Memory Modification in the Brain: Computational and Experimental Investigations

Samuel J. Gershman

Any aspect of cognition that is concerned with learning faces the same basic question: *when are old memories modified, and when are new memories formed?* For example, since Pavlov's first studies of classical conditioning, it has been believed that extinction of an associative memory entails new learning rather than modification of an old memory (Pavlov, 1927; Bouton, 2004). The same question arises in the domain of episodic memory, where the notion of an 'episode' hinges crucially on the ability to parse the stream of experience into distinct chunks. The question attains vivid clinical significance in the psychopathology of fear and addiction, where patients struggle to modify or erase a persistent, maladaptive memory.

My thesis explores how and when memory traces are modified by new experience. The centerpiece of the thesis is a theoretical framework for understanding memory modification, erasure, and recovery. According to Marr's taxonomy (Marr, 1982), the framework is situated primarily at the "computational level of analysis"—it formalizes the information processing task faced by the memory system, and derives a rational solution to this task (Anderson, 1990). The information processing task is posed as inductive reasoning in a probabilistic generative model of the environment, for which the rational solution is Bayesian inference. I present the results of behavioral and brain imaging experiments to support various aspects of this theory. To demonstrate the generality of the theory, these experiments cut across different species, tasks and stimuli. I emphasize that this theory is truly a *framework*—its details vary to accommodate the variety of domains to which it is applied, but at its core is a set of computational ideas that are postulated to hold across domains. These computational ideas are applied to animal learning, visual perception, and human memory; the thesis also devotes considerable attention to how these ideas might be implemented in neural circuitry.

The central argument of the thesis is that memories reflect inferences about the structure of the world. In particular, memories reflect the assignment of events to latent (hidden) causes. A new event modifies an existing memory trace if it is probable that the event was caused by the same latent cause as that represented by the old trace; otherwise, a new memory trace is formed. I show that probabilistic inference over latent causes, or *structure learning*, provides a parsimonious explanation of many phenomena in human and animal learning, and may guide us towards developing new treatments for pathological memories like trauma and addiction.

Chapter 1: background

Chapter 1 summarizes previous theoretical and experimental work on memory modification, arguing that earlier theories are limited in fundamental ways. In particular, psychological models that assume separate storage of traces (i.e., photographic snapshots) are biologically implausible, whereas neural network models are more biologically plausible but suffer from the assumption that memories tend to overwrite each other irreparably. I introduce a new framework for thinking about this problem, developed more formally in Chapter 3, that occupies an intermediate positions between these extremes by using probabilistic reasoning to determine when new memories should be create or old ones modified.

Chapters 2 and 3: the latent cause framework

The material in these chapters was published as Gershman et al. (2010), Gershman and Blei (2012) and Gershman and Niv (2012).

Chapter 2 provides mathematical background for the theory developed in later chapters, summarizing basic technical concepts from probability theory and Bayesian nonparametrics. The latter is an area of research that has received considerable recent attention in cognitive science, since it has supplied theoretical machinery for modeling structure learning. Bayesian nonparametric models, in particular those based on the Dirichlet process, allow the representational complexity of a model to grow with more data; this idea has been invoked to explain a variety of behavioral phenomena in categorization (Anderson, 1991; Sanborn et al., 2010), language learning (Goldwater et al., 2009), and cognitive control (Collins and Frank, 2013).

Chapter 3 uses the technical concepts introduced in Chapter 2 to formulate a latent cause framework, instantiated as a model of classical conditioning. The key idea of this framework is that learning in classical conditioning consists of a form of adaptive clustering: an animal receives patterns of stimulus configurations, and tries to group these together by assigning similar configurations to a common latent cause. Consider an animal in a fear conditioning experiment. During the training phase, the animal observes a series of tone-shock pairs; during the extinction phase, the animal observes the tone by itself. It has traditionally been thought that the animal learns an association between tone and shock over the course of training (Pearce and Bouton, 2001), which leads to the erroneous prediction that extinction results in the unlearning of the association. The latent cause framework provides a very different interpretation: observational data (tone-shock or tone-alone trials) are generated by latent causes, drawn from a distribution $P(\text{data}|\text{cause})$. The latent causes needn't have a direct physical meaning; it is better to think of them as hypothetical entities posited by the animal as a means of organizing its observational data.

Given some observational data, the animal computes the conditional distribution over possible causes given the observation—commonly known as the *posterior* distribution (shown schematically in Figure 1). This distribution may include previously inferred causes, as well as the hypothesis that a completely new cause generated the data. Mathematically, the posterior is given by the

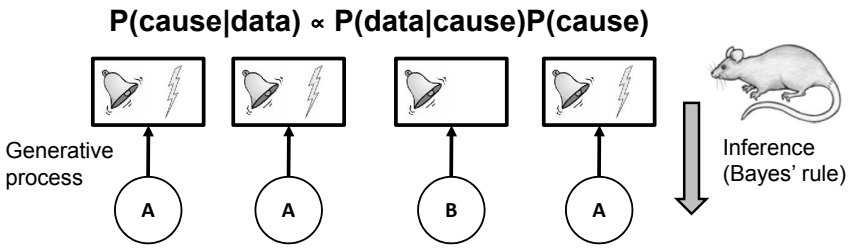

Figure 1: **Schematic of the latent cause theory**. Each box represents the animal's observations on a single trial. The circles represent latent causes, labeled according to their identity. The upward arrows denote probabilistic dependencies: observations are *generated* by latent causes. The animal does not get to observe the latent causes; it must infer these by inverting the generative model using Bayes' rule, as indicated by the downward arrow. As shown at the top of the schematic, Bayes' rule defines the probability of latent causes conditional on observations, which is obtained (up to a normalization constant) by multiplying the probability of observations given hypothetical causes (the likelihood) and the probability of the hypothetical latent causes (the prior).

axiom of probability theory known as *Bayes' rule*:

$$P(\text{cause}|\text{data}) \propto P(\text{data}|\text{cause})P(\text{cause}). \tag{1}$$

The second term in Eq. 1 is known as the *prior*—it encodes the animal's "inductive bias" (Griffiths et al., 2010) about which latent causes are likely *a priori*. In later chapters, I go into much greater detail about what kinds of inductive biases the brains of humans and animals might be using.

When several observations are assigned to the same latent cause, the summary statistics of these observations become associated with that cause. For example, when all the training trials are assigned to a single latent cause, that latent cause's distribution over observations becomes concentrated on tone-shock pairs. During extinction, this distribution is a poor predictor of tone-alone trials; because the posterior distribution must sum to 1, reducing the probability of assigning these trials to the training cause results in a corresponding increase in the probability of assigning them to a new, "extinction" cause. This is a precise formalization of the frequently proposed idea that extinction involves new learning (e.g., Bouton, 1993; Delamater, 2004). Thus, we can think of each latent cause as encoding a trace of a set of observations, and new causes are inferred when none of the previous traces are good predictors of incoming observations. The Bayesian framework provides a rational answer to the question of when a new memory should be formed. The rest of the thesis is devoted to a wide-ranging exploration of this basic idea.

Chapter 4: understanding memory reconsolidation

Chapter 4 develops a variant of the latent cause framework designed to explain the phenomenon of memory reconsolidation (Spear, 1973): retrieving a memory appears to render it temporarily labile. This finding has revolutionized our understanding of memory by showing that memories, once consolidated, are not calcified forever into a fixed form; rather, memories can be returned to a

malleable form by retrieval (e.g., the memory can be impaired by the injection of protein synthesis inhibitors following retrieval).

In order to understand reconsolidation, it is necessary to first review the concept of *consolidation*, the apparent time-dependent stabilization of memory traces (Muller and Pilzecker, 1900). The key finding motivating the concept of consolidation is the temporal gradient of retrograde amnesia (RA): new memories tend to be more susceptible to disruption than old memories (see Wixted, 2004, for a review). This gradient is seen both in experimental amnesia (e.g., induced by electroconvulsive shock; Quartermain et al., 1965; Kopp et al., 1966) and amnesia resulting from insult to the medial temporal lobes (Brown, 2002), although this assertion has not gone undisputed (Nadel et al., 2007). There is also evidence for a temporal gradient of RA for emotional memories in the amygdala (Schafe and LeDoux, 2000). The "standard model of consolidation" explains these findings by postulating that new memories exist in a temporarily labile state in the hippocampus until they are gradually transferred into a stable neocortical representation (Squire and Alvarez, 1995).

The standard model of consolidation was challenged by findings that ostensibly stable memories could be rendered labile by an appropriately timed "reminder" treatment (Lewis et al., 1968; Misanin et al., 1968; Mactutus et al., 1979). For example, administering electroconvulsive shock within a short time window after a single unreinforced conditioned stimulus (CS) presentation resulted in RA for the (putatively consolidated) CS-US association. Such reminder treatments not only made memories susceptible to interference by amnestic agents, but also allowed memories to be enhanced, for example by stimulation of the reticular formation (Devietti et al., 1977). These findings indicated that the temporal gradient of RA is at least partially determined by the activation state of a memory. When reactivated by a reminder treatment, memories must undergo a phase of reconsolidation to achieve stability.

After a flurry of experimental activity in the 1970s, this idea smoldered for several decades until Nader et al. (2000) showed, using a fear conditioning paradigm, that injection of the protein synthesis inhibitor (PSI) anisomycin into the basolateral amygdala following re-exposure to the training cues caused RA for the earlier fear memory. Thus, reactivated memories require new protein synthesis to reconsolidate into a stable state. This finding ushered in a new era of reconsolidation studies using pharmacological treatments (see Nader and Hardt, 2009, for a recent review). Some of the most compelling evidence for the proposition that reconsolidation induces memory modification (rather than memory formation) comes from subsequent work by Duvarci and Nader (2004) showing that several signatures of new learning during extinction reviewed above (spontaneous recovery, reinstatement, renewal) are absent following post-reactivation protein synthesis inhibition.

I show that the major phenomena of reconsolidation can be explained in terms of the latent cause framework framework, including variables such as memory strength, age, and various timing parameters of the conditioning protocols. I also present new experimental data testing some of the theory's predictions. In particular, I show that partial extinction prior to a reminder fails to induce memory malleability, consistent with the idea that partial extinction induces the inference of a new "extinction" latent cause, which is subsequently retrieved by the reminder (rather than modification of the original fear memory).

Chapters 5-8: testing the theory

The material in these chapters was published as Sederberg et al. (2011), Gershman et al. (2013a), Gershman et al. (2013b), and Gershman and Niv (2013).

Chapter 5 presents experimental tests of the latent cause framework—in particular, the idea that the latent causes inferred by rats depend on the reinforcement schedule during extinction. I show that extinction resembles unlearning when the extinction trials are scheduled in such a way that rats will assign extinction trials to the same latent cause as the acquisition trials.

To understand why this works, consider what happens computationally during extinction. The onset of extinction training produces a large prediction error—a discrepancy between the predicted outcome (e.g., shock) and the experienced outcome (no shock). Traditional models of associative learning propose that such prediction errors serve as a learning signal, driving the modification of predictions (e.g., Rescorla and Wagner, 1972). According to these accounts, the absence of shocks during the extinction procedure should reduce the strength of the original fear memory. However, recent models (such as the ones described in Chapters 3 and 4) propose that persistently large prediction errors might also serve as a segmentation signal, indicating to the animal a novel situation that demands new associations (Redish et al., 2007; Gershman et al., 2010). This can explain why the traditional extinction procedure leads to formation of a new, competing, "no-fear" memory, all the while allowing the original fear memory to persist unmodified.

The idea that large prediction errors are signals for segmentation suggests that one could modify the original fear memory if prediction errors were small or infrequent enough to not induce formation of a new memory, but large enough to drive some learning. To test this prediction, I designed a "gradual extinction" paradigm in which the aversive event (a foot shock) was gradually and progressively eliminated. The idea was to change the association of the cue from a shock to no shock gradually enough so as to avoid persistent, large prediction errors. If we could prevent the creation of a new memory trace, all learning would affect the old fear memory, which would gradually be weakened and erased. Two fear conditioning experiments with rats support this prediction: we found no evidence for fear recovery using a gradual extinction protocol, despite the fact that the rats were actually getting shocked *more* during extinction than during conditioning.

Similar results were found using an analogous reconstructive memory task in humans (Chapter 6). Thus, the same principles of gradual memory modification apply across a range of tasks and species.

Using a simple perceptual estimation task, I further explore the inductive biases underlying latent causal inference (Chapter 7). I show that the way in which people assign visual stimuli to latent causes influences their perceptual estimates. New causes are posited only when the properties of different stimuli are statistically distinguishable (a probabilistic manifestation of Occam's razor).

Finally, I present neuroimaging evidence that the reinstatement of "mental context" predicts memory misattributions (Chapter 8). Using a list-learning paradigm, I show that I can measure neurally the reinstatement of mental context from the study phase of one list during the study phase of another list. The amount of reinstatement is predictive, on an item-by-item basis, of the probability that a list 2 item will subsequently be misattributed to list 1. Interpreted within the latent cause

framework, this suggests that misattributions are governed by the degree to which list 2 items are believed to have been generated by the same latent cause as list 1 items.

Conclusion

Taken together, these experimental and theoretical results support the idea that memory modification can be understood as a process of structure learning. My thesis demonstrates the ability of a single theoretical framework to capture a wide range of empirical phenomena concerning the nature of memory modification. The framework motivated a series of experiments, using a variety of methods (classical conditioning, reconstructive memory, list learning, neuroimaging) that provide direct evidence for the theory. These results provide new computational answers to old questions about how memories are modified and created.

References

Anderson, J. (1990). *The adaptive character of thought*. Lawrence Erlbaum.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin*, 114:80–99.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning and Memory*, 11:485–494.

Brown, A. (2002). Consolidation theory and retrograde amnesia in humans. *Psychonomic Bulletin & Review*, 9(3):403.

Collins, A. G. and Frank, M. J. (2013). Cognitive control over learning: Creating, clustering, and generalizing task-set structure. *Psychological Review*, 120(1):190–229.

Delamater, A. (2004). Experimental extinction in pavlovian conditioning: Behavioural and neuroscience perspectives. *Quarterly Journal of Experimental Psychology Section B*, 57(2):97–132.

Devietti, T., Conger, G., and Kirkpatrick, B. (1977). Comparison of the enhancement gradients of retention obtained with stimulation of the mesencephalic reticular formation after training or memory reactivation. *Physiology & Behavior*, 19(4):549–554.

Duvarci, S. and Nader, K. (2004). Characterization of fear memory reconsolidation. *Journal of Neuroscience*, 24(42):9269.

Gershman, S. and Blei, D. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56:1–12.

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

Gershman, S. J., Jones, C. E., Norman, K. A., Monfils, M.-H., and Niv, Y. (2013a). Gradual extinction prevents the return of fear: implications for the discovery of state. *Frontiers in Behavioral Neuroscience*, 7.

Gershman, S. J. and Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & behavior*, 40:255–268.

Gershman, S. J. and Niv, Y. (2013). Perceptual estimation obeys occam's razor. *Frontiers in psychology*, 4.

Gershman, S. J., Schapiro, A. C., Hupbach, A., and Norman, K. A. (2013b). Neural context reinstatement predicts memory misattribution. *The Journal of Neuroscience*, 33:8590–8595.

Goldwater, S., Griffiths, T., and Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Griffiths, T., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364.

Kopp, R., Bohdanecky, Z., and Jarvik, M. (1966). Long temporal gradient of retrograde amnesia for a well-discriminated stimulus. *Science*, 153(3743):1547.

Lewis, D., Misanin, J., and Miller, R. (1968). Recovery of memory following amnesia. *Nature*, 220:704–705.

Mactutus, C., Riccio, D., and Ferek, J. (1979). Retrograde amnesia for old (reactivated) memory: some anomalous characteristics. *Science*, 204(4399):1319.

Marr, D. (1982). *Vision*. "W. H. Freeman and Company, San Francisco, CA.

Misanin, J., Miller, R., and Lewis, D. (1968). Retrograde amnesia produced by electroconvulsive shock after reactivation of a consolidated memory trace. *Science*, 160(3827):554.

Muller, G. and Pilzecker, A. (1900). Experimentelle bietrage zur lehre bom gedachtnisses. *Zeits. Fr Psych.*, 1:1–288.

Nadel, L., Winocur, G., Ryan, L., and Moscovitch, M. (2007). Systems consolidation and hippocampus: two views. *Debates in Neuroscience*, 1(2):55–66.

Nader, K. and Hardt, O. (2009). A single standard for memory: the case for reconsolidation. *Nature Reviews Neuroscience*, 10(3):224–234.

Nader, K., Schafe, G., and Le Doux, J. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406(6797):722–726.

Pavlov, I. (1927). *Conditioned Reflexes*. Oxford University Press.

Pearce, J. M. and Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology*, 52:111–139.

Quartermain, D., Paolino, R., and Miller, N. (1965). A brief temporal gradient of retrograde amnesia independent of situational change. *Science*, 149(3688):1116.

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

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

Sanborn, A., Griffiths, T., and Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4):1144–1167.

Schafe, G. and LeDoux, J. (2000). Memory consolidation of auditory pavlovian fear conditioning requires protein synthesis and protein kinase A in the amygdala. *Journal of Neuroscience*, 20(18):96.

Sederberg, P., Gershman, S., Polyn, S., and Norman, K. (2011). Human memory reconsolidation can be explained using the temporal context model. *Psychonomic Bulletin & Review*, 18:455–468.

Spear, N. (1973). Retrieval of memory in animals. *Psychological Review*, 80(3):163–194.

Squire, L. and Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology*, 5(2):169–177.

Wixted, J. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, 55(1):235.