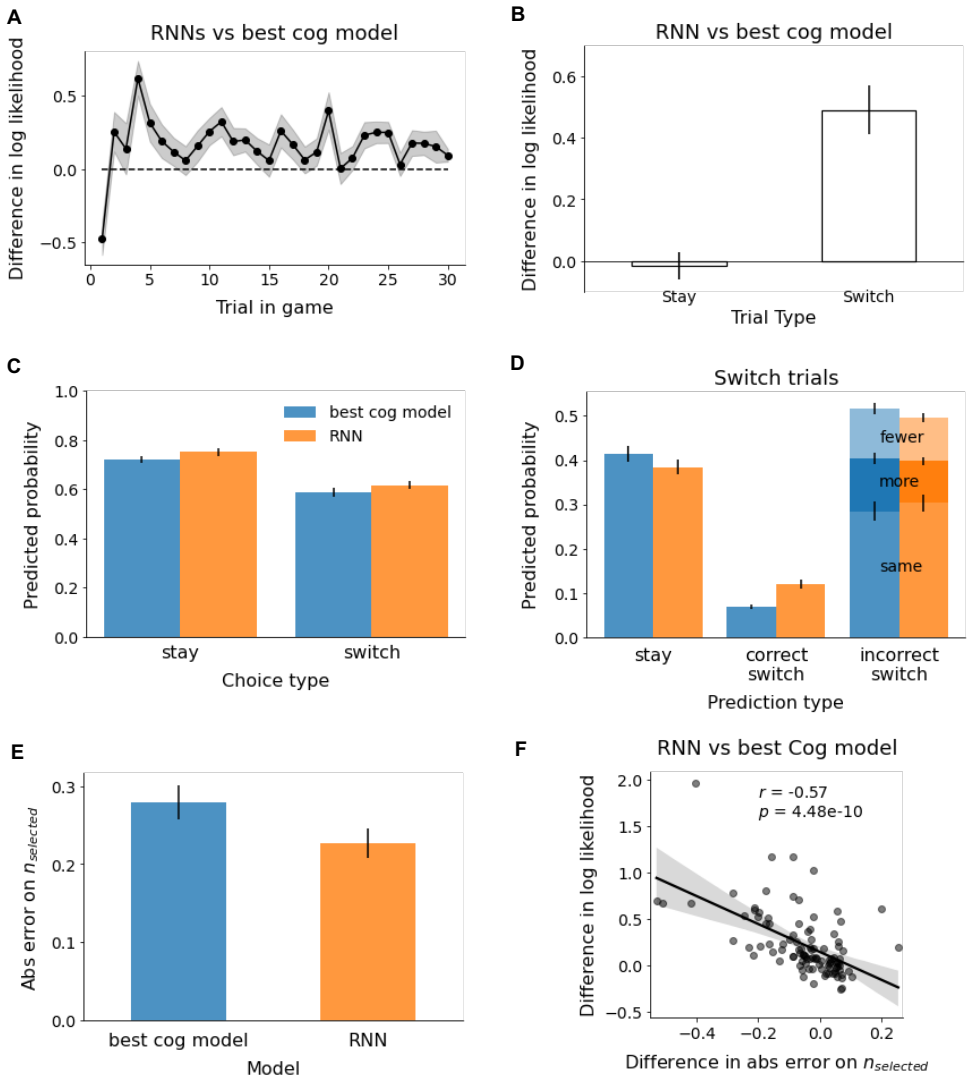

Figure 5.3: **Comparison between the RNN and the best cognitive model reveals the advantage of RNN in predicting switch trials and the number of features selected per trial.** **(A)** Log likelihood difference between the RNN and the best cognitive model, over the course of a game. Positive values indicate better prediction by RNN. **(B)** Log likelihood differences between the models for stay and switch trials respectively. **(C)** Probability that the models assigned for staying or switching on "stay" or "switch" trials, respectively. For "switch" trials, the probability was calculated as the sum of all possible switch choices. Blue bars: best cognitive model; orange bars: RNN; same in (D,E). **(D)** Model prediction on switch trials only, showing separately the probability that the model predicted for stay, correct switch, and incorrect switch choices (further divided into "fewer", "more" and "same", depending on the number of features selected in each choice, n_{selected}, relative to the true choice). **(E)** Mean absolute error in the number of features selected by the two models, calculated by taking the expectation over all possible choices with respect to the predicted probability of each choice. **(F)** Difference in log likelihood between the RNN and best cognitive model as a function of their difference in absolute error on n_{selected}, for each participant. The better the prediction on the number of features selected, the better the RNN did in fitting the participants' overall behavior in the test game. Shaded area (A) and error bars (B-E): 1 s.e.m. across participants.

better at correctly identifying how the next trial depended on the previous one (stay or switch).

We focused on switch trials to further investigate the RNN's better performance (Figure 5.3D). We categorized all possible choices into mistakenly predicting stay, predicting the correct switch, or predicting a switch to an incorrect choice. In the last case, we further categorized the choices based on whether they involved selecting fewer, the same number of, or more features than did the participant. Consistent with the results above, the RNN made fewer mistakes on switch versus stay (lower predicted probability for stay, and higher for correct switch). When it did make a mistake, the RNN was more likely to at least correctly predict the number of features selected. This was confirmed by an overall lower absolute error on predicting the number of features selected (Figure 5.3E). Finally, across participants, a lower absolute error in predicting the number of features selected by the RNN as compared to the cognitive model was correlated with a greater log likelihood advantage (i.e., a better fit) for the RNN model (Figure 5.3F).

Taken together, these analyses showed that the RNN was better at correctly predicting the number of features selected. This is particularly useful when the choice space is very large (as in this task), as it helps the RNN identify the correct subspace of the true choice. This is, however, not the complete story: the RNN was more likely to predict the true choice even within the correct subspace (predicted probability ratio between the true choice and all switch choices with the correct n_{selected} is 0.304 for RNN and 0.241 for the cognitive model), the reason behind which remains to be investigated.

---

to predict for both models, thus their likelihood was much higher than switch trials; additional improvement when the likelihood is high contributes less to the total *log* likelihood (cross-entropy loss), explaining the small difference between the models in Figure 5.3B.

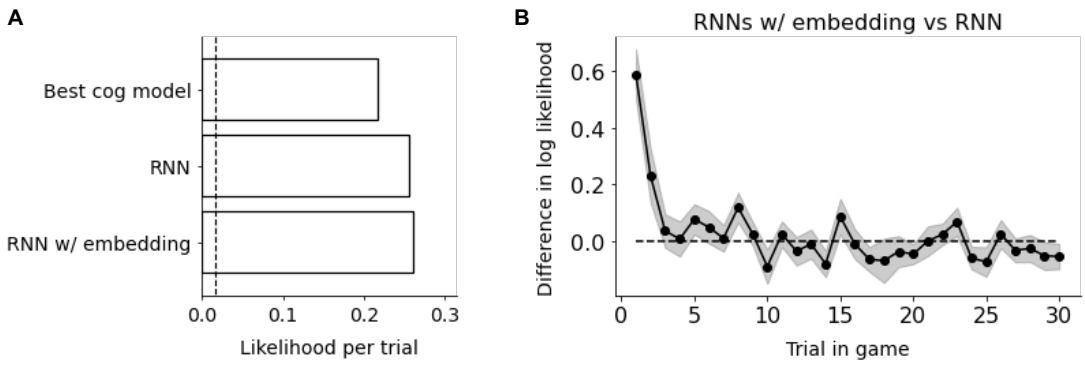

Figure 5.4: **RNN model comparison shows the effect of embedding**. **(A)** Model likelihood comparison between the best cognitive model, RNN and RNN with participant embedding. **(B)** Log likelihood difference between RNN with and without embedding, over the course of a game. Shaded area: 1 s.e.m. across participants.

5.4 Embedding captures individual differences

So far, we used the same RNN, referred to as **the RNN model**, to predict data from all participants. This works well if all participants use the same strategy. By fitting cognitive models to individual participants, however, we obtained different parameter estimates, suggesting that there might be strategy differences across individuals. Thus, we considered **the RNN with participant embedding model** (Figure 5.2B), by adding an embedding of individual participants to the input layer of the original RNN. The embedding was trained end-to-end together with the rest of the network. The size of the embedding layer was an additional hyper-parameter. As with the other hyper-parameters, its value was selected based on model performance on a validation set. The best-fit RNN with participant embedding model used the following values: learning rate of 0.001, batch size of 10000, recurrent layer size of 50, participant embedding size of 3, early stopping at epoch 23.

Adding the participant embedding improved the performance of the network (Figure 5.4A), most notably in the first two trials of each game (Figure 5.4B).

Despite the limited improvement in prediction, the embedding was crucial for reproducing individual differences. To investigate what information is encoded in

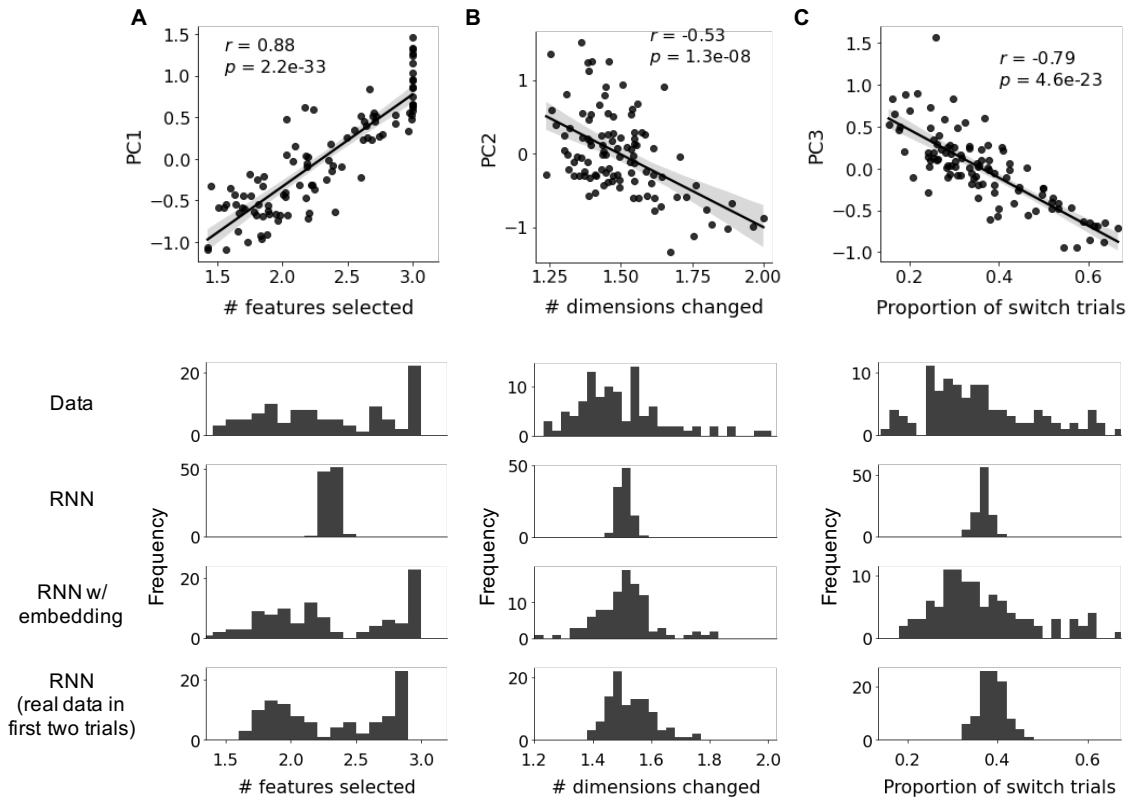

Figure 5.5: **Participant embedding encodes individual differences.** Top row: the first three principle components of the embedding layer are correlated with **(A)** average number of features selected per trial; **(B)** average number of dimensions changed on switch trials; **(C)** the proportion of switch trials. Bottom rows: histograms of the corresponding variables in participants' data and model simulations of the RNN model, the RNN with participant embedding model, and the RNN model that used participants' data in first two trials of each game.

the embedding, we performed a PCA analysis on the embedding activity. The three principal components (PCs) were correlated with the following cognitive variables of individual participants, respectively (Figure 5.5 top row): (1) PC1: average number of features selected per trial; (2) PC2: average number of dimensions changed on switch trials; (3) PC3: the overall proportion of switch trials.

We tested how well the network models captured individual differences by comparing the histograms of these variables in data with those obtained from model simulations. The RNN model failed to capture individual differences; in fact, it could only predict the mean of these variables. In contrast, the RNN with participant embedding

model, was able to capture the distribution of all three variables, demonstrating the usefulness of the embedding. Inspired by the finding that adding embedding improved model fits primarily in the first two trials (Figure 5.4B), we further tested the RNN model by providing it with participants' data (choices and outcomes) for the first two trials in the simulation. With such information, the RNN model *without* embedding was able to capture the variance of some variables (e.g., number of features selected), but still failed on others (e.g., proportion of switch trials). This suggests that the embedding layer encodes information beyond what can be extracted from first two trials of data.

Taken together, these results show the usefulness of an embedding layer in RNNs in capturing the characteristics of individual participants. The embedding layer can then be used to measure similarity between people and serve as the basis of finding subgroups in a population. In fact, similar work has been done by Dezfouli and colleagues [131], where an RNN with embedding model was fit to two-armed bandit task data. One of the two dimensions of the embedding space was found to differentiate between healthy, depressed and bipolar populations.

Moreover, it is worth noting that the embedding activity was found to encode cognitively-meaningful information in the current task. This is promising if generalizable to other tasks. Neural networks are notoriously hard to interpret despite being powerful prediction tools. The participant embedding, however, helps to make RNNs more interpretable, and can be useful in identifying important task variables that determine participants' strategies. Related, if the embedding vector is found to be correlated with certain cognitive functions (e.g., working memory capacity), the measurement of such cognitive functions can then be used to predict a participant's behavior on the task, and vice versa.

5.5 Discussion

We showed that RNN models fit human behavior better than cognitive models in a complex reward-learning task. The best cognitive model previously developed for this task used a hybrid strategy that combines reinforcement learning and serial hypothesis-testing, and considered individual differences in how participants understand instructions, their hypothesis-testing policy, and choice policy. However, such a sophisticated cognitive model still failed to fully explain participants' choices. RNN models, in contrast, made no assumption about cognitive processes, but were able to generate more precise predictions both at the group level and for individual participants. Analyses of model predictions revealed that the advantage of RNNs persisted throughout the course of learning. In particular, RNNs were better at predicting "stay" versus "switch" trials, i.e., how the current choice related to the previous one, as well as predicting what subset of the large choice space was used by participants (i.e., the number of features selected). As a next step, we can utilize the insights gained from RNNs to improve cognitive models, in order to achieve a better account of human behavior in this task.

We hope to demonstrate with this work the general value in training RNNs to predict human behavior in complex cognitive tasks. This approach has so far been under-utilized, with only a handful of applications [131, 130, 132]. This is potentially due to the large amount of data required for training neural network models. In the current work, we were able to augment the training set by utilizing the symmetric task structure. Assuming that participants' strategies did not depend on specific dimensions or features, we generated auxiliary data by shuffling the dimensions and features in each game (for both choices and stimuli), which effectively increased the training data size by about 10 times (although still less than the ideal size, and we ran the risk of under-fitting; Figure 5.6). Such data augmentation may not be possible for every task; a more general solution for using limited amount of data more efficiently

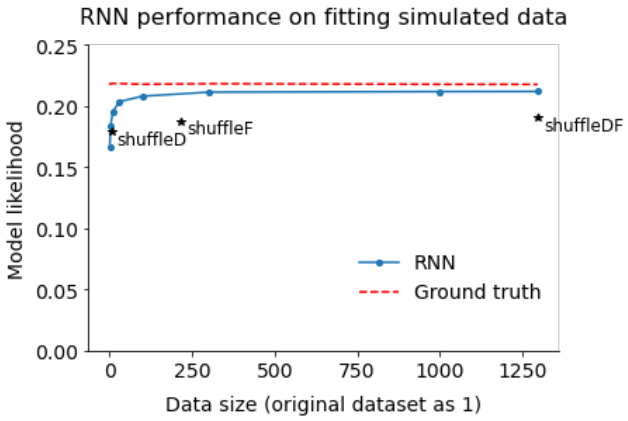

Figure 5.6: **Dependency of RNN performance on training data size, and effect of data augmentation.** We simulated "data" of varying sizes (1 to 1296 times of the original dataset) with the best cognitive model using best-fit parameters, and used the RNN model to fit these datasets. Red dashed line: ground-truth likelihood of the generative model; blue dots: likelihood of the RNN model. RNN performance improved with data size, asymptotically approaching the ground truth. We augmented the simulated data of size 1 by 6, 216, and 1296 times through shuffling dimensions (shuffleD), features (shuffleF) or both (shuffleDF; as we did for the human data), and fit the RNN model to the augmented datasets; results shown in black asterisk. Data augmentation helped increase the effective data size, but only by roughly 10 times.

may be to use cognitive models as priors for neural network models [133]. As big data becomes more and more common in cognitive science [134], the RNN approach would become more powerful. In fact, large-scale experiments, combined with the use of artificial neural networks, have already shown merits in developing interpretable models of human economic and moral decision-making [135, 136].

The RNN approach has been successfully applied in scenarios where the available cognitive models (most often RL models) are not good candidates for explaining observed behavior that is largely heuristic or stereotyped [131, 132]. The current work, in contrast, showcases another use case for this approach: complex cognitive tasks with rich individual variability. In tasks such as the current one, where the best available cognitive model cannot fully capture the richness in behavior, RNNs can be useful in (1) finding the empirical upper bound for goodness of fit; (2) revealing what is missing in the cognitive models; (3) capturing the richness of individual behavioral

differences. In this work, we achieved these goals by conducting model comparison, analyzing the winning RNN models, and developing RNN models with participant embedding. Future work can seek to apply RNNs in other similarly complex cognitive tasks to improve our understanding of human cognition and its variability.

Chapter 6

Conclusion

6.1 Contributions

I began this dissertation by demonstrating the importance of studying representation learning. In Chapter 2, I showed through model comparison that rats do not form the optimal state representation in a seemingly simple choice task. I argued that such a deviation of animals' representation from how the task was designed suggests the need to carefully examine the actual representation held by animals and humans in decision tasks, as well as its learning process. To my knowledge, this is the first study to formally test different task representations underlying animal behavior in a simple trial-and-error learning choice task – the type of task that is often thought of as the "bread and butter" reinforcement learning task where representations are trivial. My results underscore the importance of testing our assumptions—about representations, not just about the learning processes that act on those representations—rather than assuming that our experimenter-centric understanding of a decision task is shared by our experimental subjects.

To study the cognitive mechanisms underlying representation learning, I then examined two types of learning problems: grouping individual experience into states, and identifying relevant information for a task among distractor features.

In Chapter 3, I developed a computational model that explains how animals infer latent states in fear extinction experiments. Specifically, I provided a quantitative explanation on why a gradual extinction procedure is more effective than the standard procedure (or a reverse procedure) in preventing the return-of-fear in the long term. These data had not been previously explained by a computational model. Although the gradual extinction procedure was originally inspired by the latent-cause framework, which had conceptually predicted its long-lasting extinction effect, that framework on its own was unable to explain the differences between gradual extinction (which led to long-lasting extinction, as predicted) and the gradual reverse control condition (which resulted in spontaneous recovery and reinstatement of fear). My

modeling work investigated where the latent-cause framework came short, and suggested adjustments to the model that would bring it in line with empirical results. These additional assumptions enrich the original framework, and make predictions for other learning scenarios as well.

In Chapter 4, I examined how humans learn to identify relevant factors in complex environments with redundant information. This is a common problem in real-world scenarios that has only received attention in recent years. However, previous experiments only looked at scenarios where participants can choose among a small set of options that are made available to them. In the real world, we can often craft our own learning experiences, deciding what options to test, and what configurations of features to combine. To test this type of learning, I designed a novel active-learning task where human participants created multidimensional stimuli and received probabilistic feedback for their creations. I systematically examined participants' learning strategy and how it depends on task complexity. I showed that participants combine rule-based and value-based strategies in learning, and trade them off based on their costs and benefits under various task complexities. This work expands our understanding of how people learn to identify relevant information to more realistic active-learning settings. In such rich and complex environments, I showed that human learning can be characterized as a hypothesis-testing process based on value learning. My work therefore helps to unify the various findings on rule-based and value-based strategies in representation learning literature, and opens up the opportunity to further examine the integration of the two learning mechanisms in behavior and in the brain.

Finally, in Chapter 5, I proposed a novel computational approach for studying representation learning with recurrent neural networks (RNNs). By applying RNNs to the rule-learning task in Chapter 4, I showcased the utilities of RNNs in (a) setting targets for developing cognitive models, (b) providing insights on how to improve

cognitive models, and (c) studying individual cognitive variability. There is a growing interest in deep neural networks in representation learning research. Specifically, RNNs have been broadly used to model how the brain solves cognitive tasks. However, very few studies have investigated a different way of using RNNs, as flexible function approximators, to predict behavior. My work demonstrated the avenues for such use of these networks, and how they can interface with more traditional cognitive modeling research.

6.2 General Discussion

Representation learning is a key component of animal and human learning in realistic, complex environments. It comes hand-in-hand with reward learning (i.e., learning the association between stimuli/actions and outcomes): on the one hand, representation learning provides the foundation for reward learning by forming the critical state representation; on the other hand, representation learning rarely happens without feedback, and relies on reward (or punishment) as learning signals. Despite the important role of representation learning, there is little consensus on its cognitive or neural mechanisms, in contrast to a unified theory of reward learning once a state representation is assumed [6] (driven by error signals and supported by the midbrain dopaminergic system [4, 5]).

In this dissertation, I studied the computational mechanisms that underlie representation learning, and tried to uncover common cognitive principles across various learning scenarios. In Chapter 3, I studied how animals group individual experiences into hidden states in fear extinction experiments. During fear extinction, my models suggested that animals perform probabilistic inference on the underlying states based on sequences of experience. Such inference relies on animals' prior belief on how new states were generated, and my work suggested that animals use an approximation

of the posterior belief that involves occasional collapse of full posterior belief distributions to their mode. In Chapter 4, I studied how humans learn about complex rules that define states. In a multi-dimensional probabilistic reward-learning task, I showed that participants use a combination of value-learning and rule-learning strategies, biasing towards one or the other based on the instructed task complexity. The rule-based strategy is a simpler alternative to optimal Bayesian inference that reduces computation and memory load.

Despite the apparent differences between the two learning tasks, they together reveal general principles of representation learning (in addition to the common reward learning component that learns the "value" of the hidden states or candidate rules through trial-and-error): During representation learning, agents often need to consider multiple possibilities (e.g., how experience should be grouped into clusters that may change over time, or how different factors may together determine outcomes), and figure out which one is the most likely based on observations. The normative way to differentiate between alternatives and identify the correct task structure is through probabilistic inference, i.e., optimally integrating information from experience and combining it with prior knowledge. However, optimal inference can be intractable with numerous possibilities. As shown in the above two studies, animals and humans instead use approximate inference to simplify computation and reduce memory load, either by collapsing posterior belief distributions, or by reducing the size of the hypothesis space. In addition, they form useful inductive biases (or priors) to help with the inference process, either by using a time-dependent prior that reflects the dynamics of a changing environment, or by trading off two learning strategies based on the instructed task complexity. Together, studies in this dissertation suggested that approximate inference and inductive biases serve as two general principles that underlie animal and human representation learning.

Similar principles have previously been proposed in related tasks. For example, when solving categorization tasks [137, 138] or learning about causal relationships [121], humans have been shown to use approximate inference approaches (e.g., importance sampling, Markov Chain Monte Carlo sampling); they also tend to favor simpler rules over more complex ones *a priori* [120, 139], showing an inductive bias that represents the dominance of simple relationships in nature. It is still largely unknown how these principles may be implemented in the brain (but see [140, 141]), and further, whether various learning scenarios are supported by a common neural mechanism. As a first step towards a theory of representation learning, future work can test these principles and their specific implementations in different scenarios, and study their neural implementations.

Related, these cognitive principles may be useful for developing artificial intelligence that is more sample-efficient and generalizes better across tasks. Domain knowledge or inductive biases have already been shown to help neural networks learn faster with fewer samples in tasks like reasoning about intuitive physics or mimicking to generate hand-written characters [142]. It can be promising to incorporate these cognitively-inspired principles into representation-learning tasks as well.

In this dissertation, I also demonstrated the usefulness of two general computational approaches in studying the cognitive mechanisms for representation learning. First, to understand learned task representations, I showed that it is useful to compare alternative models with different state representations. For example, in Chapter 2, I demonstrated this approach with an odor-guided decision task in rats, and showed that animals formed a task representation that prohibited efficient generalization between trial types. Further, to characterize the learning process, I showed the utility of generic recurrent neural networks (RNNs). In Chapter 5, I used the aforementioned multi-dimensional probabilistic reward learning task as an example, and demonstrated

the usefulness of RNNs in both setting a target for developing cognitive models and providing insights on the underlying cognitive mechanisms and individual variability.

In sum, this dissertation contributes to advancing our understanding of representation learning both mechanistically and methodologically. Through the study of a few learning scenarios, this dissertation reveals general principles underlying the process animals and humans learn about task structures, i.e., how they use approximate inference and inductive biases to help with reasoning over numerous possibilities. Future work can test these principles in a wider range of tasks, and further investigate their potential neural underpinnings. Additionally, this dissertation demonstrates ways to study representation learning with formal model comparison and a novel neural network approach, both of which can be widely applied to a myriad of tasks.

Bibliography

[1] P Ivan Pavlov. *Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex*. Oxford Univ. Press., 1927.

[2] Burrhus Frederic Skinner. *Science and human behavior*. Number 92904. Simon and Schuster, 1965.

[3] Robert A Rescorla. A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current research and theory*, pages 64–99, 1972.

[4] P Read Montague, Peter Dayan, and Terrence J Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of neuroscience*, 16(5):1936–1947, 1996.

[5] Wolfram Schultz, Peter Dayan, and P Read Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599, 1997.

[6] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[7] Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, 6(5):679–684, 1957.

[8] Nathaniel D Daw, Yael Niv, and Peter Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience*, 8(12):1704–1711, 2005.

[9] Robert Colin Honey and Geoffrey Hall. Acquired equivalence and distinctiveness of cues. *Journal of Experimental Psychology: Animal Behavior Processes*, 15(4):338, 1989.

[10] G Elliott Wimmer, Nathaniel D Daw, and Daphna Shohamy. Generalization of value in reinforcement learning by humans. *European Journal of Neuroscience*, 35(7):1092–1104, 2012.

[11] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

[12] John O'keefe and Lynn Nadel. *The hippocampus as a cognitive map*. Oxford university press, 1978.

[13] Nicolas W Schuck, Ming Bo Cai, Robert C Wilson, and Yael Niv. Human orbitofrontal cortex represents a cognitive map of state space. *Neuron*, 91(6):1402–1412, 2016.

[14] Nathaniel D Daw, Samuel J Gershman, Ben Seymour, Peter Dayan, and Raymond J Dolan. Model-based influences on humans' choices and striatal prediction errors. *Neuron*, 69(6):1204–1215, 2011.

[15] Alexandra O Constantinescu, Jill X O'Reilly, and Timothy EJ Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468, 2016.

[16] Robert C Wilson, Yuji K Takahashi, Geoffrey Schoenbaum, and Yael Niv. Orbitofrontal cortex as a cognitive map of task space. *Neuron*, 81(2):267–279, 2014.

[17] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

[18] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[19] James CR Whittington, Timothy H Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy EJ Behrens. The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation. *Cell*, 183(5):1249–1263, 2020.

[20] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn. *arXiv preprint arXiv:1611.05763*, 2016.

[21] Samuel Joseph Gershman, Carolyn E Jones, Kenneth A Norman, Marie-H Monfils, and Yael Niv. Gradual extinction prevents the return of fear: implications for the discovery of state. *Frontiers in behavioral neuroscience*, 7:164, 2013.

[22] Angela J. Langdon, Mingyu Song, and Yael Niv. Uncovering the "state": Tracing the hidden state representations that structure learning and decision-making. *Behavioural Processes*, 167:103891, 2019.

[23] Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

[24] Yael Niv. Learning task-state representations. *Nature Neuroscience*, 22(10):1544–1553, 2019.

[25] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[26] Matthew Botvinick, Sam Ritter, Jane X. Wang, Zeb Kurth-Nelson, Charles Blundell, and Demis Hassabis. Reinforcement learning, fast and slow. *Trends in Cognitive Sciences*, 23(5):408–422, 2021/02/23 2019.

[27] Jane X Wang, Zeb Kurth-Nelson, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Remi Munos, Charles Blundell, Dharshan Kumaran, and Matt Botvinick. Learning to reinforcement learn, 2017.

[28] Jane X. Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z. Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature Neuroscience*, 21(6):860–868, 2018.

[29] Guangyu Robert Yang, Madhura R Joglekar, H Francis Song, William T Newsome, and Xiao-Jing Wang. Task representations in neural networks trained to perform many cognitive tasks. *Nature Neuroscience*, 22(2):297–306, 2019.

[30] Matthew R. Roesch, Adam R. Taylor, and Geoffrey Schoenbaum. Encoding of time-discounted rewards in orbitofrontal cortex is independent of value representation. *Neuron*, 51(4):509–520, 2006.

[31] Yuji K. Takahashi, Angela J. Langdon, Yael Niv, and Geoffrey Schoenbaum. Temporal specificity of reward prediction errors signaled by putative dopamine neurons in rat vta depends on ventral striatum. *Neuron*, 91(1):182–193, 2016.

[32] Matthew R. Roesch, Teghpal Singh, P. Leon Brown, Sylvina E. Mullins, and Geoffrey Schoenbaum. Ventral striatal neurons encode the value of the chosen action in rats deciding between differently delayed or sized rewards. *Journal of Neuroscience*, 29(42):13365–13376, 2009.

[33] Amanda C. Burton, Gregory B. Bissonette, Daniela Vazquez, Elyse M. Blume, Maria Donnelly, Kendall C. Heatley, Abhishek Hinduja, and Matthew R. Roesch. Previous cocaine self-administration disrupts reward expectancy encoding in ventral striatum. *Neuropsychopharmacology*, 43(12):2350–2360, 2018.

[34] Bob Carpenter, Andrew Gelman, Matthew D Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. Stan: A probabilistic programming language. *Journal of statistical software*, 76(1), 2017.

[35] Andrew Gelman, Hal S Stern, John B Carlin, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[36] Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *J. Mach. Learn. Res.*, 11:3571–3594, December 2010.

[37] Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27(5):1413–1432, September 2017.

[38] Jingfeng Zhou, Matthew P.H. Gardner, Thomas A. Stalnaker, Seth J. Ramus, Andrew M. Wikenheiser, Yael Niv, and Geoffrey Schoenbaum. Rat orbitofrontal ensemble activity contains multiplexed but dissociable representations of value and task structure in an odor sequence task. *Current Biology*, 29(6):897 – 907.e3, 2019.

[39] Kevin J Miller, Matthew M Botvinick, and Carlos D Brody. Dorsal hippocampus contributes to model-based planning. *Nature Neuroscience*, 20(9):1269–1276, September 2017.

[40] Brian M. Sweis, Samantha V. Abram, Brandy J. Schmidt, Kelsey D. Seeland, Angus W. MacDonald, Mark J. Thomas, and A. David Redish. Sensitivity to "sunk costs" in mice, rats, and humans. *Science*, 361(6398):178–181, 2018.

[41] Thomas L. Griffiths, Falk Lieder, and Noah D. Goodman. Rational use of cognitive resources: Levels of analysis between the computational and the algorithmic. *Topics in Cognitive Science*, 7(2):217–229, 2015.

[42] Falk Lieder and Thomas L. Griffiths. Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, 43:e1, 2020.

[43] Sean B. Ostlund and Bernard W. Balleine. Orbitofrontal cortex mediates outcome encoding in pavlovian but not instrumental conditioning. *Journal of Neuroscience*, 27(18):4819–4825, 2007.

[44] Kate M. Wassum, Sean B. Ostlund, Bernard W. Balleine, and Nigel T. Maidment. Differential dependence of pavlovian incentive motivation and instrumental incentive learning processes on dopamine signaling. *Learning Memory*, 18(7):475–483, 2011.

[45] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3):229–256, 1992.

[46] Daniel Bennett, Yael Niv, and Angela J Langdon. Value-free reinforcement learning: policy optimization as a minimal model of operant behavior. *Current Opinion in Behavioral Sciences*, 41:114–121, 2021.

[47] David M. Ferrero, Jamie K. Lemon, Daniela Fluegge, Stan L. Pashkovski, Wayne J. Korzan, Sandeep Robert Datta, Marc Spehr, Markus Fendt, and

Stephen D. Liberles. Detection and avoidance of a carnivore odor by prey. *Proceedings of the National Academy of Sciences*, 108(27):11235–11240, 2011.

[48] Daniel W Wesson, Ryan M Carey, Justus V Verhagen, and Matt Wachowiak. Rapid encoding and perception of novel odors in the rat. *PLOS Biology*, 6(4):1–13, 04 2008.

[49] Naoshige Uchida and Zachary F Mainen. Speed and accuracy of olfactory discrimination in the rat. *Nature Neuroscience*, 6(11):1224–1229, November 2003.

[50] Yuji K. Takahashi, Hannah M. Batchelor, Bing Liu, Akash Khanna, Marisela Morales, and Geoffrey Schoenbaum. Dopamine neurons respond to errors in the prediction of sensory features of expected rewards. *Neuron*, 95(6):1395–1405.e3, 2017.

[51] Jingfeng Zhou, Chunying Jia, Marlian Montesinos-Cartagena, Matthew P H Gardner, Wenhui Zong, and Geoffrey Schoenbaum. Evolving schema representations in orbitofrontal ensembles during learning. *Nature*, 590(7847):606–611, 2021.

[52] Anne GE Collins and Michael J Frank. Cognitive control over learning: creating, clustering, and generalizing task-set structure. *Psychological review*, 120(1):190, 2013.

[53] Aaron C. Courville, Nathaniel D. Daw, and David S. Touretzky. Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, 10(7):294–300, 2006. Special issue: Probabilistic models of cognition.

[54] A David Redish, Steve Jensen, Adam Johnson, and Zeb Kurth-Nelson. Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling., 2007.

[55] Terry E. Robinson, Lindsay M. Yager, Elizabeth S. Cogan, and Benjamin T. Saunders. On the motivational properties of reward cues: Individual differences. *Neuropharmacology*, 76:450 – 459, 2014. NIDA 40th Anniversary Issue.

[56] J.M. Koolhaas, R.F. Benus, and G.A. VAN OORTMERSSEN. Individual differences in behavioural reaction to a changing environment in mice and rats. *Behaviour*, 100(1-4):105 – 121, 01 Jan. 1987.

[57] Yael Niv, Jeffrey A Edlund, Peter Dayan, and John P O'Doherty. Neural prediction errors reveal a risk-sensitive reinforcement-learning process in the human brain. *Journal of Neuroscience*, 32(2):551–562, 2012.

[58] Yael Niv. Learning task-state representations. *Nature neuroscience*, 22(10):1544–1553, 2019.

[59] Stan Development Team. Pystan: the python interface to stan. `http://mc-stan.org`, 2018.

[60] Mingyu Song, Carolyn Jones, Marie-H. Monfils, and Yael Niv. Explaining the effectiveness of fear extinction through latent-cause inference, Oct 2021. `psyarxiv.com/2fhr7`.

[61] Joseph E Dunsmoor, Yael Niv, Nathaniel Daw, and Elizabeth A Phelps. Rethinking extinction. *Neuron*, 88(1):47–63, 2015.

[62] Robert A Rescorla. Spontaneous recovery. *Learning & Memory*, 11(5):501–509, 2004.

[63] Robert A Rescorla and C Donald Heth. Reinstatement of fear to an extinguished conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, 1(1):88, 1975.

[64] Mark E Bouton. Context and behavioral processes in extinction. *Learning & memory*, 11(5):485–494, 2004.

[65] Samuel J Gershman, David M Blei, and Yael Niv. Context, learning, and extinction. *Psychological review*, 117(1):197, 2010.

[66] Youssef Shiban, Jasmin Wittmann, Mara Weißinger, and Andreas Mühlberger. Gradual extinction reduces reinstatement. *Frontiers in behavioral neuroscience*, 9:254, 2015.

[67] Alina Thompson, Peter M McEvoy, and Ottmar V Lipp. Enhancing extinction learning: Occasional presentations of the unconditioned stimulus during extinction eliminate spontaneous recovery, but not necessarily reacquisition of fear. *Behaviour Research and Therapy*, 108:29–39, 2018.

[68] Najwa C Culver, Stephan Stevens, Michael S Fanselow, and Michelle G Craske. Building physiological toughness: Some aversive events during extinction may attenuate return of fear. *Journal of behavior therapy and experimental psychiatry*, 58:18–28, 2018.

[69] Samuel J Gershman and Yael Niv. Exploring a latent cause theory of classical conditioning. *Learning & behavior*, 40(3):255–268, 2012.

[70] David M Blei and Peter I Frazier. Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12(8), 2011.

[71] David J Aldous. Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII—1983*, pages 1–198. Springer, 1985.

[72] Nathan S Jacobs, Jesse D Cushman, and Michael S Fanselow. The accurate measurement of fear memory in pavlovian conditioning: resolving the baseline issue. *Journal of neuroscience methods*, 190(2):235–239, 2010.

[73] Kevin J Miller, Amitai Shenhav, and Elliot A Ludvig. Habits without values. *Psychological review*, 126(2):292, 2019.

[74] Brian Lau and Paul W Glimcher. Dynamic response-by-response models of matching behavior in rhesus monkeys. *Journal of the experimental analysis of behavior*, 84(3):555–579, 2005.

[75] Samuel J Gershman, Marie-H Monfils, Kenneth A Norman, and Yael Niv. The computational nature of memory modification. *Elife*, 6:e23763, 2017.

[76] Kevin Lloyd and David S Leslie. Context-dependent decision-making: a simple bayesian model. *Journal of The Royal Society Interface*, 10(82):20130069, 2013.

[77] Fabian A Soto, Samuel J Gershman, and Yael Niv. Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological review*, 121(3):526, 2014.

[78] Alan A Stocker and Eero P Simoncelli. A bayesian model of conditioned perception. *Advances in neural information processing systems*, 2007:1409, 2007.

[79] Makoto Ito and Kenji Doya. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. *Journal of Neuroscience*, 29(31):9861–9874, 2009.

[80] Joshua I Gold, Chi-Tat Law, Patrick Connolly, and Sharath Bennur. The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning. *Journal of neurophysiology*, 100(5):2653–2668, 2008.

[81] Daeyeol Lee, Benjamin P McGreevy, and Dominic J Barraclough. Learning and decision making in monkeys during a rock–paper–scissors game. *Cognitive Brain Research*, 25(2):416–430, 2005.

[82] Rei Akaishi, Kazumasa Umeda, Asako Nagase, and Katsuyuki Sakai. Autonomous mechanism of internal choice estimate underlies decision inertia. *Neuron*, 81(1):195–206, 2014.

[83] Lili X Cai, Katherine Pizano, Gregory W Gundersen, Cameron L Hayes, Weston T Fleming, Sebastian Holt, Julia M Cox, and Ilana B Witten. Distinct signals in medial and lateral vta dopamine neurons modulate fear extinction at different times. *Elife*, 9:e54936, 2020.

[84] Burrhus Frederic Skinner. *The behavior of organisms: an experimental analysis.* Appleton-Century, 1938.

[85] John O Cooper, Timothy E Heron, and William L Heward. *Applied behavior analysis*. Pearson Education, Incorporated, 2020.

[86] Karyn M Myers, Kerry J Ressler, and Michael Davis. Different mechanisms of fear extinction dependent on length of time since fear acquisition. *Learning & memory*, 13(2):216–223, 2006.

[87] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

[88] Samuel J Gershman. A unifying probabilistic view of associative learning. *PLoS computational biology*, 11(11):e1004567, 2015.

[89] Eran Eldar, Robb B Rutledge, Raymond J Dolan, and Yael Niv. Mood as representation of momentum. *Trends in cognitive sciences*, 20(1):15–24, 2016.

[90] Robert Stickgold. Sleep-dependent memory consolidation. *Nature*, 437(7063):1272–1278, 2005.

[91] Paul Smolen, Yili Zhang, and John H Byrne. The right time to learn: mechanisms and optimization of spaced learning. *Nature Reviews Neuroscience*, 17(2):77, 2016.

[92] Hermann Ebbinghaus. *Memory: A contribution to experimental psychology.* Dover, Oxford, England, 1964.

[93] K Matthew Lattal. Trial and intertrial durations in pavlovian conditioning: issues of learning and performance. *Journal of Experimental Psychology: Animal Behavior Processes*, 25(4):433, 1999.

[94] Sean Commins, Lorretto Cunningham, Deirdre Harvey, and Derek Walsh. Massed but not spaced training impairs spatial memory. *Behavioural brain research*, 139(1-2):215–223, 2003.

[95] Nicholas J Cepeda, Harold Pashler, Edward Vul, John T Wixted, and Doug Rohrer. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, 132(3):354, 2006.

[96] Long Luu and Alan A Stocker. Post-decision biases reveal a self-consistency principle in perceptual inference. *Elife*, 7:e33334, 2018.

[97] Mark E Bouton, Amanda M Woods, and Oskar Pineño. Occasional reinforced trials during extinction can slow the rate of rapid reacquisition. *Learning and Motivation*, 35(4):371–390, 2004.

[98] Karolien van den Akker, Remco C Havermans, and Anita Jansen. Effects of occasional reinforced trials during extinction on the reacquisition of conditioned responses to food cues. *Journal of Behavior Therapy and Experimental Psychiatry*, 48:50–58, 2015.

[99] Elisabetta Baldi, Carlo Ambrogi Lorenzini, and Corrado Bucherelli. Footshock intensity and generalization in contextual and auditory-cued fear conditioning in the rat. *Neurobiology of learning and memory*, 81(3):162–166, 2004.

[100] Samuel J Gershman and Catherine A Hartley. Individual differences in learning predict the return of fear. *Learning & behavior*, 43(3):243–250, 2015.

[101] Charles R Gallistel, Stephen Fairhurst, and Peter Balsam. The learning curve: implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, 101(36):13124–13131, 2004.

[102] Nathaniel Daw and Aaron Courville. The pigeon as particle filter. *Advances in neural information processing systems*, 20:369–376, 2008.

[103] Michael L Mack, Bradley C Love, and Alison R Preston. Dynamic updating of hippocampal object representations reflects new conceptual knowledge. *Proceedings of the National Academy of Sciences*, 113(46):13203–13208, 2016.

[104] Ian Ballard, Eric M Miller, Steven T Piantadosi, Noah D Goodman, and Samuel M McClure. Beyond reward prediction errors: Human striatum updates rule values during learning. *Cerebral Cortex*, 28(11):3965–3975, 2017.

[105] Yael Niv, Reka Daniel, Andra Geana, Samuel J Gershman, Yuan Chang Leong, Angela Radulescu, and Robert C Wilson. Reinforcement learning in multidimensional environments relies on attention mechanisms. *Journal of Neuroscience*, 35(21):8145–8157, 2015.

[106] Dimitrije Marković, Jan Gläscher, Peter Bossaerts, John O'Doherty, and Stefan J Kiebel. Modeling the evolution of beliefs using an attentional focus mechanism. *PLoS computational biology*, 11(10):e1004558, 2015.

[107] Klaus Wunderlich, Ulrik R Beierholm, Peter Bossaerts, and John P O'Doherty. The human prefrontal cortex mediates integration of potential causes behind observed outcomes. *Journal of neurophysiology*, 106(3):1558–1569, 2011.

[108] Oh-hyeon Choung, Sang Wan Lee, and Yong Jeong. Exploring feature dimensions to learn a new policy in an uninformed reinforcement learning task. *Scientific reports*, 7(1):17676, 2017.

[109] Katherine Duncan, Bradley B Doll, Nathaniel D Daw, and Daphna Shohamy. More than the sum of its parts: a role for the hippocampus in configural reinforcement learning. *Neuron*, 98(3):645–657, 2018.

[110] F Gregory Ashby, Leola A Alfonso-Reese, Elliott M Waldron, et al. A neuropsychological theory of multiple systems in category learning. *Psychological review*, 105(3):442, 1998.

[111] F Gregory Ashby and W Todd Maddox. Human category learning. *Annu. Rev. Psychol.*, 56:149–178, 2005.

[112] Angela Radulescu, Yael Niv, and Ian Ballard. Holistic reinforcement learning: the role of structure and attention. *Trends in cognitive sciences*, 2019.

[113] Lee W Gregg and Herbert A Simon. Process models and stochastic theories of simple concept formation. *Journal of Mathematical Psychology*, 4(2):246–276, 1967.

[114] Robert M Nosofsky, Thomas J Palmeri, and Stephen C McKinley. Rule-plus-exception model of classification learning. *Psychological review*, 101(1):53, 1994.

[115] Robert C Wilson and Yael Niv. Inferring relevance in a changing world. *Frontiers in human neuroscience*, 5:189, 2012.

[116] Rei Akaishi, Nils Kolling, Joshua W Brown, and Matthew Rushworth. Neural mechanisms of credit assignment in a multicue environment. *Journal of Neuroscience*, 36(4):1096–1112, 2016.

[117] Shiva Farashahi, Katherine Rowe, Zohra Aslami, Daeyeol Lee, and Alireza Soltani. Feature-based learning improves adaptability without compromising precision. *Nature communications*, 8(1):1–16, 2017.

[118] Shaoming Wang and Bob Rehder. Multi-attribute decision-making is best characterized by an attribute-wise reinforcement learning model. *BioRxiv*, page 234732, 2017.

[119] Angela Jones, Eric Schulz, and Bjoern Meder. Active function learning. In *the 40th Annual Meeting of the Cognitive Science Society*, 2018.

[120] Noah D Goodman, Joshua B Tenenbaum, Jacob Feldman, and Thomas L Griffiths. A rational analysis of rule-based concept learning. *Cognitive science*, 32(1):108–154, 2008.

[121] Neil R Bramley, Peter Dayan, Thomas L Griffiths, and David A Lagnado. Formalizing neurath's ship: Approximate algorithms for online causal learning. *Psychological review*, 124(3):301, 2017.

[122] Joshua Klayman and Young-Won Ha. Confirmation, disconfirmation, and information in hypothesis testing. *Psychological review*, 94(2):211, 1987.

[123] Bradley C Love, Douglas L Medin, and Todd M Gureckis. Sustain: a network model of category learning. *Psychological review*, 111(2):309, 2004.

[124] Anne GE Collins and Michael J Frank. Within-and across-trial dynamics of human eeg reveal cooperative interplay between reinforcement learning and working memory. *Proceedings of the National Academy of Sciences*, 115(10):2502–2507, 2018.

[125] Mingyu Song, Yael Niv, and Ming Bo Cai. Using recurrent neural networks to understand human reward learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43, 2021.

[126] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[127] Shan Shen and Wei Ji Ma. A detailed comparison of optimality and simplicity in perceptual decision making. *Psychological review*, 123(4):452, 2016.

[128] Jane X Wang, Zeb Kurth-Nelson, Dharshan Kumaran, Dhruva Tirumala, Hubert Soyer, Joel Z Leibo, Demis Hassabis, and Matthew Botvinick. Prefrontal cortex as a meta-reinforcement learning system. *Nature neuroscience*, 21(6):860–868, 2018.

[129] Qihong Lu, Uri Hasson, and Kenneth A. Norman. When to retrieve and encode episodic memories: a neural network model of hippocampal-cortical interaction. *bioRxiv*, 2021.

[130] Amir Dezfouli, Kristi Griffiths, Fabio Ramos, Peter Dayan, and Bernard W Balleine. Models that learn how humans learn: the case of decision-making and its disorders. *PLoS computational biology*, 15(6):e1006903, 2019.

[131] Amir Dezfouli, Hassan Ashtiani, Omar Ghattas, Richard Nock, Peter Dayan, and Cheng Soon Ong. Disentangled behavioural representations. In *Advances in neural information processing systems*, pages 2254–2263, 2019.

[132] Matan Fintz, Margarita Osadchy, and Uri Hertz. Using deep learning to predict human decisions and cognitive models to explain deep learning models. *bioRxiv*, 2021.

[133] David D Bourgin, Joshua C Peterson, Daniel Reichman, Stuart J Russell, and Thomas L Griffiths. Cognitive model priors for predicting human decisions. In *International conference on machine learning*, pages 5133–5141. PMLR, 2019.

[134] Jordan W. Suchow, Thomas L. Griffiths, and Joshua K. Hartshorne. Workshop on scaling cognitive science. In *the 42nd Annual Virtual Meeting of the Cognitive Science Society*, 2020.

[135] Mayank Agrawal, Joshua C Peterson, and Thomas L Griffiths. Scaling up psychology via scientific regret minimization. *Proceedings of the National Academy of Sciences*, 117(16):8825–8835, 2020.

[136] Joshua C Peterson, David D Bourgin, Mayank Agrawal, Daniel Reichman, and Thomas L Griffiths. Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, 372(6547):1209–1214, 2021.

[137] Adam N Sanborn, Thomas L Griffiths, and Daniel J Navarro. Rational approximations to rational models: alternative algorithms for category learning. *Psychological review*, 117(4):1144, 2010.

[138] Lei Shi, Thomas L Griffiths, Naomi H Feldman, and Adam N Sanborn. Exemplar models as a mechanism for performing bayesian inference. *Psychonomic bulletin & review*, 17(4):443–464, 2010.

[139] Ian Ballard, Eric M Miller, Steven T Piantadosi, Noah D Goodman, and Samuel M McClure. Beyond reward prediction errors: Human striatum updates rule values during learning. *Cerebral Cortex*, 28(11):3965–3975, 2018.

[140] Lei Shi and Thomas Griffiths. Neural implementation of hierarchical bayesian inference by importance sampling. *Advances in neural information processing systems*, 22:1669–1677, 2009.

[141] Jeffrey M Beck, Peter E Latham, and Alexandre Pouget. Marginalization in neural circuits with divisive normalization. *Journal of Neuroscience*, 31(43):15310–15319, 2011.

[142] Brenden M Lake, Tomer D Ullman, Joshua B Tenenbaum, and Samuel J Gershman. Building machines that learn and think like people. *Behavioral and brain sciences*, 40, 2017.