Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Behavioral Sciences

Discovering latent causes in reinforcement learning

Samuel J Gershman[1], Kenneth A Norman[2] and Yael Niv[2]

Effective reinforcement learning hinges on having an appropriate state representation. But where does this representation come from? We argue that the brain discovers state representations by trying to infer the latent causal structure of the task at hand, and assigning each latent cause to a separate state. In this paper, we review several implications of this latent cause framework, with a focus on Pavlovian conditioning. The framework suggests that conditioning is not the acquisition of associations between cues and outcomes, but rather the acquisition of associations between latent causes and observable stimuli. A latent cause interpretation of conditioning enables us to begin answering questions that have frustrated classical theories: Why do extinguished responses sometimes return? Why do stimuli presented in compound sometimes summate and sometimes do not? Beyond conditioning, the principles of latent causal inference may provide a general theory of structure learning across cognitive domains.

Addresses
[1] Department of Psychology and Center for Brain Science, Harvard University, United States
[2] Department of Psychology and Princeton Neuroscience Institute, Princeton University, United States

Corresponding author: Gershman, Samuel J (gershman@fas.harvard.edu)

Current Opinion in Behavioral Sciences 2015, **5**:43-50

This review comes from a themed issue on **Decision making/ neuroeconomics**

Edited by **John O'Doherty** and **Colin Camerer**

http://dx.doi.org/10.1016/j.cobeha.2015.07.007

Introduction

During his travels in Java, the 14th-century merchant Marco Polo encountered a rhinoceros. Lacking the concept of a rhinoceros, Marco Polo was faced with a dilemma: Is this a strange exemplar of a familiar category (unicorns), or a fundamentally new category? While ultimately judging it to be a unicorn, he acknowledged that 'they are not of that description of animals which suffer themselves to be taken by maidens, as our people suppose, but are quite of a contrary nature' [1].

Marco Polo's dilemma is ubiquitous: When confronted with a surprising observation, one must decide whether to extend a familiar category to incorporate this unusual example, or to postulate a novel category. We refer to the process of parsing experience into groups or delineating the boundaries of generalization among examples as *structure learning*. The idea is that this parsing process attempts to follow the true causal structure in the world: Experiences that are all instances of the same cause should be grouped together and generalized over, while experiences that are due to different underlying causes should be separated in our mind. Recent work has begun to unravel the cognitive and neural mechanisms supporting structure learning (see [2,3] for reviews). Our focus here is on the role of structure in reinforcement learning.

Structure learning is fundamental to reinforcement learning because these algorithms rely on a representation of the environment as a set of states, and the nature of the state representation determines the efficiency and efficacy of the learning algorithm. However, the state representation is almost never provided to the brain by a teacher. Instead, the brain must discover an appropriate (and useful) state representation from its interactions with the environment. Computational algorithms for state discovery have been studied in machine learning (e.g., [4–6]), but research on how the brain solves this problem is still in its infancy. We will review progress on one particular theory of state discovery, which posits that states are identified with inferred *latent causes* in the environment [7••,8–11,12•,13]. This theory connects with a rich array of ideas in psychology and neuroscience, ranging from classical conditioning to categorization and episodic memory [14]. Thus, the principles underlying state discovery in reinforcement learning may function as a core computational system in the brain.

Latent cause models

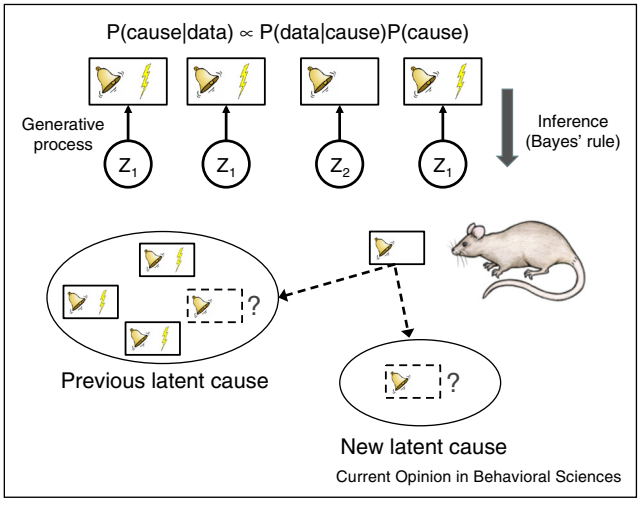
As a paradigmatic example of reinforcement learning, consider a simple Pavlovian conditioning experiment, in which a cue (e.g., tone) is paired repeatedly with an outcome (e.g., shock), resulting in the acquisition of a conditioned response (e.g., freezing). If the cue is then presented repeatedly without the outcome (extinction), the conditioned response gradually decreases and returns to baseline. Classical theories such as the Rescorla–Wagner model [15] view conditioning as the process of learning an association between cue and outcome, while extinction is viewed as unlearning of this association, with the conditioned response following the strength of the learned association. However, empirical data show that the extinguished response can return under a variety of circumstances (indeed, it is rather difficult to *prevent* the response from returning, as will be described below),

contradicting the assumption that the association was truly unlearned [16].

The latent cause framework addresses this contradiction by rethinking the representation underlying Pavlovian conditioning: Rather than learning about associations between cues and outcomes, which do not necessarily conform to the true causal structure of the environment, an agent learns about associations between 'causes,' some of them unobservable (latent), and observable stimuli (both cues and outcomes). In other words, the latent cause model conceptualizes associative learning as a form of 'clustering,' whereby observations are clustered together according to their hypothetical latent causes. By positing that animals infer the existence of different latent causes during conditioning and extinction, the latent cause model can explain both why and when the conditioned response survives extinction [8]. The model also addresses the role of the hippocampus in latent causal inference, developmental trajectories, and a host of related empirical phenomena [8,9].

Formally, the model posits that the animal computes the posterior distribution over latent causes given the observed sensory data such as cues and outcomes (Figure 1). This computation is stipulated by Bayes' rule, which states that the posterior probability of a set of causes is proportional to the likelihood of the sensory data assuming this causal structure, and the *a priori* probability of these causes. In the model, the likelihood is determined by the similarity between observations attributed to a single latent cause: the more similar the current observation is to previous observations attributed to a particular latent cause, the more likely it is that this latent cause is active on the current trial.

Figure 1



Conditioning as clustering. (Top) The latent cause model asserts that an animal will infer the latent causes of its observations using Bayes' rule. The rectangles show the animal's observations on different trials of a conditioning experiment, each generated by a single latent cause (denoted by Z_1 and Z_2). While only two latent causes appear here, the model can in principle accommodate an unbounded number. (Bottom) When presented with an extinction trial, the animal must determine whether this trial should be assigned to the existing latent cause or to a new one. In the model, this assignment is probabilistic rather than binary, and the extinction trial can ultimately be assigned to both the old and a new latent cause, with different posterior probabilities.

The prior distribution over latent causes expresses the agent's inductive biases about which latent structures are more or less plausible. Most commonly, the prior expresses a simplicity bias (cf. [17,18]) favoring latent structures with a small number of latent causes. Importantly, since in the real world the number of causes is often unknown, the prior must be able to accommodate an unbounded number of them. This combination of simplicity and flexibility can be formalized using concepts from Bayesian nonparametric statistics [19]. Box 1 provides an overview of several Bayesian nonparametric priors over latent causes.

While we have so far framed the latent cause model in terms of learning, the model also provides a framework for thinking about memory formation. Every time a new latent cause is inferred, a memory representation must be formed to store the parameters of the latent cause (e.g., the probabilities of each of the observations given that the cause is active, which are learned from experience). Memories, according to this view, constitute part of the agent's internal model of the world, with the structure of memories mirroring the inferred latent structure of the world. Bayesian inference negotiates the tension between updating existing memories and forming new ones when new information is encountered [10].

In the remainder of this review, we explore two implications of this latent cause framework for Pavlovian conditioning. First, we describe the effect of different extinction procedures on inferences about latent causes, leading to a surprising prediction that has been recently confirmed [20]. Second, we describe an extension of the model in which multiple latent causes can be simultaneously active (i.e., jointly generate the observed stimuli), allowing the model to capture a complex set of behavioral phenomena in compound conditioning [13].

Understanding the effects of different extinction procedures

To make the latent cause model concrete, we now describe its application to Pavlovian conditioning and extinction. In this paradigm, the sensory data consist of two variables, cue and outcome. Following our earlier work [8,9], we assume that on each trial of the experiment, only a single latent cause can be active (but see below for a relaxation of this constraint). The model assigns all cue-outcome conditioning trials unambiguously to a single cause, but the first

Box 1 Bayesian infinite-capacity ('nonparametric') priors over latent causes

The field of Bayesian nonparametric statistics has furnished cognitive science with several probability distributions over an unknown number of latent causes (see [19] for an introduction). These distributions have two psychologically plausible characteristics: (1) they can flexibly generate an unbounded number of latent causes, and (2) the number of generated latent causes tends to be small, thereby favoring 'simpler' structures and more parsimonious explanations of observed data.

Gershman and colleagues [8,9] used a simple sequential stochastic process for sampling latent causes, known as the *Chinese restaurant process* [56]. This name refers to a culinary metaphor that can be used to describe the sampling process (Figure 4a): customers (datapoints or trials) enter the restaurant one at a time, choosing a table (latent cause) with probability proportional to the number of other customers already seated at the table. With some probability proportional to α, a customer may choose a new table. Thus, the number of latent causes can grow as new data are observed, and the parameter α controls the rate of this growth. The Chinese restaurant process has also appeared in many areas of cognitive science, ranging from categorization [42,44] to visual perception [18] and reconstructive memory [10].

The Chinese restaurant process generates a single latent cause for each datapoint. A related stochastic process can generate multiple latent causes for each datapoint (Figure 4b), and it too can be described by a culinary metaphor. In the *Indian buffet process* [57], customers enter the buffet and sample each dish (latent cause) with probability proportional to the number of other customers who have previously sampled the dish. In addition, the nth customer samples a Poisson(λ/n) number of new (untried) dishes. The parameter λ plays the same role as α, controlling how the number of latent causes grows as more customers enter, and hence the 'complexity' of the model. Soto and colleagues used the Indian buffet process in their latent-cause model of compound conditioning [13], and it has also been used in models of perceptual feature learning [58–60].

How can we determine whether these stochastic processes are good models of human (or animal) priors? One promising approach is to use recently developed techniques for behaviorally eliciting samples from the prior. In a setting in which only one latent cause is relevant for each trial/observation, Austerweil [61•] used the Markov chain Monte Carlo with People technique [62] to collect samples from humans' priors over clusterings of simple visual stimuli (i.e., one latent cause per observation). This technique uses human judgments about clusterings to construct a Markov chain whose stationary distribution is the prior. Austerweil showed that the Chinese restaurant process is well-matched to elicited samples. A similar technique could be fruitfully applied to evaluating models with multiple simultaneous latent causes, like the Indian buffet process.

cue-no outcome extinction trial presents an ambiguous situation: Is the outcome absent because the true (unknown) parameters of the 'conditioning' cause specify a probability less than 1 of emitting an outcome, or should the absence of the outcome be interpreted as evidence of a new latent cause?

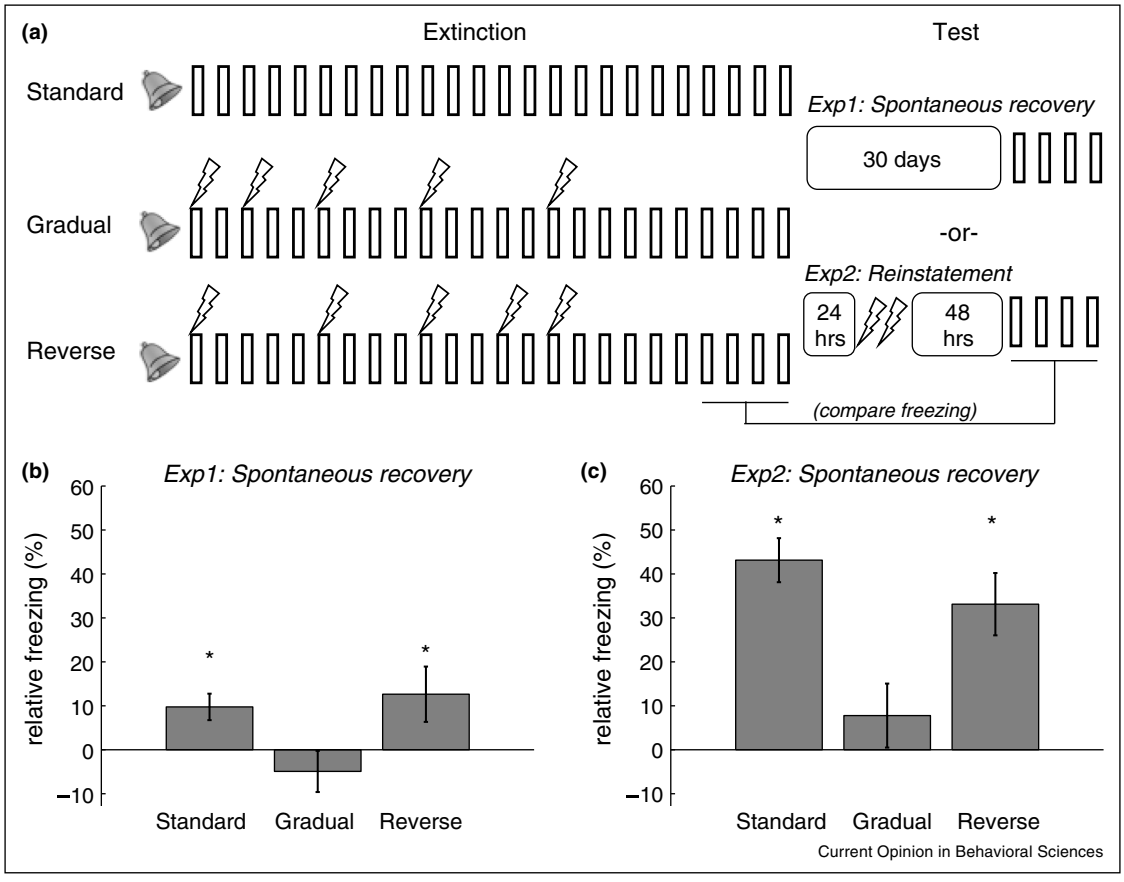
Bayes' rule balances both the (dis-)similarity between the extinction trial and the conditioning trials (captured by the likelihood) and the inductive bias for reusing previously inferred causes (captured by the prior), to determine the posterior probability of different explanatory structures: the probability that all trials have been generated by a single latent cause, the probability that all conditioning trials were generated by one cause, and all extinction trials by another, the probability that each trial was generated by a different latent cause, and so forth. In the case of extinction, this means that factors that reduce the similarity between extinction and acquisition (e.g., a context change; [8]) make it more likely that extinction trials will be assigned to a new latent cause.

To the extent that the posterior distribution over latent causes favors assigning the extinction trial to the old latent cause, the parameters of that cause will be modified (memory is updated), whereas to the extent that a new latent cause is inferred, a new memory will formed to store its parameters. According to this account, extinction fails to erase the original conditioning memory because the large change from conditioning to extinction invokes memory formation rather than memory updating. Thus, if we are interested in durably eliminating the conditioned response, we must thread a needle: new experiences should be sufficiently different to drive learning (memory updating), but not so different that they invoke memory formation.

We tested this idea by designing a *gradual extinction* procedure [20] for Pavlovian fear conditioning. Rather than abruptly shifting from cue-outcome (tone-shock) to cue-alone trials (standard extinction), we gradually reduced the frequency with which cues and outcomes are paired during extinction (Figure 2a). We compared this procedure to a standard extinction condition and a *gradual reverse* condition, in which cue-outcome pairs occurred on the same proportion of trials, but gradually increasing in frequency. In both control conditions, we expected that large differences between the beginning of extinction and the end of acquisition would lead to the inference of a new latent cause, and thus extinction training would not be effective at modifying the old (conditioning) latent cause. In contrast, with gradual extinction we hoped to encourage the animal to assign new observations to the old latent cause, thereby slowly modifying its parameters such that activation of the old cause predicted an ever shrinking probability of a shock.

We tested memory recovery with two assays (in different groups of rats). In experiment 1, we measured *spontaneous recovery* of fear 30 days following the end of extinction (Figure 1b) [21]. In experiment 2, we measured *reinstatement* of fear by presenting the shock 24 hours after extinction and then 24 hours later testing the conditioned response to the tone (Figure 1c) [22,23]. Spontaneous recovery and reinstatement tests typically show that the conditioned response returns after extinction. Indeed, we found such recovery of the fear memory in both the standard and gradual reverse conditions. However, consistent with our predictions, we found no evidence of

Figure 2



Varieties of extinction. **(a)** Three extinction procedures studied in [20]. The conditioned stimulus is a tone, denoted by bars, and the unconditioned stimulus is a shock. **(b)** Results of a spontaneous recovery experiment (test 30 days after extinction). **(c)** Results of a reinstatement experiment (two shocks without tones 24 hours after extinction, followed by a test 24 hours later). The conditioning and extinction procedures were identical across the two experiments. Both tests consisted of four presentations of the tone with no outcomes. Plotted are the differences between the conditioned response (freezing) at test and the response in the last four trials of extinction. Positive values indicate recovery of fear. Error bars represent standard error of the mean. Asterisks indicate effects significantly greater than zero ($p < 0.05$).

recovery of fear in the gradual extinction condition (Figure 2b,c). Thus, manipulating similarity can be an effective way to produce a long-lasting reduction of the conditioned response (the question of whether the memory was truly *erased* is fraught with difficulty; see [24]). Interestingly, similar effects of gradual change have been reported in motor [25–27] and visual [28–30,10] learning in humans.

Understanding compound generalization

So far we have been considering only the possibility of one latent cause per trial or per unit of experiment. However, this is clearly an artificial limitation and the true causal structure governing observed events can involve multiple latent causes that are active at once. Indeed, a rich area of research in animal conditioning concerns the nature of interactions between multiple cues that have independently been trained to signal rewards [31–34].

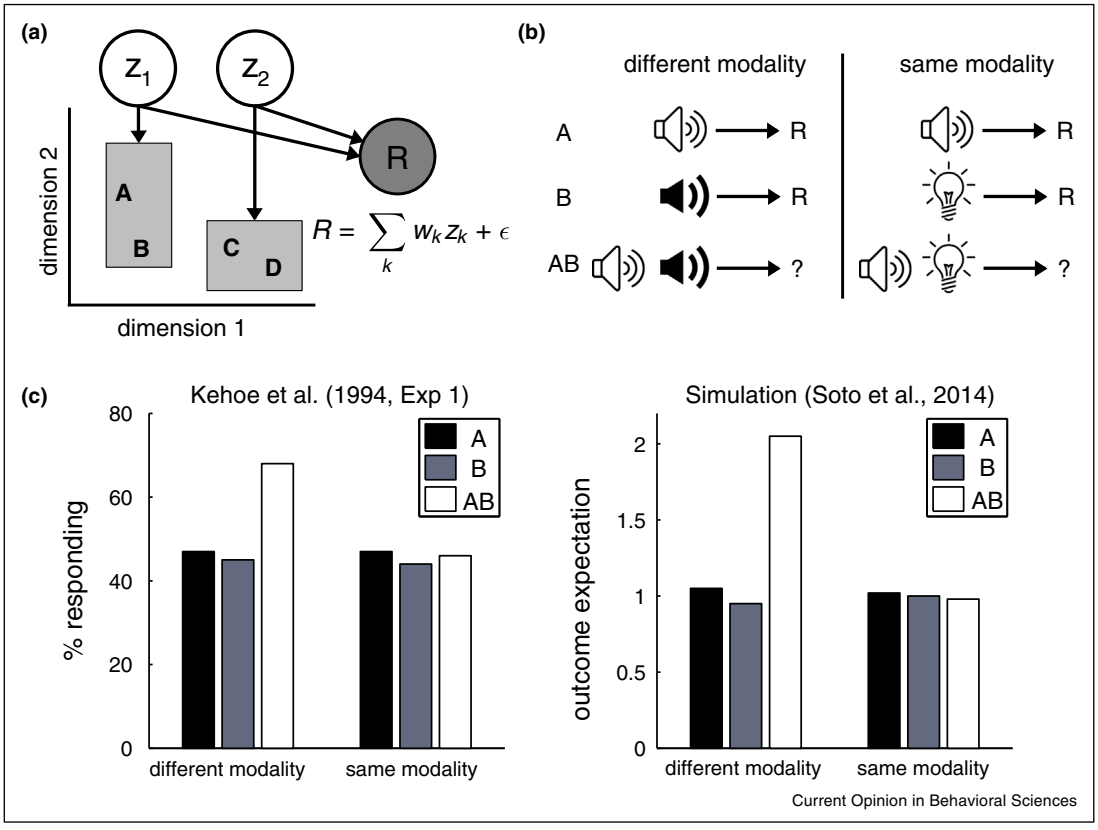
A simple paradigm for studying such interactions is known as *compound generalization*: Two cues (A and B) are separately paired with an outcome, and then tested in compound (AB). Elemental models (e,g., [15,32]) assume that each cue possesses its own association with the outcome, and these associations summate when the cues are presented in compound. Therefore the animal would expect two outcomes and accordingly emit double the conditioned response. Configural models (e.g., [31,35]), on the other hand, assume that associations are formed between whole cue configurations and outcomes. Here the outcome prediction is a weighted sum of outcome predictions for the training configurations, where the weight depends on the similarity between the training and test configurations. In compound generalization, cues A and B both have a similarity of 0.5 to the compound AB. This leads to averaging of the responses to each individual cue, rather than summation.

Despite extensive efforts to determine which of the two theories is correct, compound generalization has remained puzzling, sometimes acting in accordance with elemental models (i.e., showing summation effects), and sometimes in accordance with configural models (showing averaging effects). In particular, a number of factors influence the extent to which generalization appears elemental or configural, including spatial contiguity, temporal contiguity, and perceptual similarity between cues [36–39]. Existing mechanistic models have failed to account for the diversity of these effects.

In a recent paper [13], we formalized the principles underlying compound generalization using a variant of the latent cause model that allows multiple latent causes to be simultaneously active (see the description of the 'Indian buffet process' in Box 1). Here, as in the model of extinction above, latent causes generate the cues as well as the outcomes. Similarly to elemental models, in this model we assumed that the total outcome is a linear combination of outcomes caused by each of the active latent causes. Thus, the model predicts summation when the animal infers that each cue resulted from a distinct latent cause. However, because the latent causes — rather than cues — are combined to make predictions, the model does not predict a summation effect in cases in which the animal infers that both cues were generated by one latent cause.

How does the animal decide whether the two cues resulted from one or from two latent causes? Building on earlier theories of inductive generalization [40,41], the model assumes that each latent cause is associated with a region of stimulus space (termed a 'consequential region'), from which cues are drawn uniformly when the latent cause is active (Figure 3a). Importantly, different cues can be drawn from a single consequential region when only one cause is active, or from different consequential regions if several latent causes are active. When inferring whether two observed cues are more likely a

Figure 3

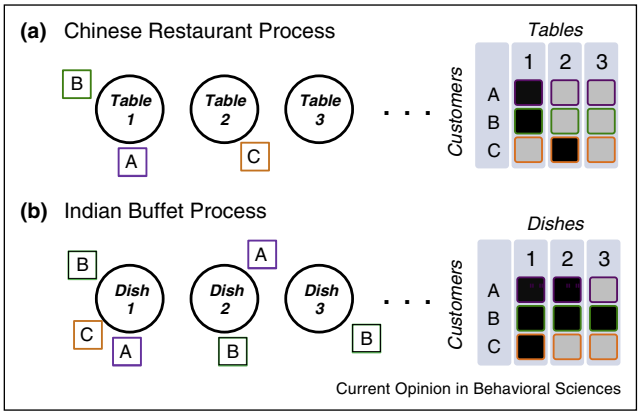


Modeling compound generalization in Pavlovian conditioning. **(a)** The latent cause model used by Soto and colleagues [13] assumes that multidimensional cues (denoted by A, B, C and D) are generated by first selecting a latent cause (denoted by Z_1 and Z_2) and then drawing uniformly from the rectangular consequential region associated with the selected latent cause. The perceived outcome (R) is a sum of the outcomes generated by the active latent causes ($z_k = 1$ when latent cause k is active, otherwise $z_k = 0$) plus Gaussian noise (ϵ). **(b)** Pavlovian conditioning experiment in which cues are trained separately and then tested in compound. In the 'different modality' condition, cue A is a tone and cue B is a light. In the 'same modality' condition, cue A is a tone and cue B is another sound. **(c)** Experimental results from [36] and simulations of the latent cause model from [13]. Summation is only observed (both experimentally and in the model) when the cues (labeled A and B) are from different modalities.

result of one or two latent causes, a natural consequence of Bayesian inference is the 'size principle' [41]: Small consequential regions have higher posterior probability than large consequential regions, all other things being equal, because smaller regions place more probability mass on the observed data. Thus, cues that are far apart in stimulus space will tend to be assigned to different latent causes, because small, localized consequential regions are more likely than a single large region spanning all the cues.

By assuming that spatial, temporal and perceptual factors are dimensions of the stimulus space, we were able to explain why compounds sometimes act elementally and sometimes configurally. When cues occupy nearby positions in stimulus space, or lie on a line parallel to one of the axes of the space, they can be explained by a single consequential region (one latent cause), for instance, a very narrow rectangle along the dimensional axis, which has a very small volume. When cues are far apart in stimulus space and do not lie parallel to axes of the space, it is more likely that they belong to different consequential regions (multiple latent causes), because if they were generated by one latent cause the consequential region associated with that cause would have to have a large volume, making the observed cues very unlikely. This simple principle is sufficient to capture a wide range of compound generalization phenomena [13]. An example is illustrated in Figure 3, where compounds from the same or different modalities are compared. When the cues are from different modalities (far apart in stimulus space), the conditioned response shows a summation effect, but this effect disappears when the cues are from the same modality [36], consistent with the model's predictions.

Figure 4



Two generative processes for latent causes. **(a)** In the Chinese restaurant process, a single latent cause generates each observed datapoint. Here each latent cause is associated with a 'table' and datapoints are associated with 'customers' (colored squares, labeled by letters). Customers are probabilistically seated at tables, indicating their latent cause assignment. Another representation of the seating arrangement is shown on the right: Shaded cells indicate that a customer is seated at a particular table. **(b)** In the Indian buffet process, multiple latent causes generate the different observations that comprise each datapoint. Here each latent cause is associated with a 'dish,' and customers can sample multiple dishes.

Conclusions

The problem of state space discovery in reinforcement learning has, until recently, been neglected by computational theories. This picture is now rapidly changing, with new models and new data to elucidate the principles governing state discovery. We have reviewed recent developments that illustrate the richness of these principles.

While we have focused on Pavlovian conditioning, the same principles of state discovery, and structure learning more generally, have been invoked across numerous cognitive domains [42–45,18,20,10,46••,47•]. For example, the latent cause framework provides insight into the conditions under which old memories are updated and new memories are formed [10]. Similarly, the process of forming and updating memories can be used to understand the unsupervised discovery of visual categories [18]. These applications suggest that structure learning may function as a core computational system that is shared across domains, with the hippocampus and orbitofrontal cortex potentially playing a central role [48,8,10,49,50].

While it is still unclear how structure learning is implemented neurally, the available data suggest that the hippocampus plays an important role in discovering latent causes, and in structure learning more generally [51]. For example, lesions of the hippocampus impair an animal's ability to distinguish between different contexts, resulting in reduced fear recovery following Pavlovian extinction (see [52] for a review). Some of these effects of hippocampal lesions can be captured by forcing the latent cause model to assign all observations to a single latent cause [8]. Another example comes from work on spatially tuned neurons in the hippocampus ('place cells'), which 'remap' (i.e., reconfigure their spatial receptive fields) when the context changes dramatically, possibly reflecting a form of structure learning sensitive to the statistics of environmental dynamics [48,10].

Wilson and colleagues [50] have suggested that the orbitofrontal cortex is crucial for representing rich structure in learning tasks. Animals with an orbitofrontal lesion perform a task as though they are using the same reinforcement learning algorithms as non-lesioned animals, but operating on an impoverished representation of the task structure — i.e., a representation lacking latent causes [49]. We speculate that the orbitofrontal cortex may be the repository for consolidated representations of the causal structure of a task, initially acquired by the hippocampus.

Finally, Crossley and colleagues [53] have suggested that tonically active cholinergic interneurons in the striatum act as gates on procedural learning, possibly functioning as a mechanism by which inferences about latent causes modulate reinforcement learning in the basal ganglia. This idea is consistent with recent rodent experiments showing that disruption of cholinergic signaling in the striatum interferes with adaptation to changes in action-outcome contingencies [54••], an effect that has been interpreted in terms of latent causal inference [55]. Future work will hopefully reveal a deeper understanding of the interplay between these brain regions and the underlying circuit mechanisms that implement structure learning.

Conflict of interest statement

Nonething declared.

Acknowledgements

This research was supported in part by the National Institute of Mental Health under Award Number R01MH098861, an Alfred P. Sloan Research Fellowship (YN), and a grant from the John Templeton Foundation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or the John Templeton Foundation.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Polo M: *The Travels of Marco Polo*. JM Dent & Sons; 1918.

2. Braun DA, Mehring C, Wolpert DM: **Structure learning in action**. *Behav Brain Res* 2010, **206**:157-165.

3. Gershman SJ, Niv Y: **Learning latent structure: carving nature at its joints**. *Curr Opin Neurobiol* 2010, **20**:251-256.

4. McCallum RA: **Hidden state and reinforcement learning with instance-based state identification**. *IEEE Trans Syst Man Cybernet, Part B* 1996, **26**:464-473.

5. Givan R, Dean T, Greig M: **Equivalence notions and model minimization in Markov decision processes**. *Artif Intell* 2003, **147**:163-223.

6. Roy N, Gordon GJ, Thrun S: **Finding approximate POMDP solutions through belief compression**. *J Artif Intell Res* 2005, **23**:1-40.

7. •• Courville AC, Daw ND, Touretzky DS: **Bayesian theories of conditioning in a changing world**. *Trends Cogn Sci* 2006, **10**:294-300.

The paper that first popularized latent cause accounts of classical conditioning phenomena, including partial reinforcement, latent inhibition, and overshadowing.

8. Gershman DJ, Blei DM, Niv Y: **Context, learning, and extinction**. *Psychol Rev* 2010, **117**:197-209.

9. Gershman SJ, Niv Y: **Exploring a latent cause theory of classical conditioning**. *Learn Behav* 2012, **40**:255-268.

10. Gershman DJ, Radulescu A, Norman KA, Niv Y: **Statistical computations underlying the dynamics of memory updating**. *PLOS Comput Biol* 2014, **10**:e1003939.

11. Gershman SJ, Hartley C: **Individual differences in learning predict the return of fear**. *Learn Behav* 2015, **43**:243-250.

12. • Lloyd K, Leslie DS: **Context-dependent decision-making: a simple Bayesian model**. *J R Soc Interface* 2013, **10**:20130069.

A Bayesian nonparametric model of instrumental learning that explains how animals infer the occurrence of changes in instrumental contingencies (which function as latent 'contexts'). This model can account for a variety of instrumental learning phenomena, including spontaneous recovery, the partial reinforcement extinction effect, and the overtraining reversal effect.

13. Soto FA, Gershman SJ, Niv Y: **Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization**. *Psychol Rev* 2014, **121**:526-558.

14. Austerweil JL, Gershman SJ, Tenenbaum JB, Griffiths TL: **Structure and flexibility in Bayesian models of cognition**. In *Oxford Handbook of Computational and Mathematical Psychology*. Edited by Busemeyer JR, Townsend JT, Wang Z, Eidels A. Oxford: Oxford University Press; 2015.

15. Rescorla RA, Wagner AR: **A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement**. In *Classical Conditioning II: Current Research and Theory*. Edited by Black AH, Prokasy WF. New York, NY: Appleton-Century-Crofts; 1972:64-99.

16. Bouton ME: **Context and behavioral processes in extinction**. *Learn Memory* 2004, **11**:485-494.

17. Chater N, Vitányi P: **Simplicity: a unifying principle in cognitive science?** *Trends Cogn Sci* 2003, **7**:19-22.

18. Gershman SJ, Niv Y: **Perceptual estimation obeys Occam's razor**. *Front Psychol* 2013, **4**.

19. Gershman SJ, Blei DM: **A tutorial on Bayesian nonparametric models**. *J Math Psychol* 2012, **56**:1-12.

20. Gershman SJ, Jones CE, Norman KA, Monfils M-H, Niv Y: **Gradual extinction prevents the return of fear: implications for the discovery of state**. *Front Behav Neurosci* 2013, **7**.

21. Rescorla RA: **Spontaneous recovery**. *Learn Memory* 2004, **11**:501-509.

22. Rescorla RA, Heth CD: **Reinstatement of fear to an extinguished conditioned stimulus**. *J Exp Psychol: Anim Behav Process* 1975, **1**:88-96.

23. Bouton ME, Bolles RC: **Role of conditioned contextual stimuli in reinstatement of extinguished fear**. *J Exp Psychol: Anim Behav Process* 1979, **5**:368-378.

24. Miller RR, Matzel LD: **Retrieval failure versus memory loss in experimental amnesia: definitions and processes**. *Learn Memory* 2006, **13**:491-497.

25. Kagerer FA, Contreras-Vidal JL, Stelmach GE: **Adaptation to gradual as compared with sudden visuo-motor distortions**. *Exp Brain Res* 1997, **115**:557-561.

26. Pekny SE, Criscimagna-Hemminger SE, Shadmehr R: **Protection and expression of human motor memories**. *J Neurosci* 2011, **31**:13829-13839.

27. Taylor JA, Wojaczynski GJ, Ivry RB: **Trial-by-trial analysis of intermanual transfer during visuomotor adaptation**. *J Neurophysiol* 2011, **106**:3157-3172.

28. Wallis G, Bülthoff HH: **Effects of temporal association on recognition memory**. *Proc Natl Acad Sci U S A* 2001, **98**:4800-4804.

29. Preminger S, Sagi D, Tsodyks M: **The effects of perceptual history on memory of visual objects**. *Vis Res* 2007, **47**:965-973.

30. Preminger S, Blumenfeld B, Sagi D, Tsodyks M: **Mapping dynamic memories of gradually changing objects**. *Proc Natl Acad Sci U S A* 2009, **106**:5371-5376.

31. Pearce JM: **A model for stimulus generalization in Pavlovian conditioning**. *Psychol Rev* 1987, **94**:61-73.

32. Wagner AR: **Context-sensitive elemental theory**. *Q J Exp Psychol: Sect B* 2003, **56**:7-29.

33. Harris JA: **Elemental representations of stimuli in associative learning**. *Psychol. Rev.* 2006, **113**:584-605.

34. Melchers KG, Shanks DR, Lachnit H: **Stimulus coding in human associative learning: flexible representations of parts and wholes**. *Behav Process* 2008, **77**:413-427.

35. Pearce JM: **Similarity and discrimination: a selective review and a connectionist model**. *Psychol Rev* 1994, **101**:587-607.

36. Kehoe EJ, Horne AJ, Horne PS, Macrae M: **Summation and configuration between and within sensory modalities in classical conditioning of the rabbit**. *Anim Learn Behav* 1994, **22**:19-26.

37. Rescorla RA, Coldwell SE: **Summation in autoshaping**. *Anim Learn Behav* 1995, **23**:314-326.

38. Glautier S: **Spatial separation of target and competitor cues enhances blocking of human causality judgements**. *Q J Exp Psychol: Sect B* 2002, **55**:121-135.

39. Livesey EJ, Boakes RA: **Outcome additivity, elemental processing and blocking in human causality judgements**. *Q J Exp Psychol: Sect B* 2004, **57**:361-379.

40. Shepard RN: **Toward a universal law of generalization for psychological science**. *Science* 1987, **237**:1317-1323.

41. Tenenbaum JB, Griffiths TL: **Generalization, similarity, and Bayesian inference**. *Behav Brain Sci* 2001, **24**:629-640.

42. Anderson JR: **The adaptive nature of human categorization**. *Psychol Rev* 1991, **98**:409-429.

43. Goldwater S, Griffiths TL, Johnson M: **A Bayesian framework for word segmentation: exploring the effects of context**. *Cognition* 2009, **112**:21-54.

44. Sanborn AN, Griffiths TL, Navarro DJ: **Rational approximations to rational models: alternative algorithms for category learning**. *Psychol Rev* 2010, **117**:1144-1167.

45. Kemp C, Tenenbaum JB, Niyogi S, Griffiths TL: **A probabilistic model of theory formation**. *Cognition* 2010, **114**:165-196.

46. •• Collins AGE, Frank MJ: **Cognitive control over learning: creating, clustering, and generalizing task-set structure**. *Psychol Rev* 2013, **120**:190-229.
A Bayesian nonparametric model of task set clustering, which provides insight into how and why humans spontaneously cluster tasks together. The authors show that the abstract computations of the Bayesian model can be mimicked by a biologically realistic neural network based on basal ganglia and prefrontal circuitry.

47. • Collins AGE, Cavanagh JF, Frank MJ: **Human EEG uncovers latent generalizable rule structure during learning**. *J Neurosci* 2014, **34**:4677-4685.
An EEG study demonstrating that inferences about latent rule structure are encoded in lateral prefrontal cortex. This neural signature is predictive of individuals' subsequent generalization of rule structure to novel contexts.

48. Fuhs MC, Touretzky DS: **Context learning in the rodent hippocampus**. *Neural Comput* 2007, **19**:3173-3215.

49. Takahashi YK, Roesch MR, Wilson RC, Toreson K, O'Donnell P, Niv Y, Schoenbaum G: **Expectancy-related changes in firing of dopamine neurons depend on orbitofrontal cortex**. *Nat Neurosci* 2011, **14**:1590-1597.

50. Wilson RC, Takahashi YK, Schoenbaum G, Niv Y: **Orbitofrontal cortex as a cognitive map of task space**. *Neuron* 2014, **81**:267-279.

51. Aggleton JP, Sanderson DJ, Pearce JM: **Structural learning and the hippocampus**. *Hippocampus* 2007, **17**:723-734.

52. Ji J, Maren S: **Hippocampal involvement in contextual modulation of fear extinction**. *Hippocampus* 2007, **17**:749-758.

53. Crossley MJ, Ashby FG, Maddox WT: **Erasing the engram: the unlearning of procedural skills**. *J Exp Psychol: Gen* 2013, **142**:710-741.

54. •• Bradfield LA, Bertran-Gonzalez J, Chieng B, Balleine BW: **The thalamostriatal pathway and cholinergic control of goal-directed action: interlacing new with existing learning in the striatum**. *Neuron* 2013, **79**:153-166.
This study used an heroic combination of lesion and pharmacological methods to show that cholinergic signaling in the dorsomedial striatum is crucial for encoding changes in action-outcome contingencies, possibly reflecting a latent causal inference process.

55. Schoenbaum G, Stalnaker TA, Niv Y: **How did the chicken cross the road? With her striatal cholinergic interneurons, of course**. *Neuron* 2013, **79**:3-6.

56. Aldous D: **Exchangeability and related topics**. *École d'Été de Probabilités de Saint-Flour XIII*. Berlin: Springer; 1985, 1-198.

57. Griffiths TL, Ghahramani Z: **The Indian buffet process: an introduction and review**. *J Mach Learn Res* 2011, **12**:1185-1224.

58. Austerweil JL, Griffiths TL: **A rational model of the effects of distributional information on feature learning**. *Cogn Psychol* 2011, **63**:173-209.

59. Austerweil JL, Griffiths TL: **A nonparametric Bayesian framework for constructing flexible feature representations**. *Psychol Rev* 2013, **120**:817-851.

60. Yildirim I, Jacobs RA: **A rational analysis of the acquisition of multisensory representations**. *Cogn Sci* 2012, **36**:305-332.

61. • Austerweil JL: **Testing the psychological validity of cluster construction biases**. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. 2014.
This paper demonstrated how human priors over clusterings (or latent causes) can be measured empirically. A comparison of different priors supported the Chinese restaurant process.

62. Sanborn AN, Griffiths TL, Shiffrin RM: **Uncovering mental representations with Markov chain Monte Carlo**. *Cogn Psychol* 2010, **60**:63-106.