Elsevier Editorial System(tm) for Current Opinion in Neurobiology
Manuscript Draft

Manuscript Number:

Title: Learning latent structure: Carving nature at its joints

Article Type: Cognitive neuroscience 20/2

Corresponding Author: Dr. Yael Niv, Ph.D

Corresponding Author's Institution: Princeton University

First Author: Samuel J Gershman

Order of Authors: Samuel J Gershman; Yael Niv, Ph.D

# Learning latent structure: Carving nature at its joints

Samuel J. Gershman

*Princeton Neuroscience Institute & Psychology Department, Princeton University*

Yael Niv

*Princeton Neuroscience Institute & Psychology Department, Princeton University*

**Abstract**

Reinforcement learning 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 reinforcement learning, with an emphasis on structure in perception and action.

*Keywords:* Reinforcement learning, Bayesian inference, conditioning

## 1. 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

*Email addresses:* sjgershm@princeton.edu (Samuel J. Gershman), yael@princeton.edu (Yael Niv)

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 to 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 reinforcement learning problems: perception and action.

## 2. A normative framework

The canonical description of a decision making problem in RL has four components: 1) a set of *states* of the environment, each comprised of different stimuli, 2) a set of *actions* that can be taken at these states, 3) a *transition function* denoting how actions cause the world to transition from one state to another, and 4) 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, extinction, etc.). 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 uses 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 reinforcement learning models as a form of policy search over not only an action space, but also a state space.

## 3. Structure in perception

In this section, we focus on two types of perceptual structure that can play a role in reinforcement learning. 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 conextual 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 reinforcement learning 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], 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 on information stored in its memory as well [37]. 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” [26, 38–40] 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 unpredictive of reward. This model is consistent with the finding that prefrontal dopamine dysfunction appears to be responsible for abberant attentional control in schizophrenia [38, 41–43].

Recently, Todd, Niv and Cohen [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, 46], but direct experimental evidence supporting its central claims remains scant.

## 4. 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, 49]. 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, Pesaran and Daw [50] designed a reinforcement learning 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 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.

## 5. Conclusions

Although computational models of reinforcement learning 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.

### Box 1: Latent structures as a state space for reinforcement learning

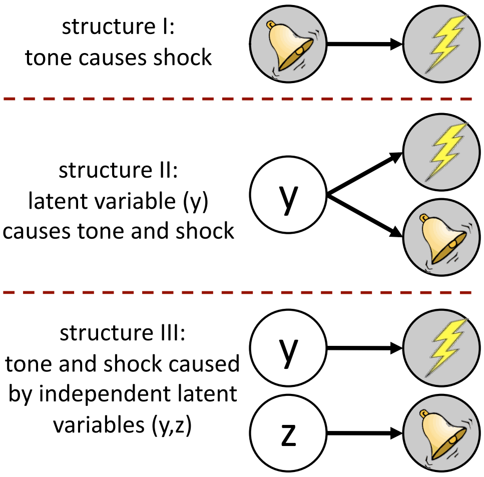

Figure 1: The graphical models of three possible causal relationships between variables in a classical conditioning experiment. By convention, observed variables are represented by shaded nodes and unshaded nodes represent unobserved (latent) variables. Arrows represent probabilistic dependencies. The 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).

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-occurence 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.

## References

[1] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[2] W. Schultz, P. Dayan, and P.R. Montague. A neural substrate of prediction and reward. *Science*, 275(5306):1593, 1997.

[3] J.C. Houk, J.L. Adams, and A.G. Barto. A model of how the basal ganglia generate and use neural signals that predict reinforcement. *Models of information processing in the basal ganglia*, pages 249–270, 1995.

[4] C. Kemp and J.B. Tenenbaum. Structured statistical models of inductive reasoning. *Psychological review*, 116(1):20–58, 2009.

> An elegant demonstration of how structured knowledge influences human reasoning in a large variety of domains.

[5] E. Mach. *The analysis of sensations*. 1897.

[6] H. Von Helmholtz. *Handbuch der Physiologischen Optik*. Voss, 1867.

[7] D.C. Knill and W. Richards. *Perception as Bayesian inference*. Cambridge Univ Pr, 1996.

[8] K. Doya, S. Ishii, P. Alexandre, and RPN Rao. *Bayesian brain*. MIT Press, 2007.

[9] N. Chater and M. Oaksford. *The probabilistic mind: Prospects for Bayesian cognitive science*. Oxford University Press, 2008.

[10] N. D. Daw, Y. Niv, and P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711, 2005.

[11] R. A. Rescorla and A. R. Wagner. A theory of of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black and W.F. Prokasy, editors, *Classical Conditioning II: Current Research and theory*, pages 64–99. Appleton-Century-Crofts, New York, NY, 1972.

[12] S.J. Gershman, D.M. Blei, and Y. Niv. Context, learning, and extinction. *Psychological Review*, 117(1):197–209, 2010.

[13] Aaron C. Courville, Nathaniel D. Daw, and David S. Touretzky. Similarity and discrimination in classical conditioning: A latent variable account. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 313–320, Cambridge, MA, 2002. MIT Press.

[14] Aaron C. Courville, Nathaniel Daw, Geoffrey J. Gordon, and David S. Touretzky. Model uncertainty in classical conditioning. In Sebastian Thrun, Lawrence Saul, and Bernhard Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[15] N.D. Daw, A.C. Courville, and D.S. Touretzky. Representation and timing in theories of the dopamine system. *Neural computation*, 18(7):1637–1677, 2006.

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

> 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] M.S.K. Yi, M. Steyvers, and M. Lee. Modeling Human Performance in Restless Bandits with Particle Filters. *The Journal of Problem Solving*, 2(2):5, 2009.

[18] Ed Vul, Michael Frank, George Alvarez, and Joshua Tenenbaum. Explaining human multiple object tracking as resource-constrained approximate inference in a dynamic probabilistic model. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1955–1963. 2009.

[19] Samuel Gershman, Ed Vul, and Joshua Tenenbaum. Perceptual multistability as Markov Chain Monte Carlo inference. In Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 611–619. 2009.

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

[21] Daniel Acuña and Paul R. Schrater. Structure learning in human sequential decision-making. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1–8. 2009.

> 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 expeirmenter. For instance, matching behavior can be understood given the assumption of a changing environment. Acuña 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 E. J. Behrens, Mark W. Woolrich, Mark E. Walton, and Matthew F. S. Rushworth. Learning the value of information in an uncertain world. *Nat Neurosci*, 10(9):1214–1221, 2007.

[23] A.M. Owen, A.C. Roberts, J.R. Hodges, and T.W. Robbins. Contrasting mechanisms of impaired attentional set-shifting in patients with frontal lobe damage or Parkinson's disease. *Brain*, 116(5):1159, 1993.

[24] B. Milner. Effects of different brain lesions on card sorting: The role of the frontal lobes. *Archives of Neurology*, 9(1):90, 1963.

[25] D.M. Barch, T.S. Braver, L.E. Nystrom, S.D. Forman, D.C. Noll, and J.D. Cohen. Dissociating working memory from task difficulty in human prefrontal cortex. *Neuropsychologia*, 35(10):1373–1380, 1997.

[26] R.C. O'Reilly and M.J. Frank. Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation*, 18(2):283–328, 2006.

[27] M.E. Bouton and R.C. Bolles. Contextual control of the extinction of conditioned fear. *Learning and Motivation*, 10(4):445–466, 1979.

[28] S. Nakajima, S. Tanaka, K. Urushihara, and H. Imada. Renewal of extinguished lever-press responses upon return to the training context. *Learning and Motivation*, 31(4):416–431, 2000.

[29] R.A. Rescorla and C.D. Heth. Reinstatement of fear to an extinguished conditioned stimulus. *J Exp Psychol*, 1:88–96, 1975.

[30] M.E. Bouton and R.C. Bolles. Role of conditioned contextual stimuli in reinstatement of extinguished fear. *J Exp Psychol*, 5:368–378, 1979.

[31] S.J. Robbins. Mechanisms underlying spontaneous recovery in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes*, 16(3):235–249, 1990.

[32] R.A. Rescorla. Spontaneous recovery. *Learning & Memory*, 11(5):501, 2004.

[33] M.E. Bouton. Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin*, 114:80–80, 1993.

[34] A.D. Redish, S. Jensen, A. Johnson, and Z. Kurth-Nelson. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. *Psychological review*, 114(3):784–805, 2007.

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] J. Ji and S. Maren. Electrolytic lesions of the dorsal hippocampus disrupt renewal of conditional fear after extinction. *Learning & Memory*, 12(3):270, 2005.

[36] C.S.L. Yap and R. Richardson. Extinction in the developing rat: An examination of renewal effects. *Developmental Psychobiology*, 49(6):565–575, 2007.

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] E.A. Zilli and M.E. Hasselmo. Modeling the role of working memory and episodic memory in behavioral tasks. *Hippocampus*, 18(2):193, 2008.

A computational exploration of how the brain can make use of memory mechanisms to solve reinforcement learning problems.

[38] T.S. Braver, D.M. Barch, and J.D. Cohen. Cognition and control in schizophrenia: A computational model of dopamine and prefrontal function. *Biological Psychiatry*, 46(3):312–328, 1999.

[39] T.S. Braver and J.D. Cohen. On the control of control: The role of dopamine in regulating prefrontal function and working memory. *Control of cognitive processes: Attention and performance XVIII*, pages 713–737, 2000.

[40] N.P. Rougier, D.C. Noelle, T.S. Braver, J.D. Cohen, and R.C. O'Reilly. Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences of the United States of America*, 102(20):7338, 2005.

One of the first attempts to combine structure learning and reinforcement learning using the gating framework.

[41] S. Jazbec, C. Pantelis, T. Robbins, T. Weickert, D.R. Weinberger, and T.E. Goldberg. Intra-dimensional/extra-dimensional set-shifting performance in schizophrenia: impact of distractors. *Schizophrenia research*, 89(1-3):339–349, 2007.

[42] A.E. Ceaser, T.E. Goldberg, M.F. Egan, R.P. McMahon, D.R. Weinberger, and J.M. Gold. Set-Shifting Ability and Schizophrenia: A Marker of Clinical Illness or an Intermediate Phenotype? *Biological Psychiatry*, 64(9):782–788, 2008.

[43] V.C. Leeson, T.W. Robbins, E. Matheson, S.B. Hutton, M.A. Ron, T.R.E. Barnes, and E.M. Joyce. Discrimination Learning, Reversal, and Set-Shifting in First-Episode Schizophrenia: Stability Over Six Years and Specific Associations with Medication Type and Disorganization Syndrome. *Biological Psychiatry*, 66(6):586–593, 2009.

[44] M.T. Todd, Y. Niv, and J.D. Cohen. Learning to use Working Memory in Partially Observable Environments through Dopaminergic Reinforcement. In *Neural information processing systems*, pages 1689–1696. Citeseer, 2009.

[45] L. Peshkin, N. Meuleau, and L.P. Kaelbling. Learning Policies with External Memory. In *Proceedings of the sixteenth international conference (ICML'99), Bled, Slovenia, June 27-30, 1999*, page 307. Morgan Kaufmann Pub, 1999.

[46] P. Dayan. Bilinearity, rules, and prefrontal cortex. *Frontiers in Computational Neuroscience*, 1, 2007.

[47] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[48] D.A. Braun, C. Mehring, and D.M. Wolpert. Structure learning in action. *Behavioural Brain Research*, 2010.

> 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] D.A. Braun, A. Aertsen, D.M. Wolpert, and C. Mehring. Motor task variation induces structural learning. *Current Biology*, 19(4):352–357, 2009.

> 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] S.J. Gershman, B. Pesaran, and N.D. Daw. Human Reinforcement Learning Subdivides Structured Action Spaces by Learning Effector-Specific Values. *Journal of Neuroscience*, 29(43):13524, 2009.

[51] S. Palminteri, T. Boraud, G. Lafargue, B. Dubois, and M. Pessiglione. Brain Hemispheres Selectively Track the Expected Value of Contralateral Options. *Journal of Neuroscience*, 29(43):13465, 2009.

[52] David J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.