Learn Behav (2012) 40:255–268
DOI 10.3758/s13420-012-0080-8

Exploring a latent cause theory of classical conditioning

Samuel J. Gershman · Yael Niv

Abstract We frame behavior in classical conditioning experiments as the product of normative statistical inference. According to this theory, animals learn an internal model of their environment from experience. The basic building blocks of this internal model are latent causes—explanatory constructs inferred by the animal that partition observations into coherent clusters. Generalization of conditioned responding from one cue to another arises from the animal's inference that the cues were generated by the same latent cause. Through a wide range of simulations, we demonstrate where the theory succeeds and where it fails as a general account of classical conditioning.

Keywords Classical conditioning · Bayes' rule · Particle filtering

A normative theory of behavior is one that takes as its starting point the question "What is the ecological problem facing the animal?" and, once this question has been suitably formalized, characterizes its optimal (or "rational") solution. The value of such a rational analysis (cf. Anderson, 1990) lies not in its prescription for behavior, but rather in the *functional organization* of behavior that it makes explicit. Marr (1982) expressed this point eloquently in the context of visual perception: "[T]rying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense" (p. 27). While it is natural to think of wings in terms of flight, the logic of behavior in classical (Pavlovian) conditioning experiments is less obvious. Only a relatively small number of theories have attempted to answer this question directly (e.g., Courville et al., 2006; Dayan, Kakade, & Montague, 2000; Dayan & Long, 1998; Gallistel & Gibbon, 2000; Sutton & Barto, 1990).

This article explores a simple normative theory of classical conditioning, first presented in Gershman, Blei, and Niv (2010). Unlike most associative-learning theories, which describe how animals learn to associate observed events to each other, our theory postulates that animals attempt to learn the *hidden (latent) structure* of the environment from their experience and that they employ this internal model of the environment to make predictions about unobserved or future variables (see also Schmajuk, Lam, & Gray, 1996; Sokolov, 1960). Following the seminal work of Courville, Daw, Gordon, and Touretzky (2003) and Courville, Daw, and Touretzky (2004, 2006), we assume that the basic building blocks of the animal's internal model are *latent causes*—variables inferred by the animal that partition trials into different clusters. As applied to classical conditioning experiments, we frame the conditioned response (CR) to a conditioned stimulus (CS) as resulting from a prediction about the unconditioned stimulus (US), given the other stimuli present. Generalization of conditioned responding from one cue to another arises from the animal's inference that the cues were generated by the same latent cause.

In the next section, we first describe the essence of the theory, and then present a formal mathematical description.[1] Following that, we report a wide range of simulations that are intended to illustrate the theory's strengths, as well as its

Electronic supplementary material The online version of this article (doi:10.3758/s13420-012-0080-8) contains supplementary material, which is available to authorized users.

S. J. Gershman (✉) · Y. Niv
Department of Psychology and Princeton Neuroscience Institute, Princeton University,
Princeton, NJ 08540, USA
e-mail: sjgershm@princeton.edu

[1] MATLAB code implementing the model is available at the first author's webpage: www.princeton.edu/~sjgershm.

weaknesses. Finally, in the Discussion, we suggest some promising routes toward improving the theory's explanatory reach as well as its connections to other normative accounts.

The latent cause theory

In this section, we review the latent cause theory of classical conditioning introduced by Gershman, Blei, and Niv (2010). This theory owes a major intellectual debt to the work of Courville and his colleagues (Courville et al., 2003; Courville et al., 2004, 2006), who were the first to demonstrate the usefulness of thinking about classical conditioning in terms of latent causes. We also drew inspiration from the reinforcement learning model of Redish, Jensen, Johnson, and Kurth-Nelson (2007); see Gershman et al. (2010) for a detailed comparison of these models. At a mathematical level, our theory is directly descended from work on human categorization, in particular from Anderson's (1991) rational model of categorization and its descendants (Sanborn, Griffiths, & Navarro, 2010).

In our theory, the animal combines its a priori beliefs about how the world is structured together with its current observations to make inferences about how CSs and USs are linked, and to make predictions about the possible future occurrence of a US. We use the term *observation* to refer to the set of features (CS, US, context, etc.) presented to the animal on a particular trial. We further assume that the animal combines beliefs and observations statistically correctly—that is, by using Bayes' rule to combine prior beliefs and the likelihood of an observation to form a posterior belief.

In particular, we assume that the animal divides its observations into groups or clusters, according to the hypothesized latent cause of each trial. To the extent that different trials are thought to result from the same (hidden) cause, they are combined to form an expectation for the probability of a US (and of other cues) given that cause. To determine the strength of its CR on the current trial, the animal first determines what cause is likely to be active, given the current observed cues, and then makes a prediction about the occurrence of the US according to the statistics previously observed for this latent cause. Thus, in essence, the animal is not learning to associate CSs with USs, but rather to associate latent causes with both CSs and USs. Importantly, as the latent causes are not observed, the animal must rely on its subjective inference about which causes were responsible for which trials.

To structure the animal's prior beliefs about latent causes, our theory imputes to the animal a set of probabilistic assumptions about the environment that collectively constitute a *generative process*—a stochastic "recipe" that the animal assumes has generated its observed data. Intuitively, the generative model amounts to a set of four assumptions:

1. Each trial is caused by one latent cause.
2. Each latent cause has some characteristic probability of emitting observed features (CS, US, etc.).
3. All else being equal, a prolific latent cause (i.e., one that has caused many trials) is more likely to cause another trial.
4. There is some small probability that the current trial results from a completely new latent cause (i.e., one that has not yet generated any observations).

Given such a generative model, the animal can reason backward from observations to latent causes using Bayes' rule (shown schematically in Fig. 1). Once it has inferred what latent cause is active in the current trial, the animal can expect the US insofar as this latent cause has previously emitted the US—that is, if the (inferred) characteristic emission probabilities of the latent cause suggest that a US is likely to appear—and can generate the CR appropriately. We now turn to a more formal description of the theory.

The internal model imputed to the animal

We assume that the animal's observation on trial t takes the form of a discrete-valued[2] multidimensional vector $\mathbf{f}_t = \{f_{t,1}, \ldots, f_{t,D}\}$. Each feature corresponds to the presence or absence of a particular stimulus (e.g., CS, US, context), with the first feature, in particular, corresponding to the binary occurrence or absence of a US—$f_{t,1} \in \{\text{US}, \neg\text{US}\}$—and other features varying depending on the particular experiment. Commonly, there is a cue feature representing a typical Pavlovian CS (or its absence): $f_{t,d} \in \{\text{CS}, \neg\text{CS}\}$. Some experiments[3] include a context feature, an abstraction of typical context manipulations (e.g., box color, shape, or odor), which we simplify into discrete values: $f_{t,d} \in \{\text{contextA}, \text{contextB}, \text{contextC}, \text{etc.}\}$.

As mentioned above, the generative model that we impute to the animal is one in which, on each trial, a single latent cause is responsible for generating the observations (see Courville et al., 2003, for an example of a latent cause model in which multiple latent causes can be active on a single trial). In such a *mixture model*, each trial is assumed to be generated stochastically by first sampling a cause c_t according to a mixing distribution $P(c)$ and then sampling an observation vector conditioned on the cause from an observation distribution $P(\mathbf{f} \mid c_t)$. This type of generative model is a reasonable prior belief for many environments,

[2] The choice of discrete-valued observations is not crucial to our formalism; we have used real-valued features and obtained similar results.

[3] We chose not to include this context feature in all simulations, in the interest of simplicity; this affected the quantitative, but not the qualitative, pattern of results.

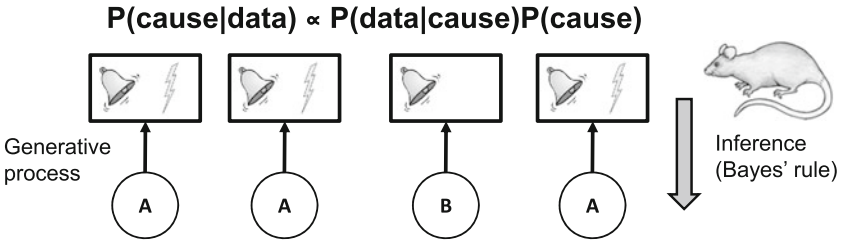

Fig. 1 Schematic of the latent cause theory. Each box represents the animal's observations on a single trial. The circles represent latent causes, labeled to distinguish different causes. The upward arrows denote probabilistic dependencies: Observations are assumed to be *generated* by latent causes. 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)

since it expresses, to a first approximation, the process by which many conditioning procedures are generated: first a phase (e.g., conditioning, extinction, or test) is selected, and then a set of stimuli are selected according to the phase.

If the animal assumes that each observation is generated by a single latent cause, then "clustering" is the process of recovering these causes on the basis of its observations. The clusters inferred by the animal may not be identical to the true causes of its observations; indeed, these are explanatory constructs that do not necessarily correspond to objects in the real world.

It seems reasonable to suppose that animals do not know (or decide) a priori how many latent causes will be involved in a certain situation. This means that, as they observe more data, animals must have the ability to flexibly expand their repertoire of latent causes. We can specify the mixture model described above such that the number of latent causes is unbounded—a so-called "infinite-capacity" mixture model.[4]

Formally, let us denote a partition of observations (trials) 1, . . . , t by the vector $\mathbf{c}_{1:t} = \{c_1, \ldots, c_t\}$. A partition specifies which observations were generated by which causes, such that $c_t = k$ indicates that the observation t was generated by cause k. In our model, before observing any data, the animal's prior belief over how likely different partitions will be is an infinite-capacity mixture model (see Gershman & Blei, 2012, for a tutorial introduction). This can be written as a sequential process that generates cause k on trial t with probability

$$P(c_t = k|\mathbf{c}_{1:t-1}) = \begin{cases} \frac{N_k}{t-1+a} \text{ if } k \text{ is an old cause} \\ \frac{a}{t-1+a} \text{ if } k \text{ is an new cause,} \end{cases} \quad (1)$$

where N_k is the number of observations already generated by cause k (by default, it is assumed that $c_1 = 1$). The number of causes generating observations 1, . . . , t is now a random variable and can be any number from 1 to t. Through determining the probability of assigning trial t to a new cause, the concentration parameter α specifies the animal's prior belief about the number of causes in the environment. When $\alpha = 0$, all observations are generated by a single cause; when α approaches ∞, each observation is generated by a unique cause. In general, for $\alpha < \infty$, the animal assumes that observations will be generated by a number of causes that is smaller than the number of observations, with fewer causes accounting for more data the lower α is. Additional information about this distribution can be found in the supplemental materials.

Once a cause has been sampled for a trial, an observation is sampled from an observation distribution conditional on the cause. In our model, each cause is linked to a multinomial distribution over features, parameterized by ϕ, where $\phi_{i,j,k}$ is the probability of observing value j (e.g., US) for feature i given latent cause k. A common assumption in mixture models (which we adopt here) is that, in the generative model, each feature is conditionally independent of all the other features, given the latent cause and its multinomial parameters. The conditional-independence assumption expresses the idea that, given the identity of the latent cause, CSs and US are generated separately, each according to its relevant probability $\phi_{i,j,k}$. In this sense, our model does not embody associations between CSs and USs, but rather between each of these and the latent causes.

At this juncture, it is worth noting several questionable assumptions of the proposed internal model of the environment. In this model, features are assumed to be conditionally independent given the latent cause. This assumption is particularly consequential with respect to the predictive relationship between cues and reward: In the model, two cues that appear simultaneously do not summate in the traditional sense of increasing the total predicted US, but rather increase the probability that a cause that tends to emit both cues is active (whether this cause is associated with one or more USs depends on the cause, not the cues). As we will discuss in the Simulations section, this prevents the model from capturing phenomena like overexpectation and

[4] For any given set of T observations, only a finite number of latent causes will actually be active (at most T).

superconditioning. Blocking is another phenomenon that is elegantly explained by different generative assumptions (e.g., the linear-Gaussian model proposed by Kakade & Dayan, 2002). Another questionable assumption is the *exchangeability* of the infinite-capacity mixture model distribution over latent causes: Changing the order of observations does not change the probability of the partition. This assumption prevents the model from learning about temporal structure in its observation sequence (Savastano & Miller, 1998).We emphasize that these assumptions are not intrinsic to our model; our goal is to develop a modeling *framework* in terms of latent causes, and to articulate one simple variant that can still capture a wide range of phenomena. In any case, in this article we explore the explanatory power of the above-specified model, replete with its specific assumptions.

Approximate inference

The inference problem facing the animal consists of two components: identifying the latent causes of its observations, and predicting the US given a partial observation (i.e., an observation consisting of cues such as CSs and context, but excluding the US). Because in our model prediction depends on inferences about latent causes, we address each of these components in turn.

Let $\mathbf{F}_{1:t} = \{\mathbf{f}_1, \ldots, \mathbf{f}_t\}$ denote the observations on trials $1, \ldots, t$. According to Bayesian inference (Gelman, Carlin, Stern, & Rubin, 2004), the animal's beliefs about the latent causes of the observations up to trial t are encoded by the posterior distribution over partitions, given the observations:

$$P(\mathbf{c}_{1:t}|\mathbf{F}_{1:t}) = \frac{P(\mathbf{F}_{1:t}|\mathbf{c}_{1:t})P(\mathbf{c}_{1:t})}{\sum_{\mathbf{c}_{1:t}} P(\mathbf{F}_{1:t}|\mathbf{c}_{1:t})P(\mathbf{c}_{1:t})} \tag{2}$$

where the posterior probability of each partition $\mathbf{c}_{1:t}$ is determined both by the prior probability of this partition, $P(\mathbf{c}_{1:t})$, and the likelihood of the observed features if this partition was the true assignment of trials to latent causes, $P(\mathbf{F}_{1:t} \mid \mathbf{c}_{1:t})$. This means that, although the generative process assumes that features are generated independently given a cause, in inference the probability of a partition also depends on multiplicative interactions between features: A partition is more likely to the extent that it involves consistent feature values in each cluster.

Unfortunately, this posterior probability is computationally intractable, since the denominator in Eq. 2 involves a summation over an exponentially large number of partitions. We must therefore consider approximate inference algorithms. One approximate inference algorithm that is suitable for implementation in the brain is the *particle filter* (Fearnhead, 2004). This algorithm approximates the posterior distribution over partitions using a set of samples (or particles), which it updates in an online fashion as new observations arrive. The particle filter algorithm has been used successfully to model a number of learning phenomena (Brown & Steyvers, 2009; Daw & Courville, 2008; Sanborn et al., 2010). The essential idea in particle filtering is to create a set of m hypothetical partitions of trials into causes such that each partition is represented in the set approximately in proportion to the partition's posterior probability according to Eq. 2. Specifically, the probability of sampling a partition depends on factors such as the number of latent causes in the partition and whether similar observations are clustered together. A detailed description of the particle filter algorithm can be found in the supplemental materials.

Finally, we assume that the animal's conditioned (Pavlovian) response is proportional to the predicted probability of a US given a "test" observation that lacks the first (US or ¬US) feature. This prediction (which we denote V) is based on the animal's posterior beliefs about the structure of the task and about the cause to which the current test trial is assigned:

$$\begin{aligned} V_t &= P\left(f_{t,1} = \mathrm{US}|\mathbf{f}_{t,2:D}, \mathbf{F}_{1:t-1}\right) \\ &= \sum_{\mathbf{c}_{1:t}} P\left(f_{t,1} = \mathrm{US}|c_t, \mathbf{c}_{1:t-1}, \mathbf{f}_{1:t-1,1}\right) P\left(c_t|\mathbf{f}_{t,2:D}, \mathbf{F}_{1:t-1,2:D}, \mathbf{c}_{1:t-1}\right) \\ &\quad \times P(\mathbf{c}_{1:t-1}|\mathbf{F}_{1:t-1}). \end{aligned} \tag{3}$$

For each possible partition of trials into latent causes $\mathbf{c}_{1:t}$, this equation calculates the probability of the US assuming that the current trial was caused by c_t, and given the occurrence of USs in previous trials that were assigned to this same latent cause. In our model, this amounts to simply counting, for all previous trials assigned to latent cause c_t, how many trials had a US and how many did not (hence, the dependency on $\mathbf{c}_{1:t-1}$ and $\mathbf{f}_{1:t-1,1}$). This is then weighted by the probability of assigning trials $1{:}t$ to causes $\mathbf{c}_{1:t}$: The last term is the (recursively calculated, from the previous trial) posterior probability of the partition up to trial $t-1$, and the middle term is the probability of assigning the current trial to c_t. Specifically, the middle term involves both the likelihood of the current trial being generated by c_t, as determined by similarity between the cues in this trial and cues in previous trials assigned to the same latent cause, and the prior probability of latent cause c_t given the previous latent causes:

$$\begin{aligned} &P\left(c_t = c|\mathbf{f}_{t,2:D}, \mathbf{F}_{1:t-1,2:D}, \mathbf{c}_{1:t-1}\right) \\ &\quad = \frac{P\left(\mathbf{f}_{t,2:D}|\mathbf{F}_{1:t-1,2:D}, c_t = c, \mathbf{c}_{1:t-1}\right) P(c_t = c|\mathbf{c}_{1:t-1})}{\sum_j P\left(\mathbf{f}_{t,2:D}|\mathbf{F}_{1:t-1,2:D}, c_t = j, \mathbf{c}_{1:t-1}\right) P(c_t = j|\mathbf{c}_{1:t-1})}. \end{aligned} \tag{4}$$

As in Eq. 2, this is the normalized product of the likelihood of this trial's features (excluding the US feature) assuming previous trials assigned to the same cause (again, calculated by counting how many trials shared these features) and the prior probability of assigning this trial to this

latent cause (Eq. 1). V_t is thus the probability of a US given the assignment of trials to causes, averaged over all the possible partitions weighted by their probabilities.[5]

Figure 2 illustrates the behavior of the model in 20 trials of simple conditioning. For this scenario, we show the behavior of a single latent cause. The left panel shows the conditional probability of the US (the leftmost term in Eq. 3). It is monotonically increasing as the confidence that the US will occur in the presence of the CS and the latent cause increases (see Eq. 4 in the supplementary materials). The right panel shows the probability of the latent cause (middle term in Eq. 3), which is close to 1 for all of these trials, since the statistical evidence for a second latent cause is weak.[6] Note that we do not show the rightmost term here because in the particle filter implementation, this term is implicitly represented by the set of latent causes (see the supplementary materials).

Simulations

In this section, we present simulations of the model's behavior in a wide range of experimental paradigms. Our goal is to illustrate both the breadth of the model's explanatory reach and its limitations. Of necessity, we concentrate on only a subset of the relevant literature. The simulations are broken down into the major research categories of classical conditioning. In each category, we begin by describing key phenomena that the model accounts for, and then discuss failures of the model.

In all of the simulations reported below, the value of α (the only free parameter of the model) was set to 1. As will become apparent, for some of our simulations, the quantitative size of the effect is small, casting doubt on the model's fidelity to the data. One approach would be to fit α to summary statistics of the data rather than to set it a priori. However, the quantitative predictions of our model are also dependent on somewhat arbitrary choices in the model specifications, such as binary features and Dirichlet-multinomial priors on the observation parameters (see the supplementary materials). Instead, our goal is to expose qualitative rather than quantitative patterns. To ensure that our small effects are indeed qualitatively robust, in each case we have run simulations, not reported here in the interest of brevity, with values of α that have ranged over an order of magnitude. In all cases, we were able to achieve a credibly large effect with values in this range.

[5] Details of how V_t is calculated recursively using the particle filter are contained in the supplemental materials.

[6] The slight decrease after the first trial is a consequence of the fact that all of the particles are initialized to the first latent cause on the first trial, but the particles can assign different trials to different latent causes after that.

Acquisition, extinction, and recovery

The latent cause theory accounts trivially for the basic acquisition of the Pavlovian response (Pavlov, 1927), as well as the lower asymptote of conditioned responding following partial reinforcement. In addition, it predicts the loss of responding during an extinction treatment (see Gershman et al., 2010). An important constraint on theories of learning is the fact that extinguished or attenuated responding can be recovered following posttraining manipulations or changes in testing conditions (Bouton, 2004). Below, we simulate a selection of important phenomena in acquisition, extinction, and recovery.

Conditioning with imperfect predictors It has long been recognized that the "associability" of conditioned stimuli (i.e., the ease with which they acquire an association with the US) can vary under different training conditions (Mackintosh, 1975; Pearce & Hall, 1980). For example, according to Pearce and Hall's theory, the surprisingness of a CS increases its associability (see Hall, 1991, for a review of the evidence). The latent cause theory lacks direct CS–US associations, so it must appeal to a different principle to explain these data.

Conceptually, one can think of each latent cause as implicitly encoding an observation prototype (the central tendency of observations generated by that cause), along with an estimate of how observations tend to vary around the prototype. The greater the diversity of observations assigned to the same latent cause, the larger the estimate of the variance will be. The functional consequence of this inflated variance is a higher tolerance for outliers, and novel observations that differ from the prototype are more likely to be assigned to a latent cause when the cause's variance estimate is larger. Paradigms that modestly increase observation diversity (e.g., by using imperfect predictors) encourage the animal to assign new observations to an existing latent cause with high diversity. New learning will be accelerated by the fact that the animal is exploiting earlier knowledge, thereby explaining the apparent associability change.

As an example, Wilson, Boumphrey, and Pearce (1992) trained rats on a serial-conditioning task in which a light was always followed by a tone, which in turn was intermittently paired with a US. Half of the rats (Group C) continued to be trained on this task, while the other half (Group E) were switched to a schedule in which the tone was omitted on all nonreinforced trials, making the light an imperfect predictor of the tone. Subsequently, the light was paired directly with the US for both groups. Wilson et al. observed that Group E acquired conditioned responding to the light more rapidly than did Group C.

Simulations of this experiment using the latent cause model are shown in Fig. 3 (recall that V denotes the animal's

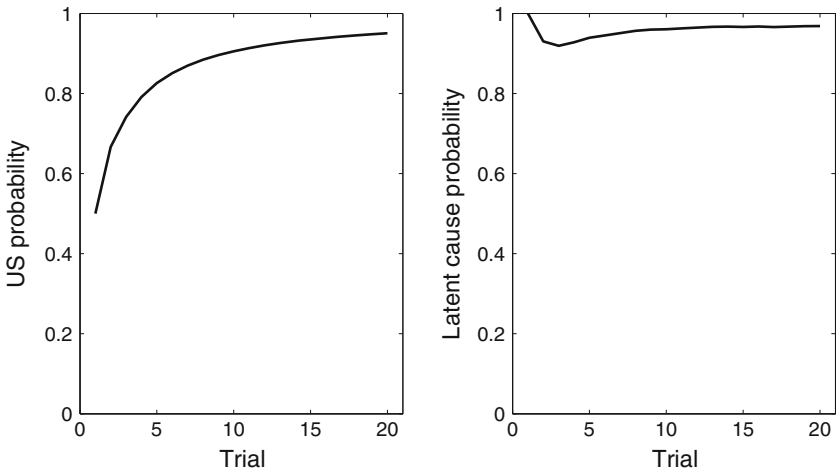

Fig. 2 Simple conditioning: Simulation of the model variables over the course of 20 presentations of the CS–US pair. The time course for each variable is plotted for a single source, averaged over 100 particles. (Left) The conditional probability of the US. (Right) The posterior probability of the latent cause

estimate of the probability that the US will occur). The model is able to capture the findings of Wilson et al. (1992) by virtue of the fact that Group E produced greater diversity within the first inferred latent cause during the

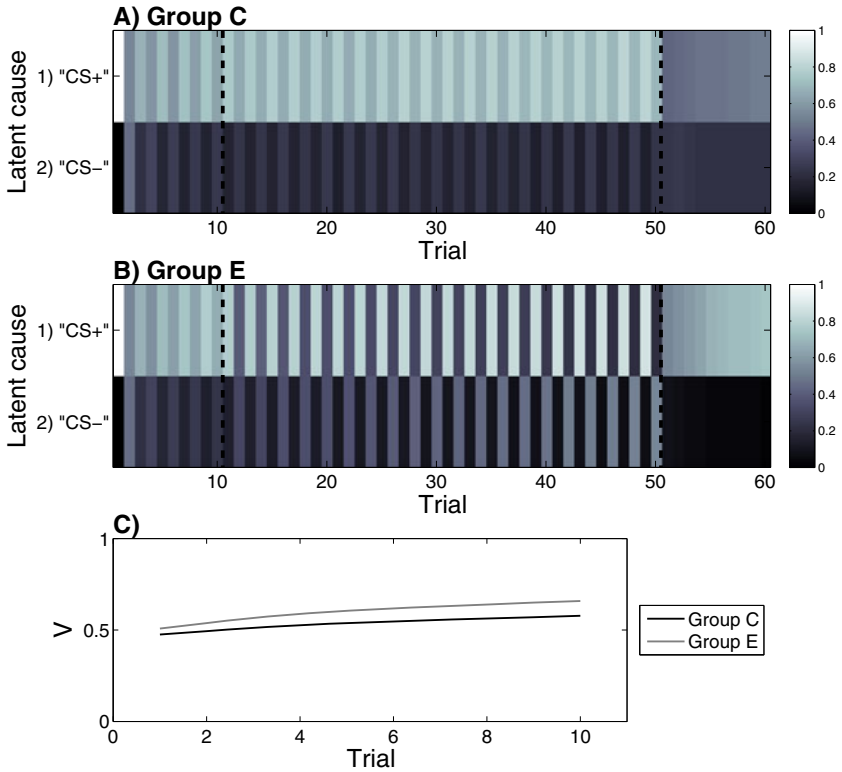

Fig. 3 Conditioning with an imperfect predictor: Simulation of the Wilson, Boumphrey, and Pearce (1992) partially reinforced serial-conditioning paradigm. In the first training phase (Trials 1–10 in our simulation), two groups of rats are presented with a light cue followed by a tone cue, which in turn is intermittently paired with a US (in our simulation, every other trial includes a US). (A–B) Both groups assign the first (reinforced) trial to Cause 1 and the second (nonreinforced) trial to both Causes 1 and 2. Henceforth, both groups assign reinforced trials predominantly to Cause 1 (which is thus associated with high probability of a US) and nonreinforced trials to both Causes 1 and 2 (the second cause being associated with low probability of a US). For illustration purposes, we have labeled the causes "CS+" and "CS–" according to their association with reinforcement. In the second phase (Trials 10–50 in our simulation), Group C continues to be trained on this task, while Group E is switched to a schedule in which the tone is omitted on all nonreinforced trials. In our simulations, this results in a third cause being inferred (not shown here) for Group C, with the new light–no-US trials being assigned to all three causes with some probability. Finally, in a test phase (Trials 50–60), the light is paired directly with the US for both groups. (C) Simulated responding corresponding to the final phase shows greater responding in group E, in agreement with the experimental results. This results from the greater diversity of the trials assigned to Cause 1 in the second phase in Group E (this cause accounts for trials with light, tone, and US; light, tone, and no US; and light and no US), which means that the light-only test trials are assigned to this cause with higher probability, thus bringing about higher expectations for the occurrence of a US

second phase of training (see the figure caption). As a consequence, the light–US trials in the third test phase were more likely to be assigned to that cause. This allowed the animal to use its preexisting predictions about the US to learn more rapidly.

The partial-reinforcement extinction effect One of the most paradoxical findings in classical conditioning is the partial-reinforcement extinction effect (PREE; Capaldi, 1957; Wagner, Siegel, Thomas, & Ellison, 1964): the finding that extinction is retarded following training in which the CS is partially reinforced. This finding is paradoxical from the perspective of most associative-learning theories because one might plausibly expect that extinction should be *faster* after partial reinforcement, since the animal presumably has a weaker association between the CS and the US.

As was first pointed out by Gallistel and Gibbon (2000), the PREE is less paradoxical when considered from a statistical perspective: Discrimination between conditioning and extinction phases is harder when they have similar rates of reinforcement (see also Courville et al., 2006). The latent cause theory offers a similar explanation: The hypothesis that conditioning and extinction phases were generated by different latent causes is less likely in the partial-reinforcement condition, as compared to training with 100 % reinforcement. Confirming this intuition, Fig. 4 shows simulations of the PREE by the latent cause theory.

Renewal In Gershman et al. (2010), we presented an extensive discussion of renewal effects. Briefly, changing the context between acquisition and extinction, or between extinction and test, has the effect of renewing the animal's CR following extinction (see Bouton, 2004, for a review). The latent cause theory explains these phenomena in terms of how contextual manipulations shift the posterior distribution over latent causes. In the most straightforward case (ABA renewal: acquisition in context A, extinction in context B, and test in context A), the context change promotes the inference of one latent cause for acquisition and one for extinction; this effectively protects the acquisition cause from extinction training. Upon return to the acquisition context, the animal infers that the original latent cause is once again active, and therefore renews its prediction that a US will occur.

Generalization and discrimination

In this section, we discuss how our model accounts for generalization from a training set to novel stimuli, as well as discrimination between stimuli in the training set. See Courville et al. (2004) for an alternative latent cause theory of generalization and discrimination.

External inhibition A simple example of generalization is the phenomenon of external inhibition (Pavlov, 1927): Conditioned responding to a (previously trained) CS is decremented when the CS is presented with an added stimulus. The latent cause theory explains this phenomenon as follows: the added stimulus reduces the posterior probability that the trial was generated by the same latent cause as the one that caused the original training trials, and hence the conditioned response is not generalized strongly from the elemental stimulus to the compound. More precisely, the added stimulus causes the animal to place more probability on a new cause, whose US prediction is initialized to .5 (the default prediction in our model, due to the uniform prior over parameters), making it lower than the prediction for the latent cause inferred during training. A simulation of external inhibition is shown in Fig. 5.

Positive and negative patterning Traditional accounts of classical conditioning, such as the Rescorla–Wagner model, generate predictions that are linear in the values of presented

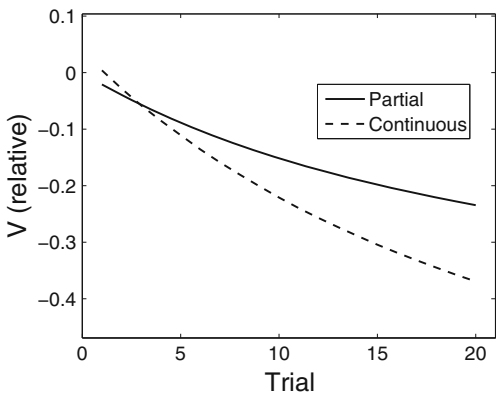

Fig. 4 Partial-reinforcement extinction effect: Simulation of extinction following partial versus continuous reinforcement. Here, V is plotted relative to the end of conditioning

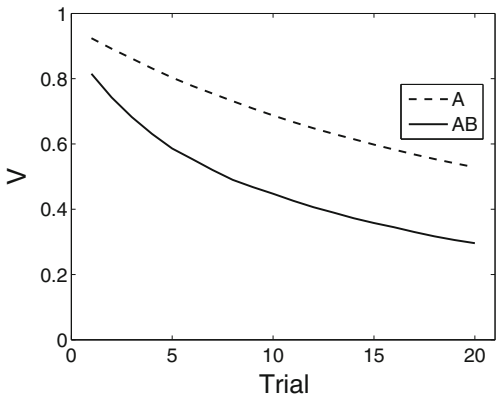

Fig. 5 External inhibition. Simulated conditioned responding is reduced when tested with the novel compound AB after training with CS A alone

stimuli. Positive and negative patterning problems are interesting in that they imply a nonlinear architecture. Bellingham, Gillette-Bellingham, and Kehoe (1985) showed that animals could solve problems of the forms {A–, B–, AB+} (positive patterning) and {A+, B+, AB–} (negative patterning), for which no linear solution exists. The latent cause model can solve these problems as well, as is shown in Fig. 6. The explanation is straightforward: The model can assign each pattern to a different latent cause, allowing it to make different predictions for different configurations. This demonstrates an important computational property of the latent cause model: It can adaptively form predictions that are nonlinear in the stimulus configuration.

Inhibitory conditioning

In this section, we describe simulations of experiments in which stimuli acquire an inhibitory potential. Whereas most earlier models viewed this inhibitory potential in terms of a negative CS–US associative weight that interacts additively with the weights of other CSs, the latent cause theory explains the inhibitory potential in terms of the evidence provided by the conditioned inhibitor for a latent cause that predicts no US.

Conditioned inhibition When two CSs are trained in a feature-negative discrimination (AX+/X–), X becomes a conditioned inhibitor, as assessed by summation and retardation tests (Pavlov, 1927). We shall return to the conditions under which conditioned inhibition arises in the Higher-Order Conditioning section. Here we discuss how various posttraining manipulations influence the conditioned-inhibition effect.

Zimmer-Hart and Rescorla (1974) observed that extinguishing the conditioned inhibitor (X–) following conditioned inhibition training has little effect on its inhibitory potential, whereas pairing it with the US (X+) attenuates its inhibitory potential. Figure 7 shows simulations of this paradigm. Inhibitory potential was assessed by pairing cue X with a test cue B that was separately reinforced during training (a so-called "summation test"). According to the latent cause theory, X– and X+ trials will both be assimilated into the cause containing X– trials that was created during conditioned-inhibition training. Extinction of X– will not greatly change the no-US prediction encoded in that cause, whereas pairing with the US will attenuate it.

It should be noted that this finding has not been reliably replicated in humans. In a causal-learning experiment, Melchers, Wolff, and Lachnit (2006) showed that extinguishing a conditioned inhibitor does reduce its inhibitory potential if the reinforcer can take on negative values. Thus, humans' judgments may be influenced by other kinds of knowledge that do not as strongly affect Pavlovian responses in animals.

While the latent cause theory correctly predicts the effects of posttraining inflation (X+) and deflation (X–) of the conditioned inhibitor, it incorrectly predicts that posttraining inflation of the conditioned excitor should attenuate the inhibitory potential of the conditioned inhibitor. Contrary to this prediction, Amundson, Wheeler, and Miller (2005) have shown that this manipulation actually *enhances* conditioned inhibition. The problem stems from the fact that the model tends to assign the A+ trials to the same latent cause as the AX– trials, and this dilutes the no-US prediction encoded in that cause.

Stimulus competition and potentiation

In this section, we discuss studies in which stimuli trained in compound either compete with or potentiate one another.

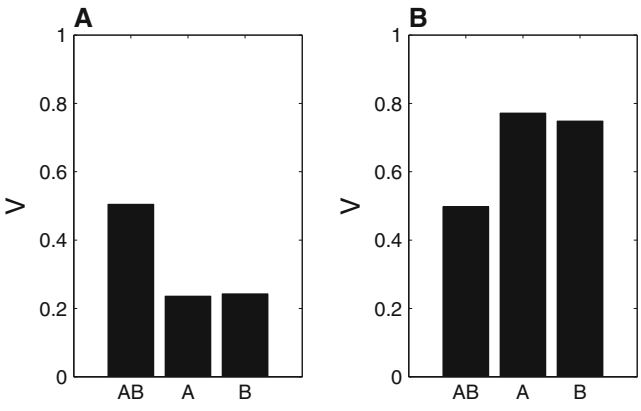

Fig. 6 Positive and negative patterning. (A) Positive patterning: Responding to the compound AB is higher than to each of its elements after {AB+, A– , B –} training. (B) Negative patterning: The opposite is true after {AB–, A+, B+} training

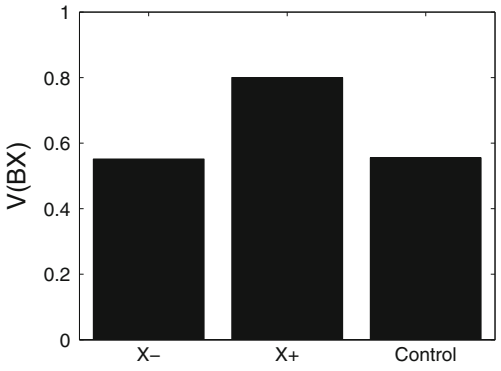

Fig. 7 Extinction of conditioned inhibition: Simulation of conditioned responding during presentation of the compound BX (i.e., a summation test) following conditioned inhibition training (AX–, A+, B+) and one of three posttraining manipulations—extinction of the conditioned inhibitor (X–), reinforcement of the conditioned inhibitor (X+), and no presentation of the conditioned inhibitor (Control)

Overshadowing Overshadowing (Pavlov, 1927) refers to the finding that compound conditioning of two CSs (AB+) results in weaker conditioning to an individual CS (A) than when each CS has been trained alone (A+, B+). The latent cause theory explains this finding in terms of generalization (Fig. 8): The individual CS test trial is more likely to be assigned to the same latent cause as the conditioning phase when training was performed to that CS individually.

Blocking and summation The classic "Kamin blocking effect" refers to the finding that conditioning to a CS1–CS2 compound results in weaker conditioning to CS2 when it is preceded by conditioning to CS1 than when it is not (Kamin, 1968). The latent cause theory fails to account for this finding, because in our theory the CSs do not directly compete to predict the US, an important principle for explaining blocking. In general, the latent cause theory as stated does not assume or result in *summation* of the predictions of different CSs, an important component in models such as Rescorla and Wagner's (1972). Instead of affecting the inferred probability of reward, different stimuli appearing together might increase or decrease the probability that a latent cause is inferred to be active. This inference can affect conditioned responding; for example, after conditioning with a CS1–CS2 compound, presentation of CS1 alone would result in less responding, but this is not because of summation (or subtraction) of predictions of the US.

Similarly, our model does not account well for other phenomena that seem to require summation of predictions. One such example is overexpectation, in which reinforced CS1–CS2 presentations following independent reinforced CS1 and CS2 presentations result in a decrement in their initial associative strengths (Rescorla, 1970). This phenomenon is very naturally explained by the Rescorla–Wagner model in terms of a negative prediction error during the compound presentation. The phenomenon of superconditioning (Rescorla, 1971) is also naturally explained by changes in prediction error due to summation across CSs.

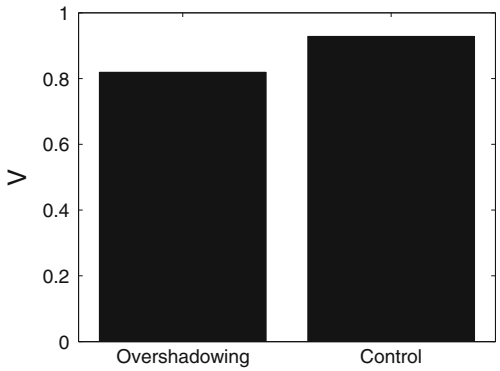

Fig. 8 Overshadowing

Although summation appears to be an important mechanism lacking in our model, it is worth noting that summation leads to erroneous predictions in certain cases. For example, it is well-known that for certain types of stimuli (e.g., from the same sensory modality), summation is not observed (Melchers, Shanks, & Lachnit, 2008). In some cases, stimuli seem to play a more modulatory role, rather than directly summating with other CSs (Holland, 1993), and in others they act more like memory retrieval cues (Bouton, 1993). The work of Courville et al. (2003; Courville et al., 2004, 2006) suggests one way of incorporating summation into a latent cause theory, by allowing for more than one latent cause per trial and for the causes (rather than the stimuli) to summate. Another idea, which we are currently exploring, is to incorporate a linear-Gaussian observation model (e.g., in the style of the Kalman filter; see the Discussion below) into our model.

Preexposure effects

In contrast to effects whose explanation calls on the idea of summation of predicted values, preexposure effects are much more readily explained by our model. In fact, these are the phenomena that are not so easily explained by associative-learning theories like that of Rescorla and Wagner (1972).

Latent inhibition One of the classic preexposure effects, latent inhibition (the CS preexposure effect), is explored at length in Gershman et al. (2010), and therefore we will only discuss it briefly here. In this paradigm, animals are preexposed to the CS prior to pairing it with the US (see Lubow, 1989, for a review). This preexposure retards learning of the CS–US relationship, as compared to a nonpreexposed CS. The latent cause theory explains this result in terms of the animal's inference that the same latent cause was active during the preexposure and conditioning phases. Because the animal learned to predict no US in the preexposure phase, this prediction retards the acquisition of the US prediction in the conditioning phase. This retardation is attenuated, however, when conditioning is performed in a context different from that of preexposure; the context change increases the posterior probability that different latent causes were active in the two phases, thereby releasing the animal from the initial prediction acquired during preexposure.

The Hall–Pearce effect Hall and Pearce (1979) found that acquisition of a CS–US association is slowed if the same CS was previously trained with a weaker US. This phenomenon has subsequently become known as "Hall–Pearce latent inhibition." Figure 9 shows simulations of the latent cause model in this paradigm. The basic explanation offered by

our theory is that the weak-shock training establishes a latent cause that is later reused in the strong-shock training. This retards acquisition because the latent cause is associated with a weak shock, which competes with the strong-shock association. In other words, at the end of acquisition, the posterior is confident that the first latent cause is associated with a weak shock, and this belief is in conflict with the strong-shock association acquired during subsequent training. If a new CS is used, there is a release from latent inhibition, since the two USs are no longer linked to the same latent cause, and hence do not compete with each other (Fig. 9, right). This form of competition is a natural consequence of Bayesian inference, where the probability distribution over hypotheses (e.g., associations) must sum to 1, and therefore increasing the posterior probability of one hypothesis must necessarily decrease the posterior probability of another hypothesis.

Preexposure effects not accounted for by the theory A number of preexposure effects are not accounted for by the latent cause theory. For example, preexposing an animal to the conditioning context facilitates subsequent contextual fear conditioning (Kiernan & Westbrook, 1993). This phenomenon is paradoxical in light of the latent-inhibition effect described above: If the context is treated like a cue, it should also exhibit latent inhibition. Clearly, then, contexts are not (or at least not always) treated as cues. Because the latent cause theory in its present form makes the assumption that contexts are cues, it cannot explain the context preexposure effect. On the other hand, Fanselow (1990) has shown that longer context preexposure can produce latent inhibition (rather than facilitation), suggesting that contexts can act as cues under certain temporal conditions in the same paradigm. This phenomenon is not captured by our model presently.

Another type of preexposure effect that challenges the latent cause theory is *learned irrelevance* (Bonardi & Hall, 1996): Random exposure to the CS and the US retards conditioning even more than the combination of CS preexposure ("latent inhibition") and US preexposure. The latent cause theory cannot account for this finding because of its assumption of *exchangeability*: Randomly permuting the order of trials leaves the probability of the whole sequence of data unchanged, and thus does not affect the US prediction. This means that interleaving random exposure to the CS and US should have the same effect as blocking the presentations of each trial type, contrary to the learned-irrelevance effect. In the Discussion, we return to ways in which the exchangeability assumption can be relaxed.

Finally, perceptual learning—that is, exposure to similar stimuli leads to faster subsequent acquisition of a discrimination between them (Channell & Hall, 1981)—is not anticipated by the latent cause theory. This is because, in our model, the similar stimuli tend to be assigned to the same latent cause, retarding later acquisition of a discrimination between them.

Higher-order conditioning

Second-order conditioning and conditioned inhibition A peculiar contradiction for a long time existed in the classical conditioning literature: Almost-identical treatments (second-order conditioning and conditioned inhibition) seemed to produce opposite results. Both treatments involve presentations of a nonreinforced compound AB– interleaved with reinforced presentations of the individual CS A+. Whereas conditioned inhibition results in B becoming inhibitory (Pavlov, 1927), second-order conditioning results in B becoming excitatory (Rizley & Rescorla, 1972). This contradiction was resolved by Yin, Barnet, and Miller (1994), who showed that with a small number of conditioning trials, one obtains second-order conditioning, whereas with a large number of conditioning trials, one obtains conditioned inhibition.

The latent cause theory explains this finding because, although fewer causes are preferred by the animal a priori, this simplicity preference can be overcome by observing

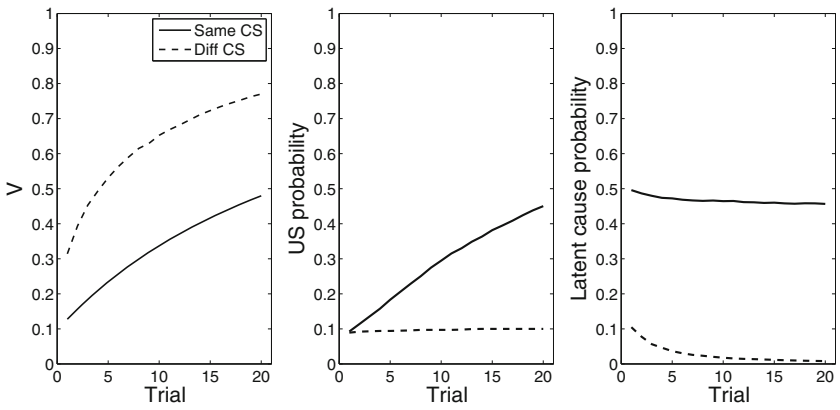

Fig. 9 The Hall–Pearce effect. (Left) Simulated conditioned responding to a CS trained with a strong shock, following training in which the same CS was paired with a weaker shock ("Same CS"). In the "Diff CS" condition, training is with a novel CS. (Middle) Conditional probability of the US under the first ("weak-shock") latent cause during strong-shock training. (Right) Posterior probability of the first latent cause during strong-shock training

more data (see also Courville et al., 2003). With few conditioning trials, the animal assigns both AB– and A+ to the same latent cause, resulting in the US prediction generalizing to CS B. With more conditioning trials, the animal accumulates evidence in favor of a more complex internal model consisting of two latent causes, one for each trial type. In this case, CS B is assigned to a cause predicting no US (Fig. 10). As mentioned above, this prediction of no US is, however, not "inhibitory" in the sense of summation, as our model does not involve summation of US predictions.

Discussion

We have shown that a simple normative theory of classical conditioning that is based on the idea that animals reason about latent causes can account for a wide range of experimental data. This is surprising for two reasons. First, the theory departs from some basic assumptions of most theories of conditioning. Below we dissect these departures, as well as the relationship of our theory to other normative accounts. Second, the theory is clearly oversimplified, as evidenced by its failure to account for a number of important phenomena. We will suggest several directions for extending and improving the theory.

Relationship to other theories

It is worth emphasizing the radicalness of the latent cause theory's departure from the major contemporary theories of classical conditioning (Mackintosh, 1975; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Schmajuk, 2010; Wagner, 1981), all of which posit that conditioned responding arises from learned associations between CSs and USs (but see Gallistel & Gibbon, 2000). According to the latent cause theory, animals instead learn associations between hypothetical latent causes and observation features (CS, US, contextual stimuli, etc.); in other words, in our model the animal's predictions about the US are *mediated* by its beliefs about which latent causes are active. The model is normative because we assume that these beliefs are updated whenever a new observation is obtained, in a statistically rational manner, by means of Bayesian inference. Conditioned responding arises not from direct CS–US associations, but rather from cause–US associations, averaged under the animal's posterior probability distribution over latent causes.

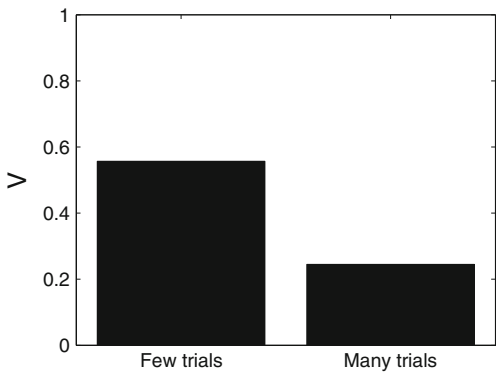

Fig. 10 Second-order conditioning and conditioned inhibition: Simulated conditioned responding to B after either few or many trials of AB–/A+ training. See the text for details

It has long been recognized that some cognitive processes operating in classical conditioning defy explanations in terms of CS–US associations (Holland, 1993; Tolman, 1948). Bouton (1993) has argued persuasively that certain types of stimuli—in particular, contextual stimuli—act to facilitate retrieval rather than to directly excite or inhibit conditioned responding. A complementary perspective is offered by Miller's comparator hypothesis (Stout & Miller, 2007), according to which associations compete in memory for control of conditioned responding at the time of test. In contrast to encoding-focused models (e.g., Rescorla & Wagner, 1972), which place the explanatory onus on processes occurring at the time of training, retrieval-focused models like the comparator hypothesis seem better suited to explain the effects of posttraining manipulations on the response to a cue without actually presenting the cue again (but see Schmajuk & Larrauri, 2006).

Our latent cause theory attempts to place some of these ideas in a statistical perspective. The probabilistic computations underlying Bayesian inference can be understood as a kind of memory retrieval process: The animal is attempting to match the current observation to the prototypes stored in memories of the latent causes that it has previously inferred. If a match is found, or several are, the animal updates the latent cause memories to reflect the current observation; if no memory matches sufficiently, the animal encodes the current observation into a new memory (see also Redish et al., 2007).

A more elaborate latent cause theory has been explored by Courville and colleagues (Courville et al., 2003; Courville et al., 2004, 2006) using sigmoid belief networks. In their model, a variety of latent causes could be active in any given trial, with the data from all trials used to infer a network-like structure of causes and their associated observations. Space precludes us from discussing the detailed similarities and differences between Courville's theory and ours; however, these theories are meant to capture different aspects of conditioning. Courville's theory was motivated by elemental and configural theories of conditioning and explores the idea that animals use rational statistical principles to decide what configurations of stimuli should be learned about. In contrast, our theory was developed as a formal distillation of the idea that some circumstances promote the formation of a new memory, whereas other conditions promote the modification of an old memory (Gershman

et al., 2010). Unlike the Courville et al. (2004) model, our model does not assume that a single causal structure underlies all of the observations (but see Courville et al., 2006, for an elaboration in which latent causes evolve through a birth–death process). The nonparametric prior over latent causes is explicitly designed to allow different latent causes to explain different observations, and these latent causes belong to a potentially infinite set. Our model is thus better suited to explaining phenomena like context-dependent renewal following extinction, which admits a natural interpretation in terms of different latent causes assigned to training and extinction (Gershman et al., 2010).

The simplicity of our theory makes it useful for investigating the idea of when a new memory is formed versus an old one modified. This, however, comes at the cost of not capturing the full richness of classical conditioning. Nonetheless, we have shown that our theory can still capture a remarkably wide array of findings. In sum, we see our model and Courville's (Courville et al., 2003; Courville et al., 2004, 2006) as two instances of a more general modeling paradigm within which particular assumptions can be tested and criticized. Its key theoretical commitment is the explanatory concept of a latent cause; the other assumptions are auxiliary (i.e., not central to the theory, but necessary for it to make quantitative predictions). Our contribution in this article has been to explore one set of auxiliary assumptions within this framework.

As mentioned in the Preexposure Effects section, the latent cause theory we have presented assumes exchangeability (invariance to ordering of the observations). Courville et al. (2006) discussed evidence that this assumption is not viable, and they presented a nonexchangeable latent cause theory as an alternative. Other work by Kakade and Dayan (Dayan et al., 2000; Kakade & Dayan, 2002) has borrowed an idea from engineering theory (the Kalman filter) to model classical conditioning as optimal online inference in a particular kind of dynamical system. Their model captures certain types of learning dynamics but does not use the concept of latent causes. In fact, the Kalman filter model can be seen as an extension of the Rescorla–Wagner (1972) and Pearce–Hall (1980) theories, and as such has some of the same limitations, such as failing to explain phenomena like the partial-reinforcement extinction effect, which are explained naturally under a latent cause account (Courville et al., 2006).

Finally, while we have used the word "cause," we do not wish to make strong causal interpretations of our model. Rather, latent causes are simply useful explanatory constructs posited by the animal. This is in contrast to work in the causal-learning literature (e.g., Griffiths & Tenenbaum, 2005), where a rigorous probabilistic interpretation of causality is explored. There is evidence that causal learning plays an important role in classical conditioning (Beckers, Miller, De Houwer, & Urushihara, 2006), but we do not address these phenomena in the present work.

Limitations and extensions

In this article, we have omitted the important area of within-trial temporal effects. These effects emerge from manipulations of the intertrial and interstimulus intervals, including distinctions between serial- and simultaneous-conditioning paradigms. Like Rescorla and Wagner's (1972) model, the latent cause theory we have presented is a *trial-level* model, treating the entire trial as a single time point; hence, within-trial timing effects lie beyond its scope. For a similar reason, the theory also does not explain the conditioned diminution of the unconditioned response after presentation of the CS with which the US was trained (Donegan, 1981), which is essentially a priming phenomenon requiring a real-time treatment. It is, however, possible to integrate the latent cause theory presented here with real-time normative theories, in particular the temporal-difference-learning theory (Sutton & Barto, 1990). We are actively exploring this direction. Courville et al. (2006) proposed a latent cause theory of within-trial temporal structure, using a mixture model over Markov chains representing different stimulus sequences. This model could be fruitfully integrated with our own work by replacing the finite mixture model with its infinite counterpart, the infinite-capacity mixture model.

A quite different class of temporal effects concerns the dynamics of trial-level phenomena—that is, how the statistics of observations change over trials. Our model makes the generative assumption that both the distribution over latent causes and the properties of latent causes do not change over trials. This assumption is unrealistic, and several authors have suggested ways in which to model gradual change in the observation statistics, as exemplified by the Kalman filter model (Behrens, Woolrich, Walton, & Rushworth, 2007; Dayan et al., 2000; Kakade & Dayan, 2002). The assumption of gradual change is not incompatible with our model; a cause's distribution over observations could be allowed to change. In addition, one could modify the generative model to allow the distribution over latent causes to change gradually. These developments are the subject of future work.

Conclusions

The latent cause theory explored in this article is clearly inadequate as a general model of classical conditioning. However, as we argued at the beginning of the article, the value of a normative theory lies primarily in its ability to reveal the logic of behavior. The latent cause theory offers a new way of conceptualizing the logic of classical conditioning, one that

will hopefully furnish new directions for experimental research. We hope to elaborate and extend these tentative first steps to provide a more general theory of classical conditioning. These elaborations include allowing the latent cause parameters to drift over time (as in the Kalman filter model) and making the latent cause prior sensitive to the temporal order of observations.

Author note We are grateful to Elliot Ludvig and Nathaniel Daw for illuminating discussions, and to David Blei for collaboration on earlier incarnations of this work. We also thank Will Brinkman and Deepa Patil for assistance with the simulations, and the reviewers for astute criticism that has improved the article. S.J.G. is supported by a Graduate Research Fellowship from the NSF. Y.N. is supported by a Sloan Research Fellowship.

References

Amundson, J., Wheeler, D., & Miller, R. (2005). Enhancement of Pavlovian conditioned inhibition achieved by posttraining inflation of the training excitor. *Learning and Motivation, 36,* 331–352.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98,* 409–429. doi:10.1037/0033-295X.98.3.409

Beckers, T., Miller, R. R., De Houwer, J., & Urushihara, K. (2006). Reasoning rats: Forward blocking in Pavlovian animal conditioning is sensitive to constraints of causal inference. *Journal of Experimental Psychology. General, 135,* 92–102. doi:10.1037/0096-3445.135.1.92

Behrens, T., Woolrich, M., Walton, M., & Rushworth, M. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10,* 1214–1221.

Bellingham, W., Gillette-Bellingham, K., & Kehoe, E. (1985). Summation and configuration in patterning schedules with the rat and rabbit. *Learning & Behavior, 13,* 152–164.

Bonardi, C., & Hall, G. (1996). Learned irrelevance: No more than the sum of CS and US preexposure effects? *Journal of Experimental Psychology. Animal Behavior Processes, 22,* 183–191. doi:10.1037/0097-7403.22.2.183

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin, 114,* 80–99. doi:10.1037/0033-2909.114.1.80

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

Brown, S., & Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology, 58,* 49–67.

Capaldi, E. J. (1957). The effect of different amounts of alternating partial reinforcement on resistance to extinction. *The American Journal of Psychology, 70,* 451–452.

Channell, S., & Hall, G. (1981). Facilitation and retardation of discrimination learning after exposure to the stimuli. *Journal of Experimental Psychology. Animal Behavior Processes, 7,* 437–446. doi:10.1037/0097-7403.7.4.437

Courville, A. C., Daw, N. D., Gordon, G. J., & Touretzky, D. S. (2003). Model uncertainty in classical conditioning. In *Advances in neural information processing systems* (Vol. 16, pp. 977–984). Cambridge, MA: MIT Press.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2004). Similarity and discrimination in classical conditioning: A latent variable account. In *Advances in neural information processing systems* (Vol. 17, pp. 313–320). Cambridge, MA: MIT Press.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences, 10,* 294–300.

Daw, N. D., & Courville, A. C. (2008). The pigeon as particle filter. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 369–376). Cambridge, MA: MIT Press.

Dayan, P., Kakade, S., & Montague, P. (2000). Learning and selective attention. *Nature Neuroscience, 3,* 1218–1223.

Dayan, P., & Long, T. (1998). Statistical models of conditioning. In *Advances in neural information processing systems* (Vol. 10, pp. 117–124). Cambridge, MA: MIT Press.

Donegan, N. (1981). Priming-produced facilitation or diminution of responding to a Pavlovian unconditioned stimulus. *Journal of Experimental Psychology. Animal Behavior Processes, 7,* 295–312.

Fanselow, M. (1990). Factors governing one-trial contextual conditioning. *Learning & Behavior, 18,* 264–270.

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Journal of Statistics and Computing, 14,* 11–21.

Gallistel, C., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review, 107,* 289–344.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2004). *Bayesian data analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.

Gershman, S. J., & Blei, D. M. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology, 56,* 1–12. doi:10.1016/j.jmp. 2011.08.004

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review, 117,* 197–210.

Griffiths, T. L., & Tenenbaum, J. (2005). Structure and strength in causal induction. *Cognitive Psychology, 51,* 334–384.

Hall, G. (1991). *Perceptual and associative learning*. New York, NY: Oxford University Press.

Hall, G., & Pearce, J. M. (1979). Latent inhibition of a CS during CS–US pairings. *Journal of Experimental Psychology. Animal Behavior Processes, 5,* 31–42. doi:10.1037/0097-7403.5.1.31

Holland, P. C. (1993). Cognitive aspects of classical conditioning. *Current Opinion in Neurobiology, 3,* 230–236.

Kakade, S., & Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychological Review, 109,* 533–544. doi:10.1037/0033-295X.109.3.533

Kamin, L. (1968). Attention-like processes in classical conditioning. In M. R. Jones (Ed.), *Miami Symposium on the Prediction of Behavior, 1967: Aversive stimulation* (pp. 9–31). Miami, FL: University of Miami Press.

Kiernan, M., & Westbrook, R. (1993). Effects of exposure to a to-be-shocked environment upon the rat's freezing response: Evidence for facilitation, latent inhibition, and perceptual learning. *Quarterly Journal of Experimental Psychology, 46B,* 271–288.

Lubow, R. (1989). *Latent inhibition and conditioned attention theory.* Cambridge, U.K.: Cambridge University Press.

Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review, 82,* 276–298. doi:10.1037/h0076778

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. San Francisco, CA: W. H. Freeman.

Melchers, K., Shanks, D., & Lachnit, H. (2008). Stimulus coding in human associative learning: Flexible representations of parts and wholes. *Behavioural Processes, 77,* 413–427.

Melchers, K. G., Wolff, S., & Lachnit, H. (2006). Extinction of conditioned inhibition through nonreinforced presentation of the inhibitor. *Psychonomic Bulletin & Review, 13,* 662–667. doi:10.3758/BF03193978

Pavlov, I. P. (1927). *Conditioned reflexes (G. V. Anrep, Trans)*. London, UK: Oxford University Press.

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review, 87,* 532–552. doi:10.1037/0033-295X.87.6.532

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

Rescorla, R. A. (1970). Reduction in the effectiveness of reinforcement after prior excitatory conditioning. *Learning and Motivation, 1,* 372–381.

Rescorla, R. A. (1971). Variation in the effectiveness of reinforcement and nonreinforcement following prior inhibitory conditioning. *Learning and Motivation, 2,* 113–123. doi:10.1016/0023-9690(71)90002-6

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Rizley, R. C., & Rescorla, R. A. (1972). Associations in second-order conditioning and sensory preconditioning. *Journal of Comparative and Physiological Psychology, 81,* 1–11. doi:10.1037/h0033333

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

Savastano, H. I., & Miller, R. R. (1998). Time as content in Pavlovian conditioning. *Behavioural Processes, 44,* 147–162.

Schmajuk, N. A. (2010). *Mechanisms in classical conditioning: A computational approach*. New York, NY: Cambridge University Press.

Schmajuk, N. A., Lam, Y.-W., & Gray, J. A. (1996). Latent inhibition: A neural network approach. *Journal of Experimental Psychology. Animal Behavior Processes, 22,* 321–349. doi:10.1037/0097-7403.22.3.321

Schmajuk, N. A., & Larrauri, J. A. (2006). Experimental challenges to theories of classical conditioning: Application of an attentional model of storage and retrieval. *Journal of Experimental Psychology. Animal Behavior Processes, 32,* 1–20. doi:10.1037/0097-7403.32.1.1

Sokolov, E. N. (1960). Neuronal models and the orienting reflex. In M. A. B. Brazier (Ed.), *The central nervous system and behavior* (pp. 187–276). New York, NY: Josiah Macy Jr Foundation.

Stout, S. C., & Miller, R. R. (2007). Sometimes-competing retrieval (SOCR): A formalization of the comparator hypothesis. *Psychological Review, 114,* 759–783. doi:10.1037/0033-295X.114.3.759

Sutton, R., & Barto, A. (1990). Time-derivative models of Pavlovian reinforcement. In M. Gabriel & J. Moore (Eds.), *Learning and computational neuroscience: Foundations of adaptive networks* (pp. 497–537). Cambridge, MA: MIT Press.

Tolman, E. (1948). Cognitive maps in rats and men. *Psychological Review, 55,* 189–208.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5–47). Hillsdale, NJ: Erlbaum.

Wagner, A. R., Siegel, S., Thomas, E., & Ellison, G. D. (1964). Reinforcement history and the extinction of conditioned salivary response. *Journal of Comparative and Physiological Psychology, 58,* 354–358. doi:10.1037/h0048721

Wilson, P. N., Boumphrey, P., & Pearce, J. M. (1992). Restoration of the orienting response to a light by a change in its predictive accuracy. *Quarterly Journal of Experimental Psychology, 44B,* 17–36. doi:10.1080/02724999208250600

Yin, H., Barnet, R. C., & Miller, R. R. (1994). Second-order conditioning and Pavlovian conditioned inhibition: Operational similarities and differences. *Journal of Experimental Psychology. Animal Behavior Processes, 20,* 419–428. doi:10.1037/0097-7403.20.4.419

Zimmer-Hart, C. L., & Rescorla, R. A. (1974). Extinction of a Pavlovian conditioned inhibition. *Journal of Comparative and Physiological Psychology, 86,* 837–845. doi:10.1037/h0036412