



ELSEVIER

Available online at [www.sciencedirect.com](http://www.sciencedirect.com)
**Current Opinion in  
Neurobiology**

## Learning latent structure: carving nature at its joints

Samuel J Gershman and Yael Niv

Reinforcement learning (RL) algorithms provide powerful explanations for simple learning and decision-making behaviors and the functions of their underlying neural substrates.

Unfortunately, in real-world situations that involve many stimuli and actions, these algorithms learn pitifully slowly, exposing their inferiority in comparison to animal and human learning. Here we suggest that one reason for this discrepancy is that humans and animals take advantage of structure that is inherent in real-world tasks to simplify the learning problem. We survey an emerging literature on 'structure learning' — using experience to infer the structure of a task — and how this can be of service to RL, with an emphasis on structure in perception and action.

### Address

Princeton Neuroscience Institute & Psychology Department,  
Princeton University, United States

Corresponding author: Gershman, Samuel J ([sjgershm@princeton.edu](mailto:sjgershm@princeton.edu))  
and Niv, Yael ([yael@princeton.edu](mailto:yael@princeton.edu))

**Current Opinion in Neurobiology** 2010, **20**:251–256

This review comes from a themed issue on  
Cognitive neuroscience  
Edited by Earl Miller and Liz Phelps

Available online 11th March 2010

0959-4388/\$ – see front matter  
© 2010 Elsevier Ltd. All rights reserved.

DOI [10.1016/j.conb.2010.02.008](https://doi.org/10.1016/j.conb.2010.02.008)

### Introduction

A major breakthrough in understanding how animals and humans learn to choose actions in order to obtain rewards and avoid punishments has been the recent framing of trial and error learning (conditioning) in the computational terms of reinforcement learning (RL [1]). RL is a powerful framework that has been instrumental in describing how the basal ganglia learn to evaluate different situations (states) in terms of their future expected rewards, and in suggesting that dopaminergic activity conveys errors in the prediction of reward which are crucial for optimal learning [2,3]. However, in their most basic form, RL algorithms are extremely limited when applied to real-world situations: when confronted with more than a handful of states and possible actions, learning becomes pitifully slow, necessitating thousands of trials to learn what animals can learn in a mere tens or hundreds.

One reason that animals and humans can rapidly learn new problems is perhaps because they take advantage of

the high degree of *structure* of natural tasks [4\*\*]. This starts with perception: although our brains are confronted with undifferentiated sensory input, our conscious perception is highly structured. As Ernst Mach remarked, 'We do not see optical images in an optical space, but we perceive the bodies round about us in their many and sensuous qualities' [5]. Helmholtz referred to the perception of such qualities as 'unconscious inference' [6], emphasizing the inductive nature of the process. The brain must go beyond the available sensory data to make inferences about the hidden structure of the world. Identifying structure in sensory data allows animals and humans to focus their attention on those objects and events that are key in obtaining reinforcement, and learn only about these while ignoring other irrelevant stimuli. Further structure lies in the way *actions* affect the environment. This can be utilized to 'divide and conquer,' and to decompose tasks to smaller (and more manageable) components.

To solve the problem of inferring structure from observations, one must determine which of a (possibly large) number of possible structures is most likely to capture correctly the causal structure of the environment. Each structure represents a hypothesis about the set of latent causes that generate observed data (in the case of RL: the latent causes of reinforcement; **Box 1**). For a given structure, one must also estimate the parameters relating the different variables in the structure, that is, the model of how it gives rise to observations. The problem of how, given an assumed structure, the brain can infer the particular latent cause that generated observations has been extensively studied in cognitive neuroscience and psychology (e.g. [7–9]). However, the problem of how the relevant causal structure is identified, which we refer to as *structure learning*, remains mysterious. The difficulty of the structure learning problem is highlighted by the fact that in many domains the number of possible structures is essentially limitless.

Here we review recent experimental and theoretical research that has begun to elucidate how the brain solves the structure learning problem, and how RL mechanisms can take advantage of such structure to facilitate learning. We first outline a normative computational approach to this problem, and then describe research that tests some of these ideas. We focus on latent structures in two different components of RL problems: perception and action.

### A normative framework

The canonical description of a decision-making problem in RL has four components: firstly, a set of *states* of the

**Box 1 Latent structures as a state space for reinforcement learning**

Graphical models provide a useful framework for describing causal structures. These depict the statistical relations between latent and observed variables. For instance, Figure 1 shows three possible relations between tones, shocks, and latent variables in a classical conditioning experiment. In RL, the emphasis is on the possible causes of reinforcement — the goal of the animal is to infer whether reinforcement will or will not occur, based on the currently observed variables and past experience. Through Bayesian inference one can use the co-occurrence of observable events to infer which structure is the most plausible, as well as what are the current settings of different latent variables. Let us suppose that before observing any data, your belief about the hidden structure *S* of your environment is encoded by a *prior distribution* over possible structures,  $P(S)$  (for instance an equal probability  $p = 1/3$  over the three structures in Figure 1). This expresses how likely you think it is that each structure accurately describes the environment *a priori*. After observing some sensory data *D*, the statistically correct way to update this belief is given by Bayes' rule:

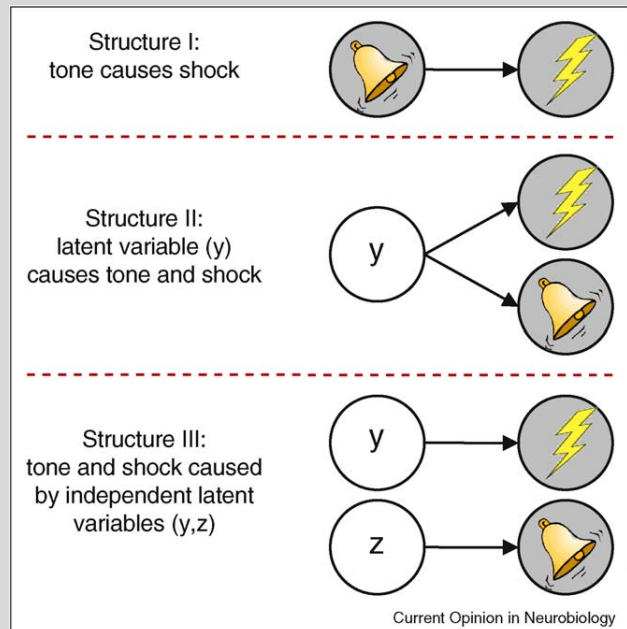
$$P(S|D) = \frac{P(D|S)P(S)}{\sum_{S'} P(D|S')P(S')} \quad (1)$$

$P(D|S)$  is known as the *likelihood* and expresses how likely it is that sensory data *D* was generated by structure *S*. The end result,  $P(S|D)$ , is the *posterior distribution* over structures given the observed data. This is the best estimate of the probability that each structure accurately describes the environment *a posteriori*. In our case, observing both trials in which tones and shocks co-occur and trials in which tones occur without shocks mitigates against structure I. Parsimony and the 'automatic Occam's razor' inherent in Bayesian inference [52] further tilts the balance towards structure II which assumes fewer latent variables.

In the context of reinforcement learning, an inferred structure (for instance, the one with the highest probability) can then be used to define a *state space* — the set of variables that have a causal relationship to reinforcement, and thus must be learned about and evaluated. These can include observed sensory variables, as well as unobserved (but inferred) variables. In this way, RL algorithms can operate over a small subset of variables that are causally related to reinforcement, and are specified by the inferred structure. In many cases, this should substantially improve adaptive behavior, since

rewards and punishments are rarely caused by all observable variables and only by these.

**Figure 1**



The graphical models of three possible causal relationships between variables in a classical conditioning experiment. By convention, observed variables are represented by shaded nodes and unshaded nodes represent unobserved (latent) variables. Arrows represent probabilistic dependencies. The parameters of a model, for instance, structure II, define the probability of each of the nodes taking on a value (e.g. absence or presence of shock) given each setting of its parent nodes (e.g. when  $y =$  acquisition or when  $y =$  extinction).

environment, each comprised of different stimuli; secondly, a set of *actions* that can be taken at these states; thirdly, a *transition function* denoting how actions cause the world to transition from one state to another; and fourthly, a *reward function* denoting the immediate reward that is available at each state. The goal of the agent is to learn a *policy* of actions at each state, that will maximize overall rewards. Model-based RL algorithms concentrate on finding an optimal policy when the transition function and the reward function are known (such as when playing chess). Model-free RL algorithms do not assume such knowledge, but rather assign values to different actions at different states through learning (such as when learning to navigate a maze) [10]. These values represent the sum of future rewards that can be expected if a particular action is taken at a particular state. Given such values, decision-making is easy: of the available actions in the current state, one should take the action with the highest value.

Importantly, both model-based and model-free RL assume that the set of states and the set of actions are

provided to the learning agent. Unfortunately, in real-world tasks this is often not the case. Rather, the relevant set of states and actions must also be inferred or learned from observations, as the space of all possible states and actions is too large to be of practical use. As an example, when learning to escape predators one must learn the values of running left, right, straight ahead, hiding in the bushes, etc. Clearly, assigning values to scratching one's nose or wiggling one's toes are irrelevant, although these are perfectly valid actions in this situation. How do animals and humans reduce the space of states and actions to a manageable (and learnable) subset?

One approach is grounded in Bayesian probability theory, which specifies how to update probabilistic beliefs about causal structures in light of new data. Through Bayesian inference one can use observed data to update an estimate of the probability that each of several possible structures accurately describes the environment (see Box 1). For example, in a typical classical conditioning experiment, an animal receives a series of tones and

shocks. Rather than (erroneously) assuming that either tones cause shocks or shocks cause tones, we suggest that animals attempt to learn about the latent causes that generate both tones and shocks (structures II and III in Box 1). Intuitively, in this case the true latent cause is the stage of the experiment, as defined by the experimenter (e.g. acquisition and extinction). If the animal knew what stage of the experiment it was in, it could perfectly predict whether it will be shocked or not following a tone. Moreover, learning about relationships between shocks and tones should be restricted such that experience is only averaged over trials that can be attributed to the same latent cause. Thus after 20 acquisition trials in which tones were followed by shocks, and 20 extinction trials in which tones were not followed by shock, the rat should not predict a shock with 50% probability, but rather predict a shock with (near) certainty if it infers that the current trial is a training trial, or predict the absence of a shock if the current trial is an extinction trial.

This perspective represents a significant departure from classical learning theory [11], which imputes to the animal the assumption that sensory variables are directly predictive of reinforcement. In contrast, we suggest that the animal use a different internal model in which sensory variables and reinforcements are both generated by latent variables. Under this assumption, the task of predicting future reinforcement requires a system for performing inference over these latent variables [12–15]. We will return in later sections to the question of what brain areas might be responsible for this inference process.

Although the Bayesian framework provides the optimal solution to the structure learning problem, in most cases computing this solution is intractable. However, the solution can be approximated, for example by representing the posterior probability over structures with a set of samples [16–19]. Another possible approximation is to use policy search methods [20], that is, to avoid representing the posterior distribution altogether, and instead try to find a good behavioral policy without knowledge of the environment's latent structure. As we describe in the next section, it is possible to interpret an influential family of neural RL models as a form of policy search over not only an action space, but also a state space.

### Structure in perception

In this section, we focus on two types of perceptual structure that can play a role in RL. The first arises in situations where multiple sensory inputs are coupled by a common latent cause, as was described above in the case of classical conditioning, and will be elaborated further below. Recent work has demonstrated that in (instrumental) bandit tasks in which a latent variable couples rewards, human behavior is consistent with Bayesian inference over the underlying coupling structure [21\*\*]. The second type of structure arises in situations where

only a subset of sensory inputs are causally related to reinforcement, such as in tasks that require selective attention to particular stimulus dimensions [23,24] or to particular stimuli in a sequence [25,26].

Continuing our example from the previous section, one perplexing observation (from the perspective of classical learning theory) is that extinction training (i.e. presenting a tone without the previously associated shock) does not result in unlearning the original association between tone and shock. Many studies have shown that returning the animal to the original acquisition context (aka 'renewal') [27,28], presenting an unpaired shock (reinstatement) [29,30], or simply waiting 48 hours before testing the animal again (spontaneous recovery) [31,32] are all sufficient to return the animal's original fear response to the tone to nearly pre-extinction levels, suggesting that the presentation of the tone still causes the animal to predict an imminent shock. This observation is rendered less perplexing if one considers the animal's conditioned response as reflecting its inferences about the latent structure of the environment at each stage of the experiment. Specifically, if the animal assumes that the pattern of tones, shocks, and contextual cues during acquisition trials were generated by one latent cause, and the distinct pattern of tones and contextual cues during extinction trials were generated by a different latent cause, then returning the animal to the acquisition context will naturally lead it to infer that the 'acquisition' cause is once again active, and hence to predict shocks (and exhibit fear). Similarly, Bouton has argued that time itself is treated by animals as a contextual cue [33], providing one explanation for spontaneous recovery of fear as a result of the mere passage of time.

Recently, Redish *et al.* [34\*\*] presented a novel computational theory of these renewal effects, synthesizing ideas from RL and neural network models. They postulated a state-splitting mechanism that creates new states when the perceptual statistics alter radically. In their model, the affinity for state-splitting is modulated by dopamine, such that tonically negative prediction errors result in a higher probability of creating a new state. This modulatory process was motivated by the idea that new states should be created when the current set of states proves inadequate for adaptive behavior. The new set of states is then used as input to a standard RL algorithm. We have elaborated upon this proposal, showing how it can be interpreted in terms of the normative structure learning framework proposed in the previous section [12]. In addition, we suggest that the hippocampus may play a particularly important role in learning about the structure of these tasks. Pre-training lesions of the hippocampus eliminate renewal [35], an effect we explain in terms of the animal's impaired ability to infer the existence of new latent causes. In essence, according to our theory, hippocampal lesions compel the animal to attribute all its

observations to a single latent cause. Intriguingly, young rats appear to display the same lack of renewal as hippocampal-lesioned adult rats [36<sup>•</sup>], suggesting that structure learning is a late-developing process, possibly requiring hippocampal maturity.

So far we have been dealing with *static* structure contained in the mosaic of sensory inputs; however, certain tasks require inferring *dynamic* structure from a sequence of sensory inputs. For example, in a sequential alternation task, an animal must alternate its response to the same sensory cues on each trial. To perform this task correctly, the animal must base its decision not just on its immediate sensory observations, but also on information stored in its memory [37<sup>•</sup>]. More complex tasks may require storing different pieces of information from arbitrarily far into the past. We interpret this as a structure learning problem: among the past sensory inputs, which of them is predictive of reward?

The 'gating framework' [38–40<sup>•</sup>,26] is one computational solution to this problem, proposing that dopaminergic prediction error signals control the contents of working memory in prefrontal cortex. Because the dopamine signal to the basal ganglia and the prefrontal cortex encodes the discrepancy between predicted and observed rewards [2,3], the prefrontal gating mechanism will tend to add the current sensory input to working memory when it coincides with unexpected rewards, and remove the contents of working memory that are unpredictable of reward. This model is consistent with the finding that prefrontal dopamine dysfunction appears to be responsible for aberrant attentional control in schizophrenia [38,41–43].

Recently, Todd *et al.* [44] showed how the gating framework can be interpreted as a policy search algorithm [45]. The state space for this policy has two components: the current sensory inputs and an internal memory register. The policy maps states onto two different kinds of actions: motor actions that directly affect the external world, and gating actions that update the contents of the memory register. RL is then used to find the policy that maximizes reward. One appealing property of this proposal is that it allows the animal to learn about reward structure without representing an explicit internal model of the world, or a posterior distribution over possible causal structures. Other recent variants of the gating framework have explored how it can support performance in a variety of complex tasks [37<sup>•</sup>,46], but direct experimental evidence supporting its central claims remains scant.

### Structure in actions

Just as animals are provided with a surfeit of perceptual inputs, they must also contend with an overwhelming bounty of motor effectors. Although such high degree of

motor control is in principle useful, it presents a formidable learning problem, known in machine learning and engineering as the *curse of dimensionality* [47]: the number of possible effector combinations grows exponentially with the number of effectors, and therefore an unrealistic amount of experience would be needed to accurately estimate and compare the value of each of these combinations. The curse of dimensionality can be alleviated if one is willing to make certain assumptions about the structure of the environment or the structure of correct motor strategies [48<sup>•</sup>,49<sup>••</sup>]. In particular, assuming that the value of an effector combination decomposes into a sum of effector-specific components drastically reduces the amount of experience required for accurate value estimates. Structure learning thus takes center stage in making action evaluation tractable.

Two recent functional magnetic resonance imaging studies have begun to examine how the human brain takes advantage of structure in the space of actions. Gershman *et al.* [50] designed a RL task in which subjects were asked to make two choices simultaneously, one with each hand, after which probabilistic reward feedback was presented. They showed behaviorally that when the rewards decomposed into hand-specific components, subjects exploited this structure to guide learning. Furthermore, prediction error signals in the ventral striatum (the main afferent of midbrain dopamine) displayed a corresponding fractionation, with the prediction error component for each hand correlating preferentially with the contralateral hemisphere. Value signals in the intraparietal sulcus also displayed hemispheric fractionation. In a related decision-making experiment, Palminteri *et al.* [51] cued subjects to make choices either with the left or with the right hand. They showed that hand-specific values were preferentially correlated with the contralateral ventral prefrontal cortex, consistent with the idea that the brain's RL system exploits structure in the space of actions rather than learning a single value function over the space of effector combinations.

### Conclusions

Although computational models of RL have greatly enriched our understanding of learning and decision-making in the brain, they have often rested upon naive assumptions about the representations over which learning operates. In particular, many studies have assumed that states are represented by simple perceptual primitives, and actions are represented by monolithic motor responses. If we have learned anything from machine learning, it is that these assumptions will not 'scale up' to real-world problems, and the very fact that humans and animals are able to learn effectively in highly complex environments suggests that these assumptions are not psychologically plausible. We have reviewed several lines of research that begin to paint a portrait of a more sophisticated learning system, based on interactions



between the basal ganglia, prefrontal cortex and hippocampus, that can deal with the challenges of a complex world by inferring its latent structure.

## Acknowledgements

We are grateful to Michael Todd for many helpful discussions. SJG is funded by a Quantitative Computational Neuroscience training grant from the National Institute of Mental Health.

## 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. Sutton RS, Barto AG: *Reinforcement Learning: An Introduction* MIT Press; 1998.
  2. Schultz W, Dayan P, Montague PR: **A neural substrate of prediction and reward.** *Science* 1997, **275(5306)**:1593.
  3. Houk JC, Adams JL, Barto AG: **A model of how the basal ganglia generate and use neural signals that predict reinforcement.** *Models of Information Processing in the Basal Ganglia.* 1995:249-270.
  4. Kemp C, Tenenbaum JB: **Structured statistical models of inductive reasoning.** *Psychol Rev* 2009, **116(1)**:20-58.  
•• An elegant demonstration of how structured knowledge influences human reasoning in a large variety of domains.
  5. Mach E: *The analysis of sensations.* 1897.
  6. Von Helmholtz H: *Handbuch der Physiologischen Optik.* Voss; 1867.
  7. Knill DC, Richards W: *Perception as Bayesian inference.* Cambridge University Press; 1996.
  8. Doya K, Ishii S, Alexandre P, Rao RPN: *Bayesian Brain.* MIT Press; 2007.
  9. Chater N, Oaksford M: *The Probabilistic Mind: Prospects for Bayesian Cognitive Science* Oxford University Press; 2008.
  10. Daw ND, Niv Y, Dayan P: **Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control.** *Nat Neurosci* 2005, **8(12)**:1704-1711.
  11. 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.
  12. Gershman SJ, Blei DM, Niv Y: **Context, learning, and extinction.** *Psychol Rev* 2010, **117(1)**:197-209.
  13. Courville Aaron C, Daw Nathaniel D, Touretzky David S: **Similarity and discrimination in classical conditioning: a latent variable account.** In *Advances in Neural Information Processing Systems, vol 17.* Edited by Saul Lawrence Kwec, Weiss Yairi., Bottou Léonect. Cambridge, MA: MIT Press; 2002:313-320.
  14. Courville Aaron C, Daw Nathaniel D, Gordon Geoffrey J, Touretzky David S: **Model uncertainty in classical conditioning.** In *Advances in Neural Information Processing Systems, vol 16.* Edited by Thrun Sebastianbsi, Saul Lawrence Kwec, Schölkopf Bernhardhr. Cambridge, MA: MIT Press; 2004.
  15. Daw ND, Courville AC, Touretzky DS: **Representation and timing in theories of the dopamine system.** *Neural Comput* 2006, **18(7)**:1637-1677.
  16. Daw N, Courville A: **The pigeon as particle filter.** *Adv Neural Inform Process Syst* 2008, **20**:369-376.  
• This paper proposes that apparently abrupt and unstable learning in animals can be the result of their using a crude approximation of the posterior distribution over causal relationships between observed events.
  17. Yi MSK, Steyvers M, Lee M: **Modeling human performance in restless bandits with particle filters.** *J Problem Solving* 2009, **2(2)**:5.
  18. Vul E, Frank M, Alvarez G, Tenenbaum J: **Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model.** In *Advances in Neural Information Processing Systems, vol 22.* Edited by Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A. 2009:1955-1963.
  19. Gershman S, Vul E, Tenenbaum J: **Perceptual multistability as Markov Chain Monte Carlo inference.** In *Advances in Neural Information Processing Systems vol 22.* Edited by Bengio Y, Schuurmans D, Lafferty J, Williams CKI, Culotta A. 2009:611-619.
  20. Williams RJ: **Simple statistical gradient-following algorithms for connectionist reinforcement learning.** *Mach Learn* 1992, **8(3)**:229-256.
  21. Acuna D, Schrater PR: **Structure learning in human sequential decision-making.** In *Advances in Neural Information Processing Systems, vol 21.* Edited by Koller D, Schuurmans D, Bengio Y, Bottou L. 2009:1-8.  
•• Humans' choices often seem suboptimal even in simple bandit tasks. Recent work (e.g. [22]) has suggested that choice behavior might be optimal, but for a different problem than the one designed by the experimenter. For instance, matching behavior can be understood given the assumption of a changing environment. Acuna and Schrater suggest another possibility, namely that humans may be learning the structure of the task concurrently with its parameters. Focusing on structure in rewards, they provide evidence that people perform structure learning in a 2-bandit task in which the bandits might be independent or coupled. The effects of structure learning in this task are evident in the resulting exploration policies.
  22. Timothy EJ, Behrens W, Mark, Woolrich E, Mark, Walton FS, Matthew, Rushworth: **Learning the value of information in an uncertain world.** *Nat Neurosci* 2007, **10(9)**:1214-1221.
  23. Owen AM, Roberts AC, Hodges JR, Robbins TW: **Contrasting mechanisms of impaired attentional set-shifting in patients with frontal lobe damage or Parkinson's disease.** *Brain* 1993, **116(5)**:1159.
  24. Milner B: **Effects of different brain lesions on card sorting: the role of the frontal lobes.** *Arch Neurol* 1963, **9(1)**:90.
  25. Barch DM, Braver TS, Nystrom LE, Forman SD, Noll DC, Cohen JD: **Dissociating working memory from task difficulty in human prefrontal cortex.** *Neuropsychologia* 1997, **35(10)**:1373-1380.
  26. O'Reilly RC, Frank MJ: **Making working memory work: a computational model of learning in the prefrontal cortex and basal ganglia.** *Neural Comput* 2006, **18(2)**:283-328.
  27. Bouton ME, Bolles RC: **Contextual control of the extinction of conditioned fear.** *Learn Motiv* 1979, **10(4)**:445-466.
  28. Nakajima S, Tanaka S, Urushihara K, Imada H: **Renewal of extinguished lever-press responses upon return to the training context.** *Learn Motiv* 2000, **31(4)**:416-431.
  29. Rescorla RA, Heth CD: **Reinstatement of fear to an extinguished conditioned stimulus.** *J Exp Psychol* 1975, **1**:88-96.
  30. Bouton ME, Bolles RC: **Role of conditioned contextual stimuli in reinstatement of extinguished fear.** *J Exp Psychol* 1979, **5**:368-378.
  31. Robbins SJ: **Mechanisms underlying spontaneous recovery in autoshaping.** *J Exp Psychol: Anim Behav Process* 1990, **16(3)**:235-249.
  32. Rescorla RA: **Spontaneous recovery.** *Learn Mem* 2004, **11(5)**:501.
  33. Bouton ME: **Context, time, and memory retrieval in the interference paradigms of Pavlovian learning.** *Psychol Bull* 1993, **114**:80-180.
  34. Redish AD, Jensen S, Johnson A, Kurth-Nelson Z: **Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling.** *Psychol Rev* 2007, **114(3)**:784-805.

An extension of reinforcement learning ideas to account for structure learning. The authors use this model to explain a number of experimental findings in instrumental conditioning and addictive behavior.

35. Ji J, Maren S: **Electrolytic lesions of the dorsal hippocampus disrupt renewal of conditional fear after extinction.** *Learn Mem* 2005, **12(3)**:270.
36. Yap CSL, Richardson R: **Extinction in the developing rat: an examination of renewal effects.** *Dev Psychobiol* 2007, **49(6)**:565-575.  
This study showed that context-specificity of conditioning does not emerge in developing rats until later in life. In fact, young rats seem to learn just as was originally envisioned by Rescorla and Wagner [11] and straightforward RL [1].
37. Zilli EA, Hasselmo ME: **Modeling the role of working memory and episodic memory in behavioral tasks.** *Hippocampus* 2008, **18(2)**:193.  
A computational exploration of how the brain can make use of memory mechanisms to solve reinforcement learning problems.
38. Braver TS, Barch DM, Cohen JD: **Cognition, and control in schizophrenia: a computational model of dopamine and prefrontal function.** *Biol Psychiatry* 1999, **46(3)**:312-328.
39. Braver TS, Cohen JD: **On the control of control: the role of dopamine in regulating prefrontal function and working memory.** *Control of Cognitive Processes: Attention and Performance XVIII*. 2000:713-737.
40. Rougier NP, Noelle DC, Braver TS, Cohen JD, O'Reilly RC: **Prefrontal cortex and flexible cognitive control: rules without symbols.** *Proc Natl Acad Sci U S A* 2005, **102(20)**:7338.  
One of the first attempts to combine structure learning and reinforcement learning using the gating framework.
41. Jazbec S, Pantelis C, Robbins T, Weickert T, Weinberger DR, Goldberg TE: **Intra-dimensional/extra-dimensional set-shifting performance in schizophrenia: impact of distractors.** *Schizophr Res* 2007, **89(1-3)**:339-349.
42. Ceaser AE, Goldberg TE, Egan MF, McMahon RP, Weinberger DR, Gold JM: **Set-shifting ability and schizophrenia: a marker of clinical illness or an intermediate phenotype?** *Biol Psychiatry* 2008, **64(9)**:782-788.
43. Leeson VC, Robbins TW, Matheson E, Hutton SB, Ron MA, Barnes TRE, Joyce EM: **Discrimination learning, reversal, and set-shifting in first-episode schizophrenia: stability over six years and specific associations with medication type and disorganization syndrome.** *Biol Psychiatry* 2009, **66(6)**:586-593.
44. Todd MT, Niv Y, Cohen JD: **Learning to use working memory in partially observable environments through dopaminergic reinforcement.** *Neural Information Processing Systems*. Citeseer; 2009.
45. Peshkin L, Meuleau N, Kaelbling LP: **Learning policies with external memory.** In *Proceedings of the Sixteenth International Conference (ICML'99); Bled, Slovenia, June 27-30, 1999, Morgan Kaufmann Pub*: 1999:307.
46. Dayan P: **Bilinearity, rules, and prefrontal cortex.** *Front Comput Neurosci* 2007:1.
47. Bellman R: *Dynamic Programming*. Princeton University Press; 1957.
48. Braun DA, Mehring C, Wolpert DM: **Structure learning in action.** *Behavioral Brain Research* 2010, **206**:157-165.  
An excellent review of structure learning with an emphasis on sensorimotor control. The authors view structure learning as learning to learn — extracting a low-dimensional manifold on which the inputs to output mapping of many problems lie. This then accelerates learning of new tasks, as well as constrains exploration of solution strategies. This is somewhat different from our focus on learning of action structure, as we are interested in actions in the context of reinforcement learning.
49. Braun DA, Aertsen A, Wolpert DM, Mehring C: **Motor task variation induces structural learning.** *Curr Biol* 2009, **19(4)**:352-357.  
By exposing human subjects to a set of randomly varying motor tasks, the authors show that subjects are facilitated at learning novel tasks and exhibit decreased task-switching interference. This study provides evidence that humans employ structure learning to guide their generalization from experience, and that task variation is an important factor in structure learning.
50. Gershman SJ, Pesaran B, Daw ND: **Human reinforcement learning subdivides structured action spaces by learning effector-specific values.** *J Neurosci* 2009, **29(43)**:13524.
51. Palminteri S, Boraud T, Lafargue G, Dubois B, Pessiglione M: **Brain hemispheres selectively track the expected value of contralateral options.** *J Neurosci* 2009, **29(43)**:13465.
52. David JC, MacKay: *Information Theory, Inference and Learning Algorithms*. Cambridge University Press; 2003.