Received: 22 June 2020 | Revised: 19 November 2020 | Accepted: 28 December 2020

DOI: 10.1002/hipo.23298

RESEARCH ARTICLE

WILEY

Extinction and discrimination in a Bayesian model of context fear conditioning (BaconX)

Franklin B. Krasne[1] | Raphael Zinn[2] | Bryce Vissel[3,4] | Michael S. Fanselow[1,5,6]

[1]Department of Psychology and Brain Research Institute, University of California, Los Angeles, Los Angeles, California

[2]Centre for Neuroscience and Regenerative Medicine, Faculty of Science, University of Technology Sydney, Sydney, New South Wales, Australia

[3]Centre for Neuroscience and Regenerative Medicine, Faculty of Science, University of Technology Sydney, Ultimo, New South Wales, Australia

[4]St Vincent's Centre for Applied Medical Research, St Vincent's Health Network Sydney, Darlinghurst, New South Wales, Australia

[5]Staglin Center for Brain and Behavioral Health, University of California, Los Angeles, Los Angeles, California

[6]Department of Psychiatry and Biobehavioral Sciences, University of California, Los Angeles, Los Angeles, California

Correspondence
Franklin B. Krasne, Department of Psychology, University of California, Los Angeles, Los Angeles, CA 90095.
Email: krasne@psych.ucla.edu

Funding information
Australian Research Council, Grant/Award Number: 200102445; National Institute of Mental Health, Grant/Award Number: RO1-MH62122; The Boyarsky Family Trust, Tony and Vivian Howland-Rose, Doug Battersby, and David King

Abstract

The extinction of contextual fear is commonly an essential requirement for successful exposure therapy for fear disorders. However, experimental work on extinction of contextual fear is limited, and there little or no directly relevant theoretical work. Here, we extend BACON, a neurocomputational model of context fear conditioning that provides plausible explanations for a number of aspects of context fear conditioning, to deal with extinction (calling the model BaconX). In this model, contextual representations are formed in the hippocampus and association of fear to them occurs in the amygdala. Representation creation, conditionability, and development of between-session extinction are controlled by degree of confidence (assessed by the Bayesian weight of evidence) that an active contextual representation is in fact that of the current context (i.e., is "valid"). The model predicts that: (1) extinction which persists between sessions will occur only if at a sessions end there is high confidence that the active representation is valid. It follows that the shorter the context placement-to-US (shock) interval ("PSI") and the less is therefore learned about context, the longer extinction sessions must be for enduring extinction to occur, while too short PSIs will preclude successful extinction. (2) Short-PSI deficits can be rescued by contextual exposure even after conditioning has occurred. (3) Learning to discriminate well between a conditioned and similar safe context requires representations of each to form, which may not occur if PSI was too short. (4) Extinction-causing inhibition must be applied downstream of the conditioning locus for reasonable generalization properties to be generated. (5) Context change tends to cause return of extinguished contextual fear. (6). Extinction carried out in the conditioning context generalizes better than extinction executed in contexts to which fear has generalized (as done in exposure therapy). (7) BaconX suggests novel approaches to exposure therapy.

KEYWORDS
amygdala, classical, conditioning, extinction, fear, hippocampus

1 | INTRODUCTION

In recent years, contextual fear conditioning has become an important topic of study. This is both because the mechanisms involved in forming contextual representations during context fear conditioning are thought to be similar to those underlying hippocampal representations of episodic memories and because contextual fear likely plays an important role in fear-related disorders. Indeed, context is increasingly entering center stage in modern models of disorders like PTSD (Brewin, Gregory, Lipton, & Burgess, 2010; Liberzon & Abelson, 2016; Maren, Phan, & Liberzon, 2013). While there has been much empirical and some relevant theoretical work on context fear conditioning itself, the extinction of context fear has been less studied experimentally and has not been addressed at all theoretically. This is important because extinction plays a critical role in ensuring fear does not become maladaptive. The purpose of this article is to begin to fill this gap.

There is considerable evidence that representations of the conjunctions of attributes that define a context are formed in the hippocampus and that these then become associated with aversive consequences in the amygdala (Fanselow, 2000). Thinking about the role of hippocampus in both episodic memory generally, and context fear conditioning in particular, derives substantially from Marr's Theory of Archicortex (Marr, 1971). Marr proposed that the neocortical activity which encodes a to-be-remembered event gets rerepresented at the level of entorhinal cortex (EC), the active cells of which recruit a small group of hippocampal cells that become the hippocampal representation of the event. The synapses of the active entorhinal cells on these hippocampal neurons and synapses in a pathway back to EC then undergo Hebbian potentiation with the consequence that thereafter even a fragment of the original neocortical pattern can activate the full hippocampal representation and, via it, activation of a neocortical pattern of activity that produces a memory of the full event.

Given this, one might expect that a US following a CS would cause formation of an episodic memory of this sequence so that in the future, the CS, as one fragment of the event, would evoke a memory of the full event including the US, and this would be the basis for the conditional response. In that case, one would expect all fear conditioning to be hippocampus-dependent. However, this is not the case. Learning to become afraid of a simple, explicit cue such as a tone or light that is followed by a shock does not require the hippocampus (Kim & Fanselow, 1992) and instead seems to involve potentiation of synapses of neocortical and thalamic inputs directly onto amygdala cells (Blair, Schafe, Bauer, Rodrigues, & Ledoux, 2001; Doron & Ledoux, 1999; McDonald, 1998). *Context fear conditioning*, however, *does* depend on the hippocampus, as well as the amygdala (Helmstetter, 1992; Kim & Fanselow, 1992; Kim, Rison, & Fanselow, 1993; Phillips & LeDoux, 1992). The hippocampus appears to be critically involved in the formation of contextual representations, while the basolateral amygdala is crucially involved in associating these representations with affect (Fanselow, 1982, 1990, 2000; Fanselow & Gale, 2000; Fanselow & Gale, 2000; Rudy et al., 2004; Stote & Fanselow, 2004; Young, Bohenek, & Fanselow, 1994; Zelikowsky, Hersman, Chawla, Barnes, & Fanselow, 2014). Subsequent work from the Bucci laboratory established that retrosplenial, postrhinal, and perirhinal cortices also provided critical extrahippocampal junctures in the processing needed for contextual fear conditioning (e.g., Bucci, Saddoris, & Burwell, 2002; Fournier, Eddy, DeAngeli, Huszár, & Bucci, 2019), and of course there has been much discussion of the role of interactions between Hippocampus, prefrontal cortex, and amygdala in fear expression and extinction (e.g. Giustino & Maren, 2015; Milad & Quirk, 2012; Orsini, Kim, Knapska, & Maren, 2011). The exact roles of the various cortical structures involved are still being resolved.

Although Marr's ideas have been applied extensively to episodic learning in general, they have only been directly applied to context fear learning in a limited way. Perhaps most notable is the work of O'Reilly and Rudy which has used versions of these ideas to explain the intriguing phenomenon of false conditioning (O'Reilly & Rudy, 2001). However, recently Marr's ideas led to the construction of BACON (Bayesian Context Fear Algorithm; Krasne, Cushman, & Fanselow, 2015), a computationally implemented neural model of context fear conditioning that creates and activates representations of contexts in much the same way as Marr envisaged for representations of episodic events, but goes one step further. It makes a comparison between its memory and what is currently being observed, and based on this comparison it generates an estimate of how certain it can be that the recalled context really is the one it is now in. *It then uses its estimate of memory validity to control aspects of BACON's further hippocampal representation formation as well as formation of associations to those representations in the amygdala*. With these additional features, BACON can simulate a wide range of context fear phenomenology including the immediate shock deficit (Fanselow, 1982), the gradual increase in expressed fear during exposure to a feared context (Bae, Holmes, & Westbrook, 2015; Fanselow, 1982; Lester & Fanselow, 1986; Lingawi et al., 2018; Wiltgen, Sanders, Anagnostaras, Sage, & Fanselow, 2006), and false conditioning (Rudy & O'Reilly, 1999; Rudy et al., 2002). It has also provided a basis for explaining seeming contradictions in experiments on the effects DG suppression during encoding and recall (Bernier et al., 2017), and it has provided plausible explanations for a number of nonintuitive effects of conditioning that was carried out at short intervals after placement in a context (Zinn et al., 2020). Moreover, it abolishes the so-called "tradeoff between pattern separation" and completion" (O'Reilly & McClelland, 1994; O'Reilly & Rudy, 2001) that arises in the absence of some reasonable mechanism to control when new representations should and should not be created.

Thus, the BACON model copes well with context fear learning, and its additions to Marr's basic ideas resolve certain limitations of the original model. However, neither BACON nor previous models deal with extinction. Of course when considering episodic memory, "extinction" as such does not seem like a meaningful concept, though such memories can presumably be degraded by forgetting or perhaps altered somewhat by further experience with similar situations. Extinction certainly is relevant, however, when considering context fear, especially from a clinical perspective. This is because traumatic events in humans are likely associated with complex situations and not simple precise predictive cues. Additionally, the rules of extinction for contexts and cues are likely to

be different because cued fear extinction is defined by its modulation by contexts whereas there is no obvious modulatory influence for contexts (though as we will see, the current model does predict, for context fear, phenomena that are analogous to cued fear renewal). Thus, the purpose of the present effort was to add extinction and discrimination learning mechanisms to the basic BACON model. We call the extended model "BaconX" (X for extinction).

In both BACON and BaconX different processes require different degrees of confidence in memory validity. Thus, the confidence required to allow fear conditioning is relatively low whereas the confidence required to allow the association of new contextual attributes to an existing representation ("updating") is set quite high. This design feature reflects the hypothesis that it is probably adaptive to associate fear with a contextual representation even if there is only modest certainty that it is valid. In contrast, if incorrect attributes were to become associated with a representation, accurate recall of that representation would become irrevocably damaged forever more. In BaconX somewhat similar logic applies to the extinction that we are attempting to model. The logic dictates that a high degree of conviction as to the validity of a recalled representation should be required to allow permanent extinction of context fear. For it would be highly maladaptive if an animal were to extinguish fear to a dangerous place as the result of being in a safe place that was similar. Interestingly, it has recently been found that when animals are conditioned to a shock after having been in a context for too short a time to have observed very much about the place, the fear they learn cannot be extinguished (Zinn et al., 2020). This is exactly what the sort of thinking used in the construction of BACON would predict. For if an animal has not learned much about a context of which it has become afraid, in the future it will be difficult for it to be sure that what seems like the same place, but now absent shock, really is the same place and that therefore it is safe to extinguish its fear. The present study describes how this logic plays out within BaconX, how extinction is affected by the degree of initial learning and by the context in which extinction occurs, and what basic and translational implications this might have.

2 | METHODS

BaconX's hippocampal model is almost the same as that of its predecessor, BACON, which is described fully in Krasne et al. (2015); its basic features are reviewed at the start of Section 3 and a few aspects that are relatively technical or had to be changed to deal with extinction are described in the present section. The circuitry of the amygdala, which must now deal with the inhibitory conditioning that mediates extinction, is more elaborate than that in BACON. The logic of BaconX's amygdala design is explained in Section 3, but some technical background is given here.

2.1 | BaconX's amygdala neurons

Excitation and inhibition affect Bacon's amygdala cells in a way that is somewhat more realistic than sometimes assumed in the neuron models of connectionist literature, and this has important consequences for the functioning of the circuit. BaconX amygdala cells have a single compartment with the conventional equivalent circuit membrane and synapse model (as explained, e.g., in Kandel, Schwartz, and Jessell (2000), chaps. 7, 10–12). We take resting potential as 0 because it simplifies formulas. Excitatory input causes postsynaptic conductances that depolarize the neurons, whereas inhibitory input produces conductances that move the membrane potential toward its resting level. We take the membrane potential toward which excitatory input drives membrane potential (the excitatory reversal potential, *E*) as 100 and let inhibitory input drive membrane potential toward 0 (the inhibitory reversal potential). Firing rates of neuron **n**, which we often refer to as the "activation" or "activity" of the neuron, we denote as A_n. The strength or "weight" of a synapse determines how great a conductance increase will be produced by a given presynaptic firing rate; denoting as *G* the conductance (*note that all conductances are taken as relative to the resting "leakage" conductance of the neuron*) produced by synaptic input of strength *W* from a neuron with activation *A*,

$$G = A \cdot W. \tag{1}$$

Denoting the total excitatory conductance of a postsynaptic cell as *Ge* and the total inhibitory conductance as *Gi*, the postsynaptic depolarization *V* is given by

$$V = E \cdot Ge/(1 + Ge + Gi). \tag{2}$$

Note that effect of both excitation and inhibition are very nonlinear. Increasing *Ge* from 1 to 2 increases *V* by about 30% whereas increasing *Ge* from 40 to 80 (also a doubling) causes *V* to increase by only about 1%. When *Ge* = 5 the effect of *Gi* = 3 is to drop *V* by about 35% whereas if *Ge* = 30 it drops it by about 6%. As we will discuss below, these nonlinearities have significant implications for generalization of fear and extinction.

We assume throughout that the firing rate (activation, *A*) produced by a depolarization *V* is proportional to the depolarization above a threshold (*Thrsh*) reaching a maximum of unity at depolarization level "*Mxat*." We refer to this as a "linear sigmoid" function which we write

$$A = Linsig\,(V\,|\,Thrsh, Mxat). \tag{3}$$

2.2 | Technical aspects of BaconX's hippocampal model

2.2.1 | Evaluating degree of confidence in an active representation's correctness (or "validity")

The current mode of operation of the hippocampus (recall, create, or update) is determined by degree of confidence in the currently active

representation's correctness. Degree of confidence in a representation's validity is indexed by the Bayesian weight of evidence (Kass & Rafter, 1995; Krasne et al., 2015), which we refer to as "B_{Rep}." Suppose that as the result of being placed in a certain context an individual were to activate a contextual representation that had associated with it a certain number (Z_{rec}) of recalled attributes and that so far during this visit the individual had observed Z_{cur} of the context's attributes. The individual could then use the number of attributes in common between its Z_{rec} recalled attributes and its Z_{cur} currently observed ones to get a sense of how likely it was that the memory it had activated was the right one. According to our model, when placed in a context, BACON activates the representation whose attributes best match those it has sampled from its current location. Suppose that Z_{com} is the number of these matching elements. BACON could then calculate the probability of getting that degree of match given that it was sampling its Z_{cur} attributes at random if it *were in fact in the place it remembered* (call this P[Z_{com}|Same, Z_{cur},Z_{rec}]) and also calculate the probability of getting that degree of match if it *were just in some random new place* (call this P[Z_{com}|Diff, Z_{cur}, Z_{rec}]). If the former number were much higher than the latter one, then it could be pretty confident that it was in the place it recalled, whereas if the reverse were true, it was probably remembering the wrong place. This ratio of probabilities is the so-called "Bayes factor," and its logarithm is the Bayesian weight of evidence (Kass & Rafter, 1995), which we here call B_{Rep}. Our model postulates that B_{Rep} is calculated by some extrahippocampal circuitry and used as a measure of its confidence that it is in fact where it thinks it is. The more positive B_{Rep}, the more certain it is that the active recalled representation is that of the actual current context, and the more negative, the more certain that it is not.

Thus

$$B_{Rep}(Z_{com}\,|Z_{cur},Z_{rec}) = \log_{10}\frac{P[Z_{com}|\text{Same},Z_{cur},Z_{rec}\,]}{P[Z_{com}|\text{Diff},Z_{cur},Z_{rec}\,]} \qquad (4)$$

where "*P*" is probability, "Same" means that the active contextual representation is that of the actual current context, "*Diff*" means that it is not, and Z_{cur} (for "$Z_{current}$") is the number of attributes of the current context that have been sampled; Z_{rec} (for "$Z_{recalled}$") is the number of attributes already associated with the currently active representation; and Z_{com} (for "Z_{common}") is the number of attributes that are the same in these two sets.

Calculations of B_{Rep} as well as of synaptic weight changes, and excitations in both hippocampal and amygdala neurons were based on calculations of the Expected Values of Z_{rec} and Z_{com} given BaconX's prior experience, assuming that attributes are sampled at random during a contextual visit.

2.2.2 | Rate of attribute sampling

As explained at the start of Section 3, at each visit to a context, BaconX, like BACON, samples the N_{attr} attributes of the context it is in in random order without replacement. We assume that the time needed to find a new, previously unsampled attribute increases as the number of attributes not yet sampled diminishes. Since extinction will develop during unreinforced periods between one sampling and the next, we need to specify the duration of those periods. Time in BaconX is measured in "intervals" of about half a second. We assume that the number of intervals needed to find a new attribute is given by $\kappa/([N_{Attr} - Z_{cur})^{\mu}]$ (the values of these, and all other parameters used for the simulations of this article are listed in Table 1).

TABLE 1 BACON and BaconX parameter definitions and values

Description	Name	Value
Basic network characteristics		
Numbers of EC′$_{in}$ and EC′$_{out}$ cells	N_{Ctx}	1,000
Numbers of DG′ and CA3′ cells	N_{Hipp}	10,000
Number of attributes per context	N_{Attr}	100
Number of EC′$_{in}$ neurons innervating each DG neuron	F	60
Number of winners of hippocampal KWTA calculations	K	60
B_{Rep} thresholds		
Negative B_{Rep} level sufficient for representation creation	B_{new}	-3
Minimum B_{Rep} level allowing addition of attributes to an existing hippocampal representation	B_{add}	15
Minimum B_{Rep} level for conditioning to occur	B_{cnd}[a]	3
B_{Rep} level at which maximal conditioning occurs	B_{mxcnd}	12
B_{Rep} threshold for expression of conditioned fear	B_f	0
B_{Rep} level of maximal expression of conditioned fear	B_{ff}	6
B_{Rep} threshold for between-session extinction	B_{xBtwn}	10
B_{Rep} level for maximum extinction consolidation	B_{xxBtwn}	13
B_{Rep} threshold for within-session extinction	B_{xWthn}	0
B_{Rep} level for maximal within-session extinction	B_{xxWthn}	13
Attribute sampling parameters		
Time to next sample = $\kappa/[(N_a\text{-}Z_{cur})^{\mu}]$		
Numerator	κ	1
Exponent	μ	3
Amygdala learning parameters		
Amygdala learning rate parameter α = $(k_{\alpha}\cdot A_u)^{\wedge}m_{\alpha}$ (see Figure 4, Box 1)	k_{α} m_{α}	0.9 5
Inhibitory learning parameter	β	5 E-6
Proportion of within-session inhibitory learning that gets consolidated	φ	0.05

[a]B_{cnd} was called "B_{old}" in Krasne et al. (2015).

2.2.3 | Updating

As will be further explained in Section 3, when there is sufficient confidence that the representation active in a context is the correct one, newly observed attributes can become associated with the representation ("updating"). In BACON it was assumed that such additions only occurred when BACON was extremely confident that it really was in the represented context (specifically, $B_{Rep} > B_{add}$, where B_{add} was quite high—see Table 1), and additions occurred on the fly as they were observed. However, as we will discuss, there is now some evidence that updating may occur when there is only moderate confidence in an active representation's validity. Therefore, to reduce chances of updating a representation in the wrong context, in BaconX, the addition of attributes newly observed during a session only occurs if $B_{Rep} > B_{add}$ at a session's end, at which time the maximal amount of information about the current context has been sampled.

3 | RESULTS

3.1 | Contextual representation creation, recall, and updating in BaconX's hippocampus

BaconX's hippocampus, like that of its predecessor, BACON, is designed, as originally proposed by Marr (1971), to form sparse multicellular hippocampal representations of cortical (specifically EC) patterns of activity that will be reactivated in their entirety when portions of the cortical patterns recur and will in turn cause reactivation of the full original cortical patterns. These hippocampal representations consist of "*K*" cells, which in rats and in BACON, are thought to be composed of about 0.5% of the total number of DG cells. As explained below, these hippocampal patterns of activity, in addition to determining EC recall activity, reach the amygdala where the conditioning of fear to them is the basis for contextual fear conditioning.

When BACON visits a context, it randomly samples the context's attributes without replacement and retains a working memory of those attributes for the rest of the session. BACON and BaconX carry out their computations every τ seconds (~0.5 s, the "computational interval"). In BaconX, the actual time between samples is important because extinction of context fear develops over time; the number of intervals needed to sample a new attribute is assumed to increase as more attributes are acquired (details in Section 2).

When a contextual representation is created, all the attributes in working memory become permanently (i.e., for as long as hippocampal memory lasts) associated with the representation. These associated attributes provide the basis for activating the same representation during future visits to the context and also for reactivating the full set of recalled attributes despite only a fraction of them having been actually observed during a current visit.

BACON's hippocampus can operate in three different modes:

1. Representation *creation* (i.e., encoding) during which the set of hippocampal cells that will represent a new context is selected and whatever attributes of the context have so far been observed become associated with the representation.
2. *Recall* of already established representations, during which whatever existing representation's associated attributes best match those that have currently been observed is activated.
3. *Updating*, during which newly discovered attributes of the currently active representation's context become associated with the representation. Control of which mode is operational depends the estimate of current representation validity. Degree of confidence in representation validity also modulates the extent to which paired activity of hippocampal representation cells and amygdala fear-producing cells can cause Hebbian potentiation and thereby successful conditioning. Thus, confidence in representation validity influences both the mode of operation of the hippocampus itself and the development of the synaptic potentiation within the amygdala thought to be the basis for associating emotions with hippocampal representations.

Contextual visits always begin with BACON's hippocampus operating in Recall mode (the default). When in this mode, whichever existing representation's associated attributes best match the set of attributes so far sampled during the visit becomes active. Non-hippocampal circuitry then makes a comparison between the active representation's associated attributes ($\mathbf{EC_{out}}$ in Figure 1) and currently observed ones ($\mathbf{EC_{in}}$), and this comparison provides a basis for determining how confident the virtual animal can be that the currently active (best-fitting) hippocampal representation is in fact that of the current context. The metric for this degree of confidence is the Bayesian Weight of Evidence (Kass & Rafter, 1995), which we denote "B_{Rep}," that the active representation is the correct one (B_{Rep} is defined in Section 2; see Krasne et al., 2015 for details). When the representation's validity is totally uncertain, $B_{Rep} = 0$; the greater the certainty that the representation is correct, the more positive B_{Rep} goes; and the greater the certainty of nonvalidity, the more negative.

B_{Rep} values, which vary as contextual attributes are sampled, control both the operational mode of the hippocampus (create, recall, or update) and aspects of amygdala function. The recall mode is the default. When B_{Rep} goes sufficiently negative (<"B_{new}") a new representation is encoded. When B_{Rep} becomes sufficiently positive that the currently active hippocampal representation is very likely to be the valid representation of the current context (>B_{add}), Updating (associating newly observed attributes with the representation) is allowed (Figure 1b). B_{Rep} values also control fear conditionability and expression of previously conditioned fear, as described below.

BACON and BaconX calculate B_{Rep}, from the number of attributes in common (Z_{com}) between the attributes associated with a representation and the attributes observed in the current context (Z_{cur}). The value of B_{Rep} obviously also depends on the number of attributes that are already known (i.e., associated with the representation—Z_{rec} for "recalled"), and the number that have been so far observed in the current context (Z_{cur}). Consistent with common sense, the Bayesian mathematics is such that the more one knows about a context, the less sampling of the current context is needed in order to decide whether it is or is not the hypothesized place, and conversely.

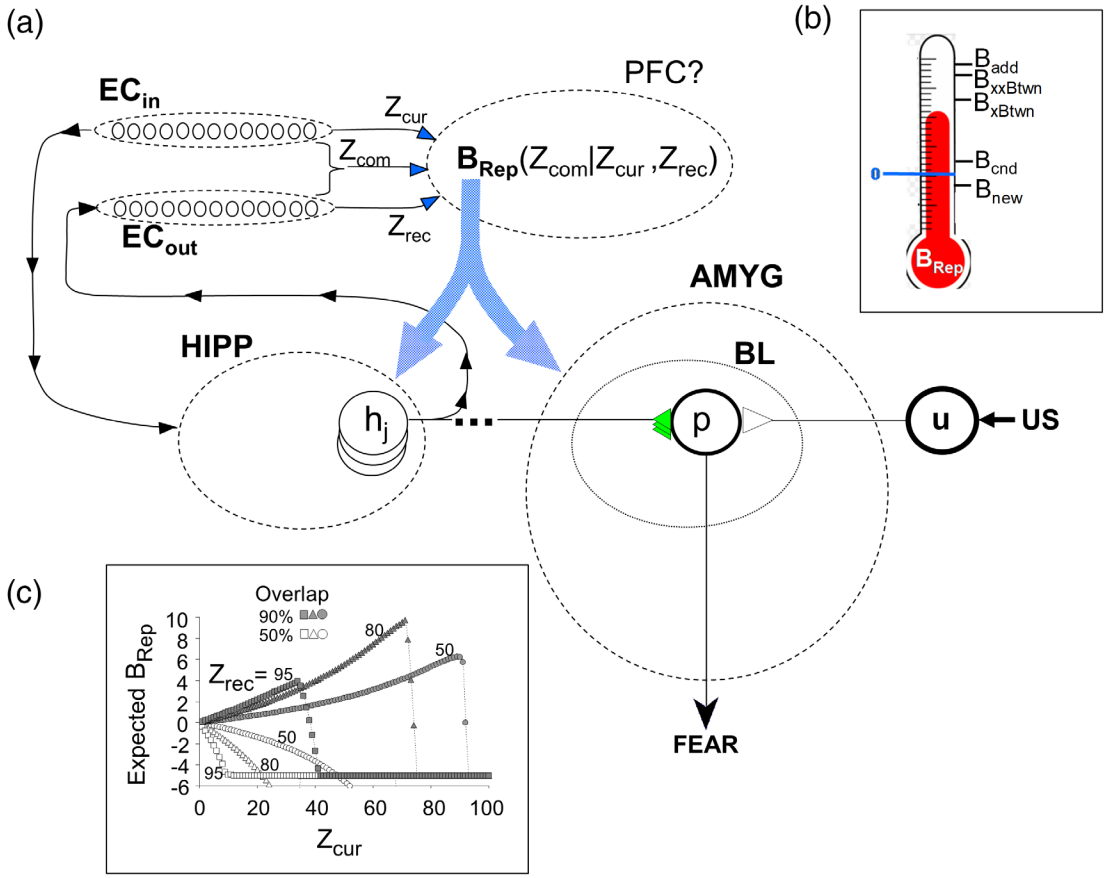

FIGURE 1 Summary of BACON model. (a) The organization of BACON. When placed into a context, $\mathbf{EC_{in}}$ cell firing tends to cause the firing of the *K* hippocampal representation cells (**h**) cells of the best matching previously encoded context. Based on the number of current attributes so far observed (Z_{cur}), the number associated with the representation (Z_{rec}), and the number in common between them (Z_{com}), BACON computes B_{Rep}, which indexes its degree of confidence that the active representation is in fact that of the current context. This value determines which of certain processes can or will occur as shown in Panel B. (b) A sufficiently negative B_{Rep} ($<B_{new}$) causes creation of a new representation, a very positive one ($>B_{add}$) allows newly observed attributes to become associated with an existing representation, a moderately positive one ($>B_{cnd}$) allows potentiation of synapses between active **h** cells and **p** cells in the amygdala should a US occur, and so forth. Note that the points marked on the thermometer are in the correct order but not quantitatively correct; actual values are shown in Table 1. (c) B_{Rep} values when BACON is placed in an unfamiliar context (B) having some similarity to a familiar one (A). If B is quite different from A, B_{Rep} soon becomes very negative, and a representation of context B gets created. However, if B is similar to, A then BACON can become quite convinced that it is actually in A before it ultimately "realizes" it is somewhere new and then creates a representation of the new context. The extent to which this happens depends on the degree of similarity between A and B and how much is known about A (i.e., its Z_{rec} value)

Moreover, as Bacon samples contextual attributes in a new context that is very similar to a known one, it can become very confident that it is in the known place (i.e., very high B_{Rep}) before it "realizes" that this is really somewhere different (i.e., before B_{Rep} goes very negative). These features are illustrated in Figure 1c. They are central to the model's ability to emulate many kinds of experimental observations.

3.2 | Design of BaconX's amygdala

3.2.1 | The amygdala circuit—general features

BaconX's amygdala circuit is shown in Figure 2 (Numbered boxes here and in later figures give quantitative relationships used for our simulations). As in BACON, context fear conditioning is due to Hebbian potentiation at the synapses (green) between hippocampal representation cells ($\mathbf{h_i}$) and principal (**p**) cells of a region analogous to the basolateral amygdala (BL). However, potentiation only occurs when there is sufficient confidence that the active representation is in fact that of the current context (specifically, if $B_{Rep} > B_{cnd}$; note: B_{cnd} was referred to as "B_{old}" in Krasne et al., 2015).

The modulation by B_{Rep} of potentiation at $\mathbf{h_i}$-**p** synapses is one of several instances of modulation by B_{Rep} in the model. We imagine that the value of B_{Rep}, which has various global effects on the operation of both hippocampus and the amygdala neurons, might be conveyed by some widely distributing neuromodulator such as nor-adrenaline, dopamine, 5-HT or ACh. However, the model is

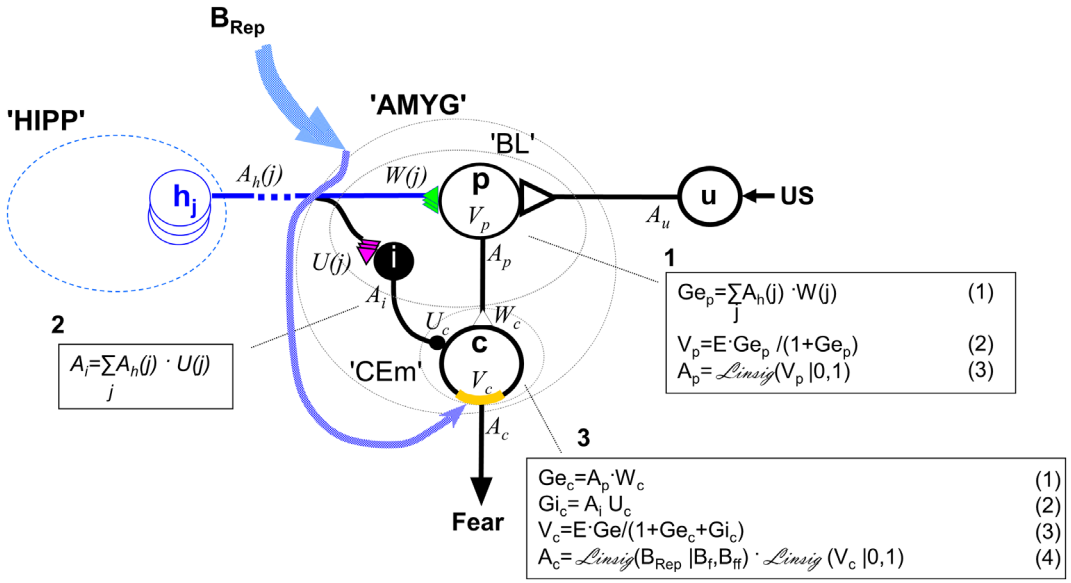

FIGURE 2 The amygdala circuit. Hippocampal representation cells innervate principal cells (**p**) and projecting inhibitory neurons (**i**) both of which innervate downstream output neurons (**c**). Synapses on **p** and **c** cells produce excitatory or inhibitory conductances (Ge_p = excitatory conductance/resting conductance of **p**; Gi_c = inhibitory conductance/resting conductance of **c**—see Section 2). These cause depolarizations (V_p and V_c) and firing rates (A_p and A_c) of the **p** and **c** cells, respectively. Numbered boxes indicate specific quantitative relationships (also in later figures)

agnostic as to the precise mechanism via which such modulations are done.

In accordance with the extensive evidence that extinction is due to learned suppression of persisting response tendencies, rather than to erasure of conditioning, extinction of fear in BaconX is mediated by active inhibition of fear causing neurons. Thus, the hippocampal **h** cells innervate inhibitory (**i**) neurons via Hebbian synapses (magenta in Figure 2), which become potentiated if fear is generated in the absence of any reinforcement (details below).

It might seem logical for the inhibition responsible for extinction to be targeted directly to the BL's **p** neurons; however, as explained below, this would lead to what we believe would be unsatisfactory characteristics for the generalization of extinction. Extinction can be made to generalize more plausibly if inhibition is applied downstream of the **p** cells. Therefore, we have introduced downstream **c** cells, somewhat analogous to CEm neurons, at which extinction-causing inhibition is applied. This makes our **i** neurons reminiscent of the Extinction neurons of Herry et al. (2008), which also project out of the region where conditioning itself is thought to occur. For reasons considered in the Discussion, we have not included prefrontal cortex (PFC) in the circuit responsible for inhibition.

Finally, again as in BACON, fear expression is modulated by B_{Rep}. Here, this is done by modulating the excitability of the **c** cells (i.e., the extent to which depolarization causes their firing), indicated by the yellow bar at the start of **c** cell axon in Figure 2. Its modulation by B_{Rep} is indicated by a blue modulatory pathway (as prescribed in Box 3, Equation (3)).

3.2.2 | Amygdala-based generalization—the need for c neurons and downstream inhibition

Generalization, both of fear itself, and of the inhibition that mediates extinction, can occur either because two similar contexts are confused with one another (discussed further below) or because of the properties of the amygdala itself (considered here). Since the pattern-separating properties of DG minimize overlap between the representations of even very similar contexts, one might suppose that amygdala-based generalization of context fear would be slight for all but the most similar contexts. However, the extent of depolarization of Bacon neurons that is produced by excitatory conductances saturates at the excitatory reversal potential E as excitatory conductance increases (see Section 2). Therefore, if synaptic weights are large, even input from a small percentage of potentiated synapses can produce almost maximal depolarizations. So, amygdala-based generalization can be great even when there is only modest representation overlap between two contexts. Figure $3A_2$ gives the depolarization of a **p** neuron as a percentage of its maximum possible value (the excitatory equilibrium potential) for various excitatory input conductances, Ge_p, and the solid curves of Figure $3A_3$ show the percentage of generalization of fear for each of these levels of Ge_p. For the largest excitatory conductances (green) generalization of **p** neuron activation (or depolarization) is 80% maximal when representation overlap is only about 2% and attribute overlap about 20%. The representations of all but the most extremely similar contexts overlap by less than 20%. For this range of overlaps, generalization is commensurate with overlap when Ge_p is small to moderate. However,

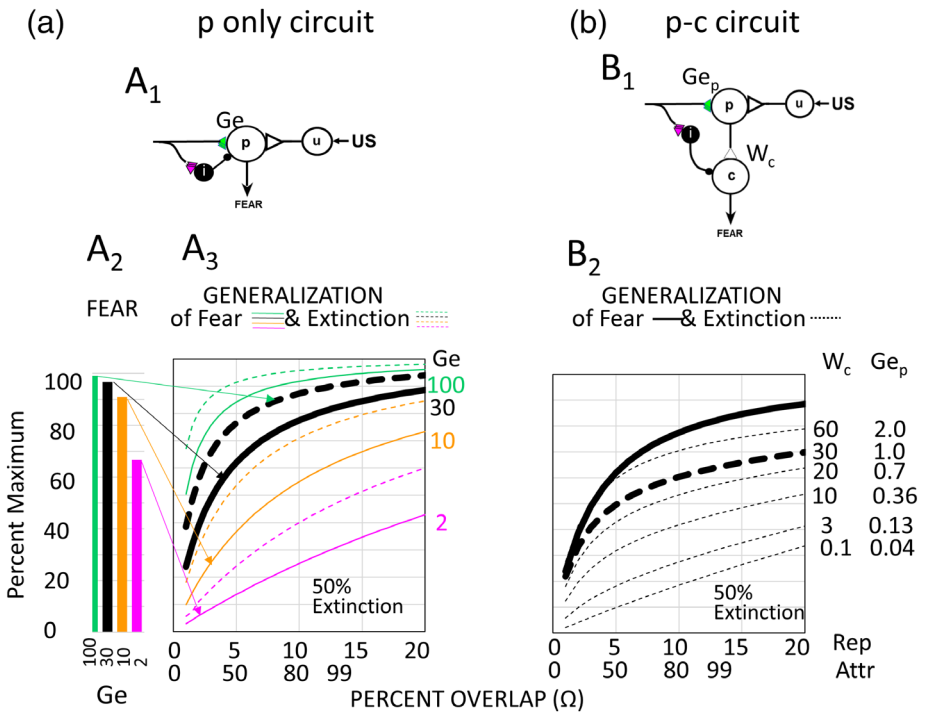

FIGURE 3 Generalization properties of amygdala circuit. We define generalization of fear (here taken as the depolarization of the output cells, V_c) to be fear in a generalization context as percentage of fear in the conditioning context. Generalization of extinction we define as degree of extinction (taken as proportion by which fear is reduced) in a generalization context as a percentage of reduction in the extinction context itself. (a) One-stage circuit. A1. Circuit. A2. Percent of maximum possible fear output (which occurs when the activation of the **p** neuron is 1.0) as function of excitatory conductance of **p** neuron (Ge_p) in the conditioning context. A3. Solid curves give the generalization of fear for the similarly colored fear levels in A2. Dashed curves give the generalization of extinction for the corresponding fear levels. Note that with the one-stage circuit extinction always generalizes more than fear itself. (b) Two-stage circuit. B1. Circuit. B2. Fear and extinction generalization for the two-stage circuit for various values of W_c with Ge_p set so that the fear generalization curve (bold solid line) is the same as that seen for the one-stage circuit when Ge_p = 30. For all simulations, W_c was taken as 30, giving the bold dashed extinction generalization curve. Note that with the two-stage circuit one can have a chosen fear generalization curve but with extinction generalizing *less* than fear to the extent determined by choice of parameters W_c and Ge_p

when conditioning is strong, there is a lot of generalization. And since there seems to us to be considerable generalization in many rodent context fear conditioning experiments, we will do many of our simulations under conditions that yield fairly considerable generalization of fear. Specifically we have chosen to set parameters so that when a single US that is 50% of the maximum possible occurs once in a context of whose validity BaconX is maximally confident (the rules that specify conditionability as a function of B_{Rep} are explained below), generalization is like that seen in the figure when Ge_p = 30 (the bold black curve). This results in about 70% generalization of conditioned fear when there is a 10% representation overlap, corresponding to an 80% attribute overlap.

It might seem logical for the inhibition responsible for extinction to be targeted to the BL's **p** neurons. However, if this were the case, then it would follow from the neuron model we are employing, whose properties are given by Equation (2), that extinction would always generalize more than conditioning itself. (In Figure $3A_3$ compare percentage generalization of extinction, which is plotted as dashed curves to fear generalization curves which are plotted as solid curves of the same color.) To us, this seems implausible. While it is well known that the extinction of cued fear is quite specific to the context in which the extinction occurred, there appears not to have been much study of the extent to which extinction of context fear generalizes. However, extreme generalization of context fear extinction would seem to be a dangerous strategy because it would often prevent fear when in fact fear might be highly adaptive. We therefore wished to design BaconX so that extinction would generalize less than conditioning itself. As shown by calculations based on Equation (2), which are graphed in Figure $3B_2$, this can be done by applying inhibition downstream at the level of the **c** rather than at the **p** cells. When this is done, extinction generalizes less, rather than more than does fear itself if parameter values are chosen appropriately. It should be noted that in the real amygdala, downstream inhibition does in fact seem to be employed to effect extinction (see Pare & Duvarci, 2012 for a review).

Figure $3B_2$ shows extinction generalization curves for a range of strengths of **p**-**c** synapses (W_c) given the fear generalization curve that we selected above. In all of them extinction generalizes less than fear, but there is not a good empirical basis for choosing between them. Therefore, we have considered what sort of generalization of context

fear seems as though it would be most adaptive. Danger can be intrinsic to the general nature of the current circumstances or particular to the specific situation of the moment. If one thinks there is a wolf in a forest because one was attacked there, but one discovers that one can now navigate that forest safely, it might be because wolves rarely attack and last time was just very bad luck, or it might be that the wolf that attacked has left the forest. In the one case, one should probably generalize one's extinction and in the other not. Unfortunately, one is not likely to know which sort of strategy is applicable. So, it seems likely that nature would have hedged its bets and chosen a degree of generalization of contextual fear extinction somewhere between the extremes. We have chosen for our simulations parameters that lead to the bold dashed curve in Figure $3B_2$.

3.2.3 | Learning rules

As said above, conditioning and extinction are due to Hebbian potentiation of **h** neuron synapses on BL **p** cells and **i** cells. The details of this are portrayed in Figure 4.

Potentiation of **h**-**p** synapses depends not only on joint presynaptic and postsynaptic activity, but also on the strength of the US and the extent to which BaconX has already learned to be afraid. Increments in the strength of the **h**-**p** synapse are proportional to a learning parameter α that increases exponentially as a function of US magnitude (Box 1, Equation (1)) and on modulation via an **R** (reinforcement) neuron whose activity depends on the extent to which the strength of fear that has already been learned has reached a level appropriate to the US magnitude (Box 1, Equations (2) and (3)); these dependencies lead to a negatively accelerated learning curve. Increments also depend on degree of confidence that the active contextual representation is valid (Box 1, Equation (3)) since there is no point in becoming afraid of a place if one does not know enough about it to identify it in the future. We imagine that such modulatory effects might be mediated by neuromodulators such as dopamine or norepinephrine (see Krasne et al., 2015).

Potentiation of **h**-**i** synapses within a session depend on **h** cell input and postsynaptic depolarization caused by input, if any, from the **p** cell, but also on modulatory input from **X** (extinction) neurons whose activity depends on both **p** cell and **c** cell activation (Box 2, Equation (1)); thus, these synapses become potentiated if **p** cells are

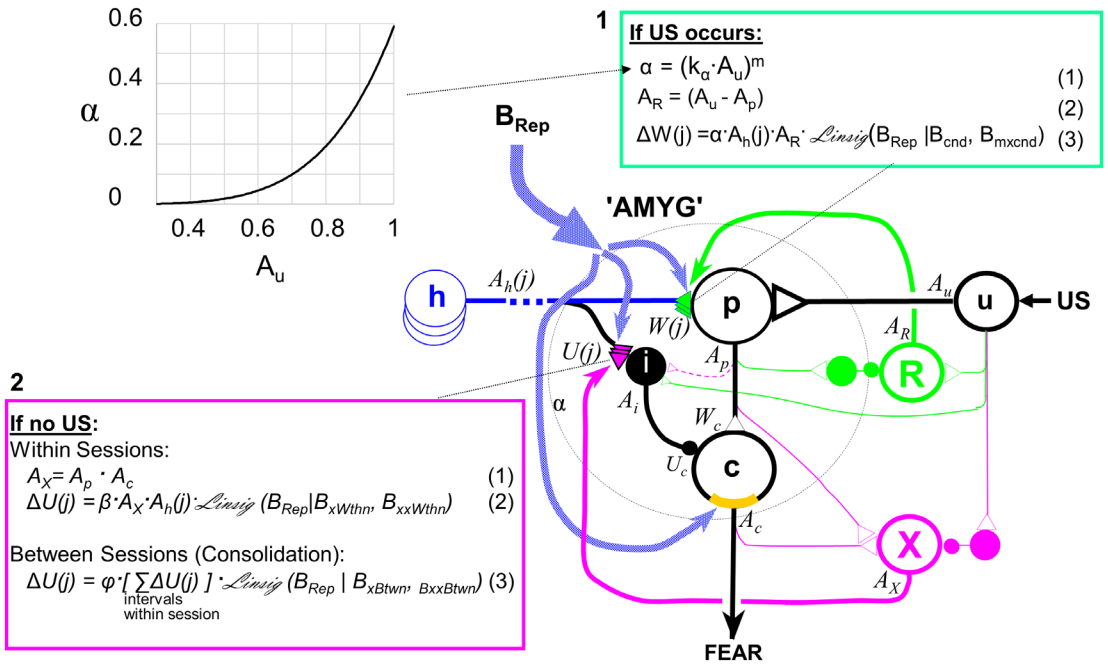

FIGURE 4 Learning rules. As described in Section 3, potentiation of **h**-**p** and **h**-**i** synapses depends on presynaptic activity plus postsynaptic depolarization as well as modulatory input from reinforcement **(R)** and extinction (**X**) neurons and B_{Rep}. Specifics are summarized in Boxes 1 and 2. (a) Release of transmitter from **c**-**X** synapses due to its activity is augmented by presynaptic input from **p** (an instance of presynaptic facilitation as in Castellucci and Kandel (1976), Li and Zhuo (1998), and MacDermott, Role, and Siegelbaum (1999)). (b) **p** depolarizes **i** dendrites so that $\mathbf{h}_j$-**i** LTP can be established when $\mathbf{h}_j$ and **p** are coactive; however, this dendritic depolarization does not cause firing of **i**. Additional considerations (not pictured) apply if fear is repeatedly conditioned and extinguished. In that case, eventually A_p would saturate at its maximum, and it would no longer be possible to reverse extinction by increasing **h**-**p** synapse strength. We therefore allow for reinforcements to decrease previously established inhibition: Within sessions $\Delta U(j) = -\gamma_{Wthn} \cdot U(j) Linsig\ (B_{Rep}|B_{xoWthn}, B_{xomxWthn})$ and between sessions $\Delta U(j) = -\gamma_{Btwn} U(j) \cdot Linsig\ (B_{Rep}|B_{xoBtwn}, B_{xomxBtwn})$ and between sessions. However, for the simulations of this article, we set to zero the parameters (γ_{Wthin} and γ_{Btwn}) that specify the degree of erasure of inhibition that results from reinforcements. As considered in Section 4, we think it is quite possible that when reinforcement and nonreinforced sessions occur repeatedly, the two situations may end up becoming considered different contexts and in that event, erasure of inhibition would not be needed. There is in fact some evidence that such erasure does not occur (Rescorla, 2001)

active, but not if enough potentiation has already developed so that **c** cells are fully suppressed. Inhibition of **X** cells by the US prevents reinforced responses from contributing to extinction.

Increments in **h-i** potentiation are also dependent on confidence as to representation validity (Box 2, Equation (2)).

Within versus between-session extinction of context fear

It is adaptive for fear being expressed within a session in a context to decline if there appears to be no current threat. Thus, the synapses of active hippocampal cells on **i** neurons become subject to potentiation whenever **p** cells are activated and fear is expressed but no US occurs, as prescribed by the above rules. However, whereas it is adaptive for fear to decline within a given session if there appears to be no current threat, it is important to be especially sure that one really is in the place that one's hippocampus "thinks" one is before permanently extinguishing fear conditioned to whatever contextual representation is active. It could be disastrous if time in a safe place that had been confused with a dangerous one was to cause permanent fear extinction to the dangerous place. Thus, extinction should only become permanent if B_{Rep} is quite high. For this reason, we distinguish within and between-session extinction. For the latter, we specify a rather stringent criterion level B_{xBtwn} that B_{Rep} must reach before any consolidation of newly developed extinction can occur, and a still more stringent criterion (B_{xxBtwn}) for a maximal degree of consolidation. If the amount of potentiation that has developed within a session at synapse $\mathbf{h_j}$-**i** is $\Delta U(j)$, then the increment in permanent inhibition after consolidation is a fraction φ of this, adjusted by degree of confidence in correctness of the currently active representation (Box 2, Equation (3)). Thus

$$\Delta U(j)_{\text{consolidted}} = \varphi \cdot Linsig\left(B_{Rep} \mid B_{xBtwn}, B_{xxBtwn}\right) \cdot \Delta U(j)_{at\ session's\ end} \quad (5)$$

3.3 | Behavior of the model

While the properties of the amygdala circuitry discussed above play a major role in BaconX's behavior, the behavior generated by the full model is also determined by which representations the hippocampus creates and activates over the course of its stay in a context as well as on BaconX's degree of confidence in the validity of its currently active representation. (i.e., B_{Rep} values).

3.3.1 | Conditions for extinction

Within versus between-session extinction: Extinction consolidates only if BaconX is confident as to its whereabouts

The effects of nonreinforced exposures to a fear-conditioned context are illustrated in Figure 5. Consider first a case (a_3) in which conditioning was at a PSI that allowed substantial but not extensive information about the context to be learned (300 intervals), and extinction sessions were fairly long (1,500 intervals). Within-session extinction then occurred during every session, and since B_{Rep} reached B_{xxBtwn} by the end of each session, the degree of extinction increased substantially from session to session. In case a_2, extinction sessions were shorter, and because less of the current situation had been

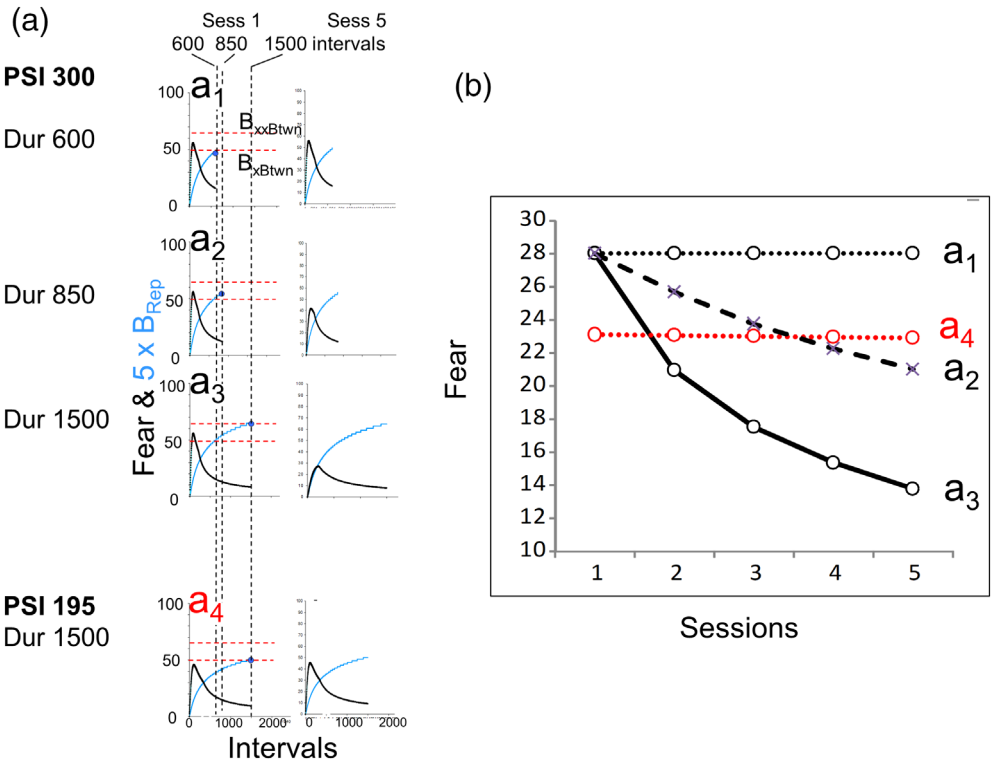

FIGURE 5 Between-session extinction fails if PSI or extinction session duration are too short. (a) First and last extinction sessions under various combinations of PSI and session duration ("Dur") (a_1–a_4). The B_{Rep} value at the end of the first session are marked (blue dots) and the B_{Rep} levels required for minimal and maximal between-session extinction indicated (red dashes). (b) Peak fear across sessions for the conditions of A. Conditions a_1 and a_4 resulted in no between-session extinction because B_{Rep} at the end of the session was too low. Condition a_2 resulted in slower between-session extinction because B_{Rep} was only slightly above B_{xBtwn} at the end of the session

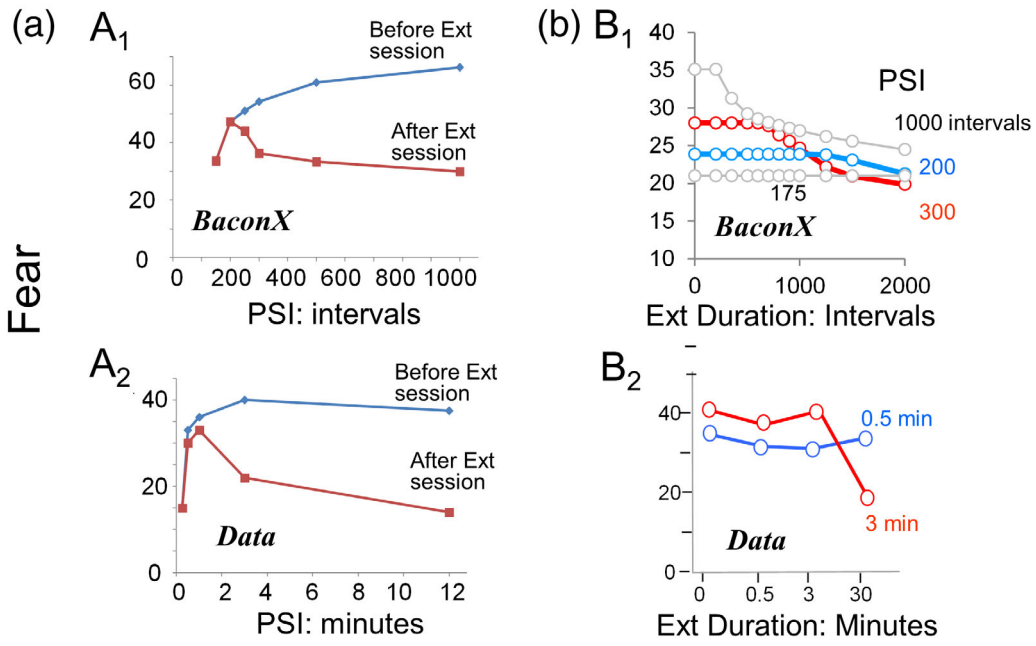

FIGURE 6 Between-session extinction as a function of PSI and extinction session duration as predicted by BaconX and as observed in Zinn et al. (2020). In each graph, fear was averaged across the test session. (a) Fear before and after an extinction session as a function of the PSI of conditioning. A_1. BaconX. (A_2) Data (Zinn et al.). (b) Fear after a single extinction session as a function of the duration of the session for conditioning at a variety of PSIs. A_1. BaconX. A_2. Data (Zinn et al.)

observed by the end of the session, BaconX was less confident that it really was in the situation where it had gotten shocked; therefore, though B_{Rep} still exceeded B_{xBtwn}, it did not reach B_{xxBtwn}, so extinction did progress over sessions, but less rapidly than in a_3. When B_{Rep} did not reach Bx_{Btwn} by the end of the session, either because PSI had been too short or because the extinction session itself was too short, no between-session extinction occurred, as in cases a_1 and a_4.

It should also be noted that, as we will discuss below, new information about a context can be acquired during extinction sessions, and this can increase the rate of extinction. However, it did not occur in the simulations of Figure 5. Cases where it did happen are discussed in the next section.

Figure 6 shows simulations of average fear as a function of PSI (Figure $6A_1$) and extinction session length (Figure $6B_1$) on a session following a single extinction session. Panels A_2 and B_2 show the corresponding relationships as actually observed in a recent study on mice (Zinn et al., 2020); the predicted and observed relationship seem to us remarkably similar.

Contextual knowledge obtained even after conditioning ("updating") promotes subsequent extinction

It is well known that increasing context exposure prior to conditioning (either by increasing the PSI in the conditioning session itself or by introducing a pre-exposure session) increases the amount of conditioning that occurs (Fanselow, 1990). In BACON and BaconX this occurs because the more exposure to the context prior to conditioning, the larger Z_{rec}, and hence the greater B_{Rep} at the time of conditioning (Krasne et al., 2015). Moreover, even after conditioning has occurred, further attributes can become associated with a representation if B_{Rep} exceeds B_{add}. This increases Z_{rec}, and hence B_{Rep}, during subsequent recall sessions, which enhances expression of previously conditioned fear (Figure 2, Box 3, Equation (4)). However, as discussed above it also enhances both within and between-session extinction (Figure 4, Box 2, Equations (2) and (3)). It enhances within session extinction because extent of within session extinction depends on both the strength of the fear response and on B_{Rep} per se, and it enhances between-session extinction both because it increases within session extinction and because consolidation of within session extinction depends on B_{Rep} per se, as discussed in the section on amygdala circuit design.

The effect of postconditioning updating is illustrated in Figure 7. As seen in Panel A, after conditioning at a PSI of 300 intervals, B_{Rep} did not quite reach B_{xBtwn} by the end of the extinction sessions, and therefore no between-session extinction occurred no matter how many sessions were given. However, in Panel B the initial extinction was made much longer. This allowed considerable updating to occur (Z_{rec} increased from 50 to 90). Thus, on later sessions, B_{Rep} was much higher and, even though the remainder of the sessions were short, between-session extinction was substantial.

Updating may not require great confidence in representation validity and may occur inappropriately

When designing BACON it was assumed that B_{add} should be set to a very high level so that attributes of a context similar to but not the same as a feared one would virtually never become associated with

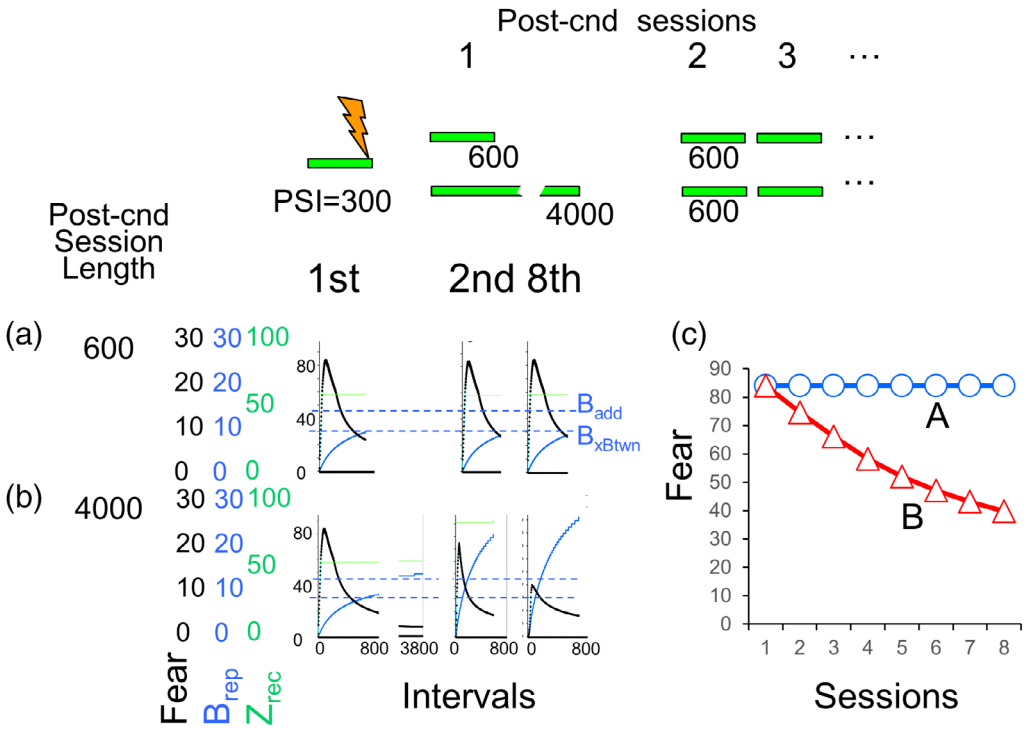

FIGURE 7 Enhancement of between-session extinction by postconditioning updating. (a) At the PSI and extinction session lengths used here, there was no between-session extinction. (b) However, a single long postconditioning session allowed updating that made between-session extinction possible. (c) Peak fear as a function of postconditioning sessions

the feared context's representation (Krasne et al., 2015). In most of the simulations done here, B_{add} was accordingly made very high and was above B_{xBtwn} (see Table 1). However, perhaps, B_{add} should not be set so high. It could be argued that updating one's information about a context has advantages (with becoming able to extinguish fear that would otherwise be inextinguishable being among them) that make it worth the risk of adding some incorrect information about a context to its representation.

Figure 8 illustrates a case where with conditioning at a PSI of 150 (allowing conditioning but only limited sampling of the context) and with B_{add} set to our standard value of 15, even a long re-exposure to the conditioning context did not restore between-session extinguishability at the extinction session length used. However, with B_{add} lowered to 10, updating became possible (Z_{rec}, not shown, went from 45 to 90), and this allowed between-session extinction to occur (Figure 8a). In this case, updating caused a large increase in expressed fear prior to extensive extinction, because fear expression is an increasing function of B_{Rep}. However, the extent to which such an increase will occur depends on the parameter (B_{ff}) that specifies the B_{Rep} at which maximal fear expression will occur, and if this is reduced, as was done in Figure 8b, only a very slight increase of fear occurs as the result of updating. There is some indication that this lower value may be more reflective of mouse behavior than those used in Figure 8a (Zinn et al., 2020).

However, there are dangers in lowering B_{add}. This is illustrated in Figure 9 where B_{add} was set to 2, a level that statisticians who deal with Bayesian weights of evidence consider a degree of certainty at

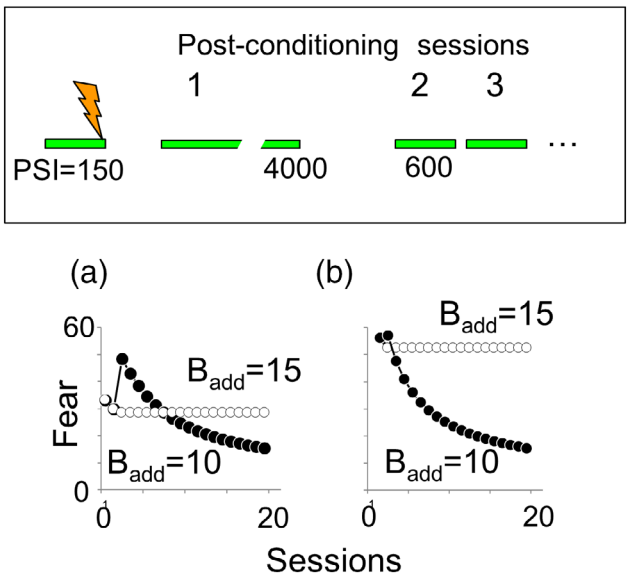

FIGURE 8 Effect of lowering B_{add} on effects of postconditioning exposure to the conditioning context. (a) When PSI was very short, even a long postconditioning exposure (4,000 intervals) did not make between session extinction possible when the B_{Rep} threshold for updating was very high (B_{add} = 15). However, lowering B_{add} somewhat (to 10) allowed updating that made between-session extinction possible. (b) The same experiment as in (a) except that the B_{Rep} threshold for maximal fear expression (B_{ff}) was lowered from 6 to 2

the low end of "substantial" (Kass & Rafter, 1995). After being conditioned in context A at a short PSI, Bacon was given time in a safe context B that was similar to A. The result of this exposure was that BaconX became more afraid of the safe context than the dangerous one, as seen in panel II. This happened because the safe context B was sufficiently similar to A so that, given the nonstringent B_{add}, enough of its attributes became associated with Rep A during the exposure to so that there were actually more context B than context A attributes associated with it after the updating (Figure 8I). As seen in panel III, when tested in context B, BaconX "thought" it was in A and expressed fear accordingly, whereas when place back in context A, there was a sufficient mismatch between the currently observed attributes of context A and the now mixed (majority B) set of recalled attributes so that a new (second) representation of context A was created. However, this new representation had only generalized fear associated with it, so less fear was now expressed in context A than in context B. Surprisingly, a very similar result has actually been seen in mice that were conditioned at low PSIs, extinguished, and then given time in a context similar to the one in which they had been conditioned (Zinn et al., 2020)!

3.3.2 | Generalization of fear and extinction

Extinction of context fear generalizes less than the fear itself; thus extinguished context fear returns in a different but similar context

As discussed above in Design of BaconX's amygdala, BaconX was designed to show some return of extinguished context fear when placed in a context different from, but similar to, the one in which fear was both conditioned and extinguished. The consequences of this design feature are shown in Figure 10.

Based on the properties of the amygdala circuit alone in Figure 10a, it can be seen that, as might be expected, extinction generalizes more when it is more complete. Also, as comparison of A_1 and A_2 shows, it generalizes less when the fear that was extinguished was greater.

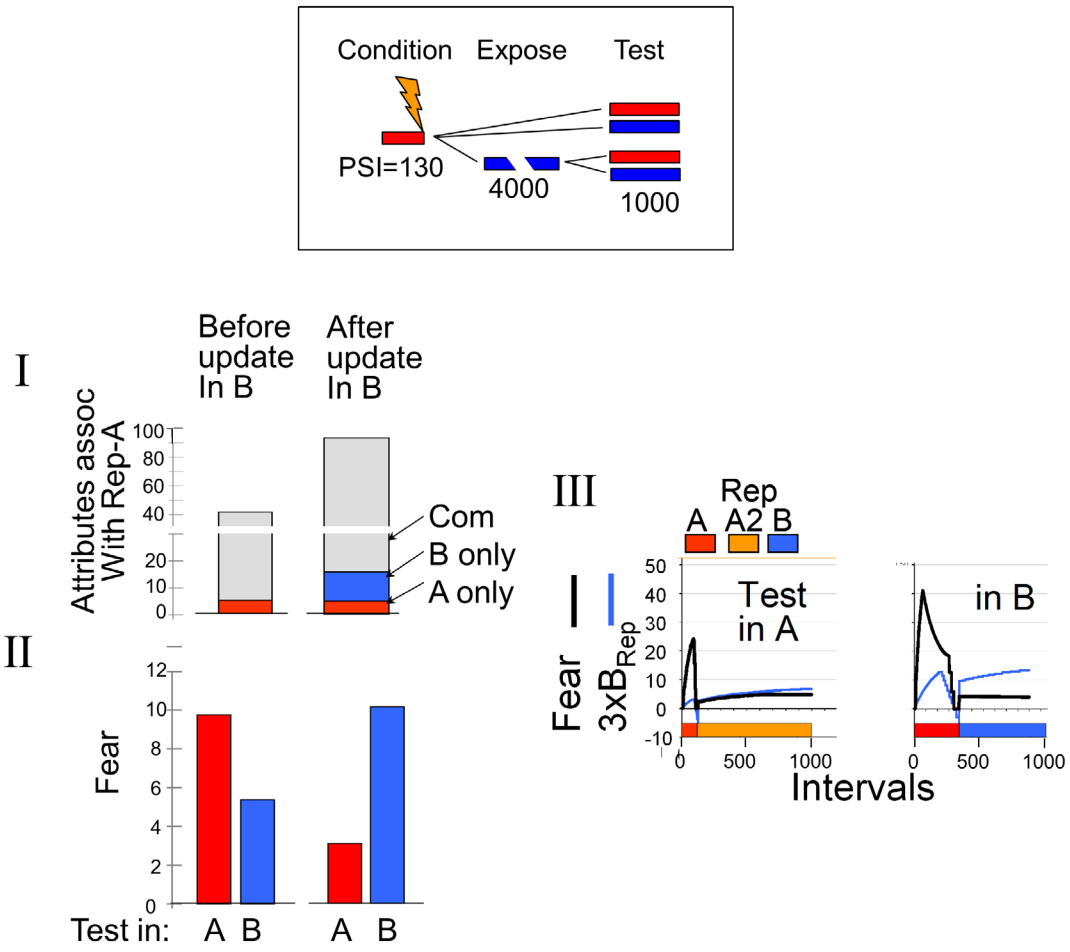

FIGURE 9 Reducing the B_{Rep} threshold for updating can cause serious recall distortions. (I) When B_{add} = 2, postconditioning exposure to context B, which was fairly, but not extremely, similar (87%) to the conditioning context (context A), caused a number of attributes of context B (blue segment of bar after update) to become associated with Rep A. (II) When tested without this postconditioning exposure there was more fear in A than B. However, after it BaconX became more afraid of context B than context A. (III) As explained in Section 3, tests in A caused a new, unfeared representation of A to be created and activated, whereas during a test in B Rep A was initially activated and caused fear, though eventually a representation of B was created

With the full model operating, Figure 10b shows the effects of extinguishing fear in the context of conditioning at levels of fear produced by three different US intensities. In all cases, there is considerable generalization of fear to context B (80% similar to A). However, as with cued fear, extinction is more context-dependent. In our simulations extinguished context fear generally returns when there is a shift to a different but similar context; however, this effect is only conspicuous when the conditioned context fear was strong. We view such returns of fear in a different context as analogous to the renewal of cued fear when an animal moves to a context different from that in which extinction was executed.

Extinction of primary fear generalizes more than extinction of generalized fear

As just said, BaconX's amygdala was designed to show some return of context fear if fear is extinguished in the conditioning context and then tested in a different similar one. It is also of considerable translational interest to ask how extinction carried out in a context similar to the conditioning context generalizes, because this is essentially the situation during exposure therapy. Figure 11 compares the degree of generalization of extinction executed in the conditioning context to that of extinction executed in a similar but different context, based on the properties of the amygdala alone. Context A is the conditioning context and contexts B and C are equi-similar "surrogate" contexts (80% attribute and 10% representation overlap) that we take as analogs of a virtual reality therapy context and a random real-world context, respectively. Panel I compares the effect of 90% extinction (V_c reduced by 90%) in A versus B. Extinction in the conditioning context itself generalizes much more to surrogate contexts than the reverse, and extinction generalizes much better from the conditioning context itself to a surrogate context than from one surrogate context to another (as in generalization from the clinic to the real world). This happens because during extinction conditioned inhibition of **c** neurons grows only until it can nullify their excitation. Since

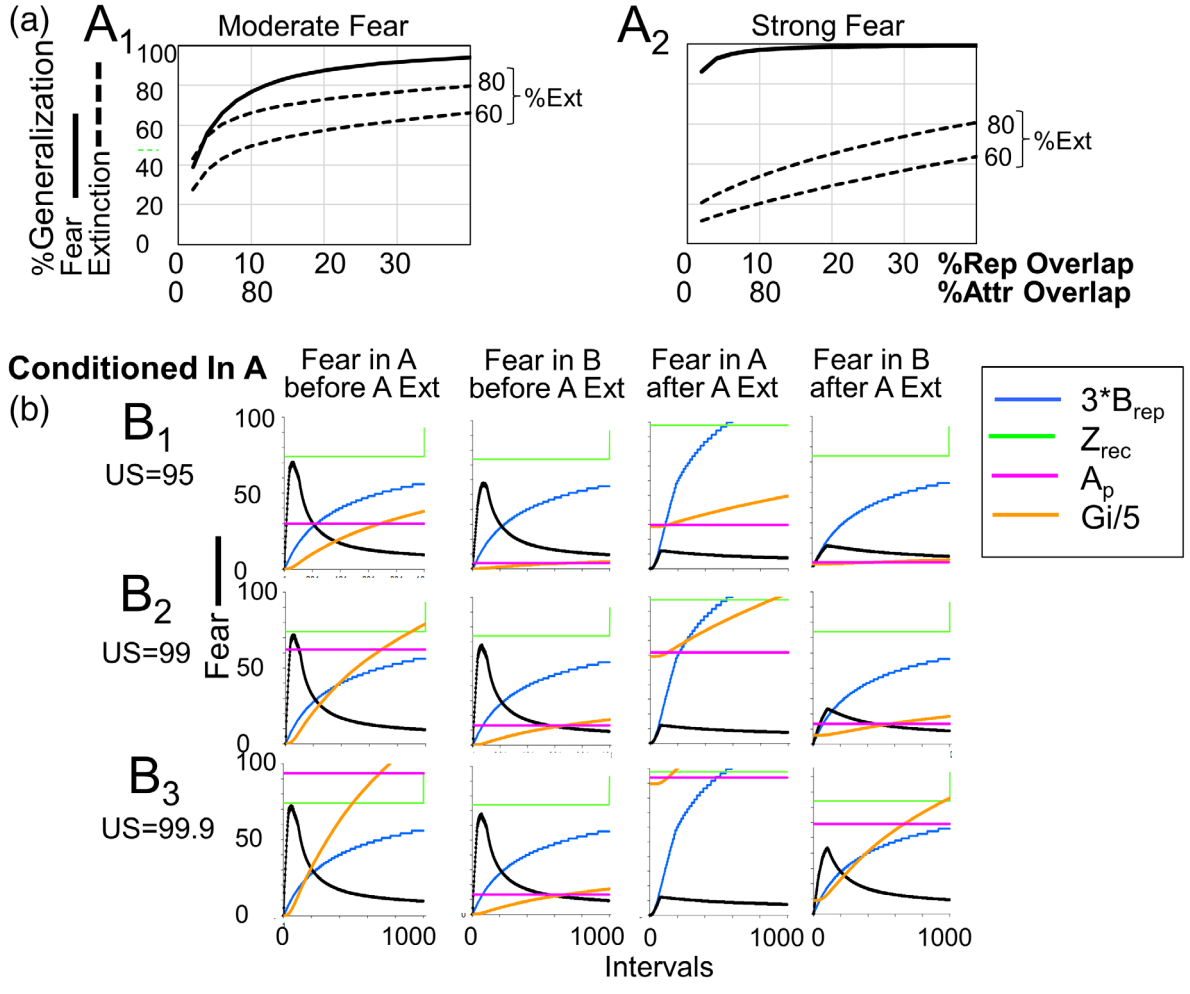

FIGURE 10 Generalization of extinction from the conditioning to other contexts. (a) Based on amygdala properties alone: Generalization of fear and of extinction as a function of attribute and representation overlap at two levels of fear. A_1 Moderate fear (Ge_p in conditioning context = 2). A_2. Strong fear (Ge_p in conditioning context = 20). %Generalization is computed as described in the caption of Figure 3. (b) Entire model operating: Generalization of extinguished fear after 20 extinction sessions. Attribute overlap 80% and three strengths of shock. Fear is the percentage of maximum possible fear. Context fear returns when BaconX is moved to a new context after extinction in a similar one; the greater the fear, the more renewal

excitation of **c** is greater in the context of conditioning than in a context to which fear has generalized, inhibition is greater when extinction is executed in the conditioning context than in one to which fear has generalized. Thus, there is more inhibition to generalize to still other contexts. Panels II shows the results of similar calculations for a range of conditions. They confirm that generalization of extinction from the conditioning context is worse, the greater the excitation of the extinguished fear and that extinction generalizes better from the conditioned context than from a surrogate context.

When the operation of the full model is considered, some additional phenomenology arises due to variations in which representations are activated during extinction. This is explored in the section below on approaches to exposure therapy.

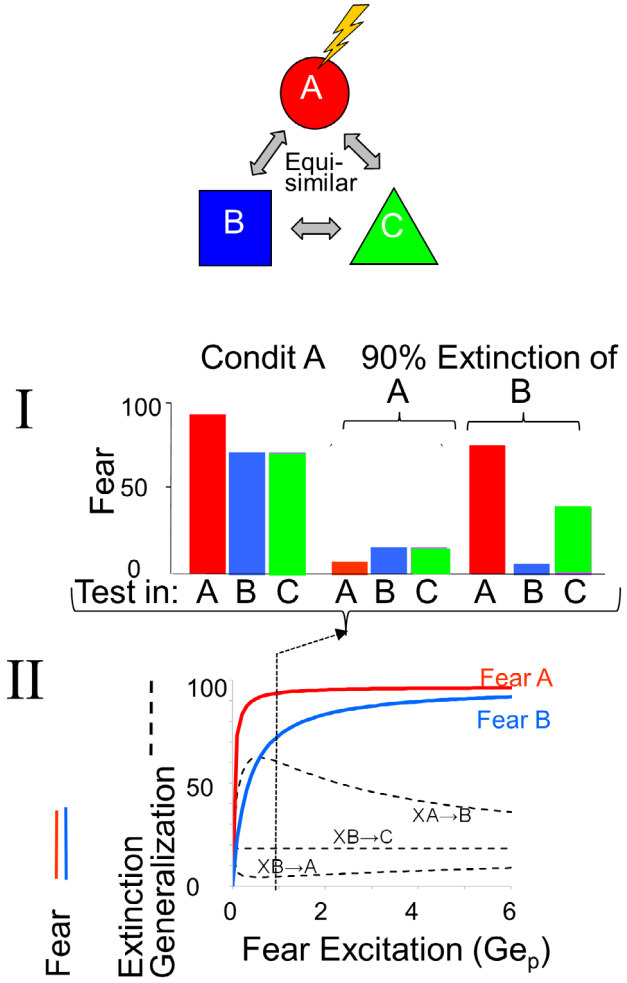

FIGURE 11 Comparison of extinction executed within versus outside the conditioning context as predicted from amygdala properties. "Fear" here is taken as depolarization of c cells (V_c) divided by their maximal possible depolarization. Contextual representations overlapped by 10%. (I) Generalization of fear extinction that was done in either the conditioning or a surrogate context; excitatory conductance of **p** cells in the conditioning context (Ge_p) was 1.0, and extinction reduced fear by 90%. As shown by the left, set of bars, prior to extinction there was considerable generalization of fear. The middle group of bars shows that extinction executed in the conditioning context generalized well (though with a slight return of fear). The right group of bars shows that extinction executed in B did not generalize well to C and generalized very poorly back to A. (II) Fear Generalization of extinction under the conditions of I for a range of levels of conditioning indexed by Ge_p in the conditioning context. The fear levels in I correspond to the values in II when Ge_p = 1.0

3.3.3 | Discrimination training

When BaconX is subjected to discrimination training in which fear is initially established in one context (A) and then multiple alternating sessions are given in A and a safe context B, there are a number of factors that can affect the development of correct differential responding to the two contexts:

1. If the contexts are similar and initial PSIs are short, a representation of B may never form. In that event, fear will become conditioned to the representation of A established during the initial conditioning session and inhibition will become conditioned to the same representation when in context B. Because the features observed during sessions in B will match learned features of context A less well than during sessions in A itself, there will be less confidence in the validity of Rep A when in B than when in A (i.e., lower B_{Rep}), and this will result in weaker fear expression in B than A; hence, there will be some differentiation, but not much.
2. Depending on initial PSI, session lengths, and context similarity, updating in A may or may not occur during sessions after the first in A, and Rep B may or may not be created. When and if Rep B does get created, there will be differential reinforcement when in A and extinction when in B, and this will greatly enhance the differentiation of fear of A versus B.
3. Long discrimination training sessions will increase the chances of Rep B getting created and will also increase the extent to which generalized fear in B gets extinguished.

Figure 12 illustrates these factors in action. In panel I, PSI during the initial conditioning session in A was context too short to allow B_{Rep} to go below B_{new} when in context B and too short to allow updating in context A that might have remedied this. Hence, Rep A was always active, but there is some differentiation of response as explained in Mechanism 1 of the previous paragraph. In panel II, the initial PSI was long enough to allow updating in A on Session 2. This in turn permitted Rep B to form during Session 3. Thus, differential conditioning of fear excitation to Rep A and inhibition to Rep B became possible as in Mechanism 2 above. In panel III a longer PSI allowed Rep B to form on Session 3 despite sessions being rather short. However, because the sessions were short, B_{Rep} at the end of sessions in context B was only barely above B_{xBtwn} (see criterion B_{Rep} levels at the right of the graphs). Therefore, only a small portion of the inhibition that developed during each such session got consolidated, so differentiation developed less rapidly than in II.

3.3.4 | Approaches to exposure therapy suggested by BaconX

A possible approach to mitigating the return of extinguished fear in different but similar contexts

A major problem with the use of exposure therapy to extinguish fear of situations in which traumatic events have occurred is that extinction which develops in the clinic may not generalize well to situations outside the clinic, as discussed above (Figure 11). This is shown for the full model operating in Figure 13 in which context A is the context of conditioning and contexts B and C may be thought of as representing a virtual reality therapy context and a similar real-world context, respectively. As will be explained in further detail below, if extinction occurred in the context of conditioning itself (context A, Row I), it generalizes quite well when tested in context C (only very slight fear resurgence). However, if extinction occurred in context B (Row II), fear substantially returns when BaconX is moved to context C. The superior generalizability of extinction carried out in the conditioning context as opposed to a place to which fear has generalized suggests a possible strategy for improving the generalizability of extinction done in a context similar to A. If context B is unfamiliar, it will initially activate the existing representation of A, itself, and it will take some time before a representation of B is created. Therefore, if extinction sessions were kept short, one could prevent Rep B from being created and perhaps extinction would proceed almost as if BaconX were in the conditioning context itself. This strategy was employed in Row III of the figure, in which it is seen that short extinction sessions executed in B resulted in extinction that generalized well to a different context, C. A detailed explanation of these results follows:

When BaconX was extinguished in context A and then moved to a similar, unfamiliar, context C (Row I), it for a moment thought it was in

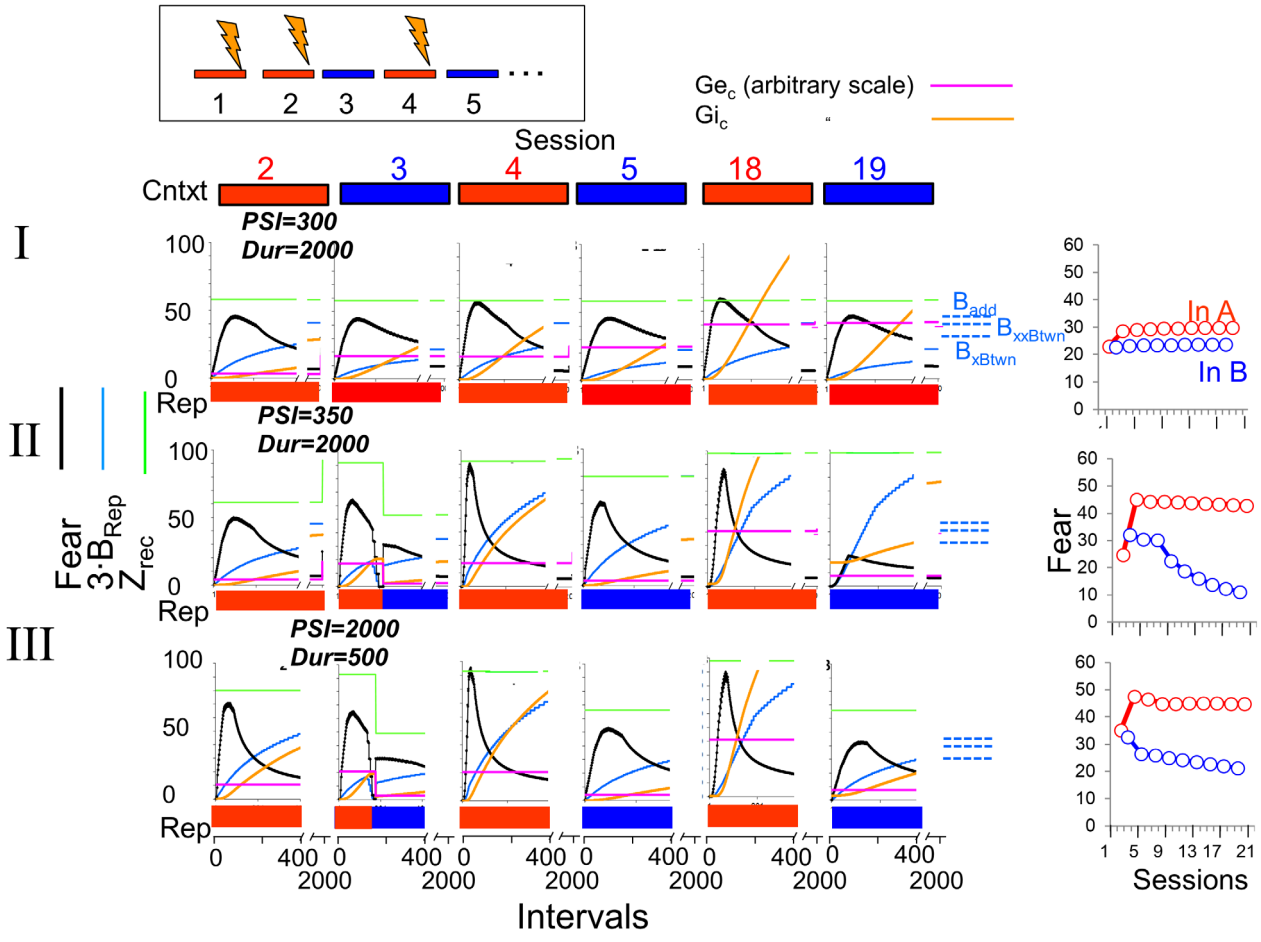

FIGURE 12 Discrimination training. BaconX was given two conditioning sessions in context A and then given alternate nonreinforced sessions in context B and reinforced sessions in context A. B_{add}, B_{xBtwn}, and B_{xxBtwn} are indicated at the right of each row of sessions. (I) PSI is so short that B_{Rep} does not reach either B_{xBtwn} or B_{add} by the end of any session, so no representation of context B develops. There is a little discrimination between A and B due to B_{Rep} being greater when BaconX is in the context the representation of which is active (A)than when it is in a different context (B). (II) A slightly longer PSI allows B_{Rep} to reach B_{add} at the end of Session 2, which allows updating and in turn formation of Rep B on Session 3. Once this happens inhibition becomes preferentially associated with context B, and discrimination markedly improves. (III) With a longer PSI but a shorter session length Rep B still gets formed on Session 3, but the sessions are so short that B_{Rep} at the end of sessions in context B just barely exceed B_{xBtwn}. Hence, only a little inhibition gets consolidated on each session in context B, and development of discrimination is somewhat slower

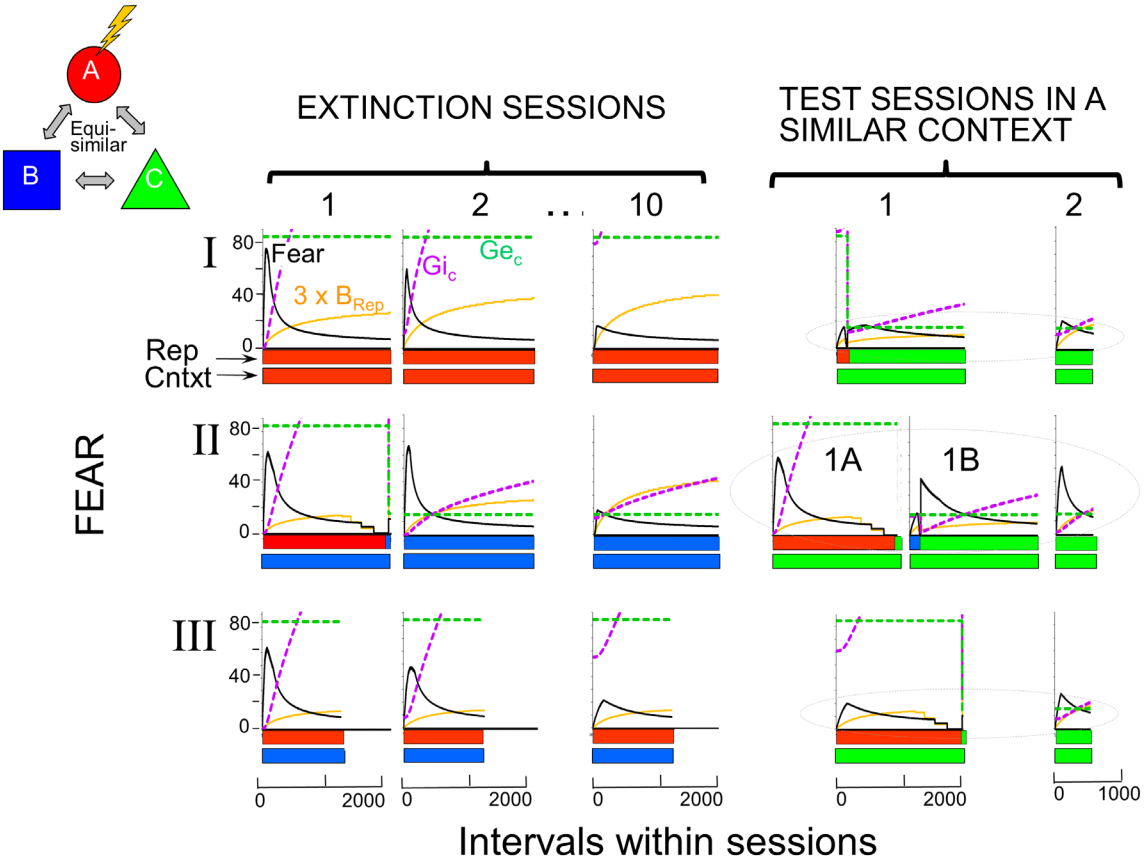

FIGURE 13 Improving generalization of extinction by preventing creation of an "exposure therapy" context representation. BaconX is conditioned in context A, extinguished in context B (the fictive therapy context) with a 43% maximal US, and tested in context C (a fictive real-world context). There were 10 extinction sessions, and pairwise context attribute overlap between contexts was 92%. Excitatory and inhibitory conductances (Ge_c and Gi_c) have arbitrary but consistent scales. (I) Extinction was in the conditioning context itself. A representation of the similar test context was created soon after the first exposure to it. Note that strong inhibition developed during extinction, and that extinction generalized quite well to the test context. (II) Extinction was in the nominal "therapy" context (B). A representation of B was created near the end of the initial session there and thereafter was always activated during the fictive exposure therapy sessions. Much less inhibition developed during extinction, and it became very slight during tests in a different context. So there was substantial return of fear during tests in Context C (the analog of the "real-world," nontherapy context. (III) Short "exposure therapy" sessions. Sessions were too short for a representation of the therapy context to form, so Rep A was always activated during extinction; consequently, extinction has generalization properties similar to those in case II, and extinction generalized well to context C

A but then formed a representation of context C. While Rep A was active, it drove both the substantial excitation of and roughly matching inhibition of **c** that had previously developed, and little fear was expressed. Once Rep C got created, the **c** neurons received only that portion of excitation and inhibition that was driven by those cells that were common to representations A and C. There was slightly better generalization of excitation than of inhibition due to the way we designed BaconX's amygdala, but excitation and inhibition remained roughly matched, and there was very little return of fear in context C. At the second test session Rep C was activated at the outset, and again excitation and inhibition were fairly well matched, so the response remained fairly well extinguished. Note that in this simulation context C was unfamiliar. If BaconX had previously formed a representation of C, then it would have behaved in Test Session 1 as it did in Test Session 2.

When extinction was executed in context B rather than context A (Row II), the inhibition that became conditioned to Rep B was much less than that which had developed when extinction was carried out in context A. This is because the amount of inhibition needed to extinguish generalized fear was much less than that needed to extinguish primary fear. The situation during the first exposure to context C was somewhat more complicated than that discussed for Row I, because in simulation of Row II, representations for both contexts A and B already existed when BaconX was first introduced into C. Either A or B could have been activated with about equal probability, so we must consider each possibility separately. If Rep A got activated (Row II, 1A), then excitation of fear was strong but inhibition of fear was generalized from the modest inhibition conditioned to Rep B, so substantial fear was expressed. At the very end of the session, Rep C got created. By this time, a substantial amount of within-session inhibition had become associated with Rep A. This plus generalized between-session inhibition from B kept fear low. If on the first session in C Rep B got activated initially, fear would have been as it was at the end of

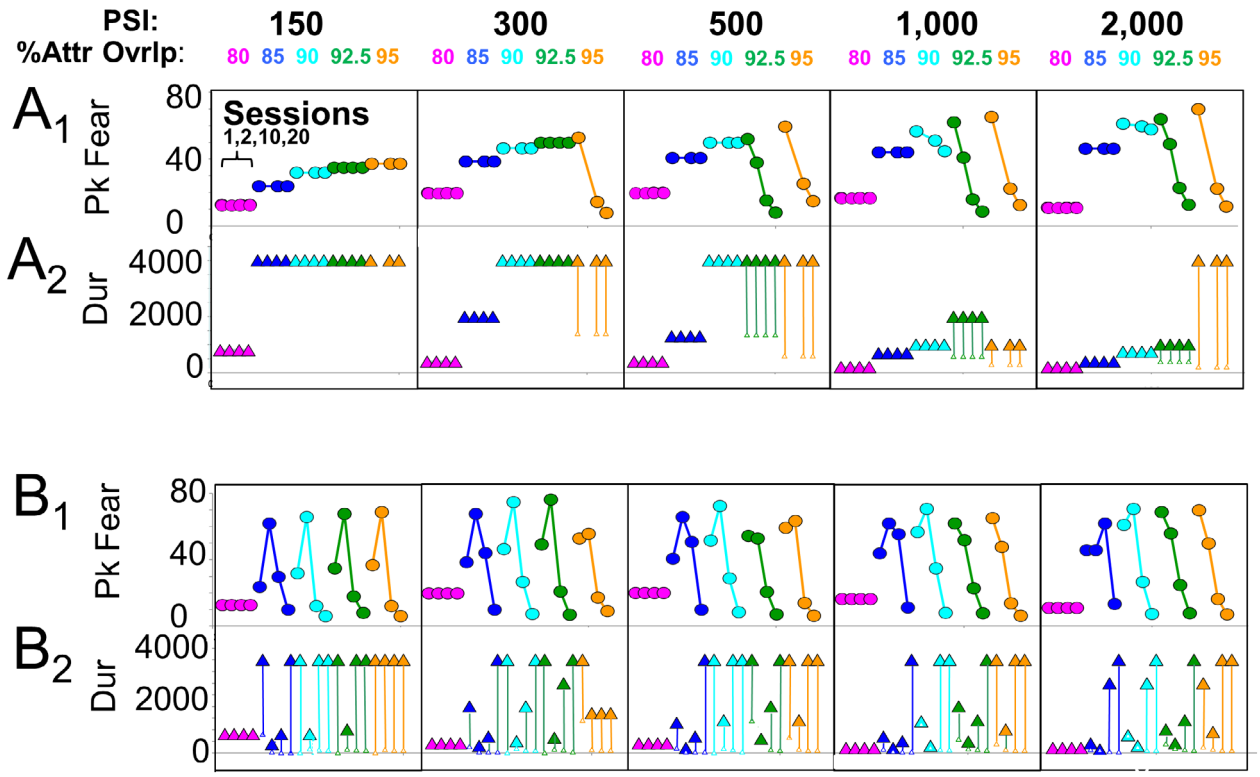

FIGURE 14 Exposure therapy with high versus low B_{add}. Fictive exposure therapy of fear conditioned with a 56% maximal US at various PSIs (*top row of figure*) was carried out in contexts having a range of overlaps with the conditioning context (*second row of figure*). In all cases, "therapy" sessions were made as long as could be done without causing creation of a representation of the therapy context itself. In A, B_{add} was our standard value of 15 and in B it was 2.5. A_1 and B_1 plot peak fear on the first, second, tenth and last of 20 extinction sessions. Comparison of fear in A and B shows clearly that exposure therapy can be much more effective if B_{add} is low, though as discussed in the text there are complications to carrying it out effectively when this is the case. The large upper triangles in A_2 and B_2 plot the session lengths used, and the small lower triangles indicates the shortest session for which B_{Rep} would reach B_{xBtwn} by the end of the session

extinction in B, and slight. However, BaconX became very familiar with B during its many extinction sessions there. So it soon "realized" that C was a new place and formed a representation of it. Once that happened, fear got promoted by the generalization-attenuated strong excitation from A and suppressed by the generalized weak inhibition from B. Fear was therefore substantial. Overall, the results on initial exposure to context C would be a mix of the situations portrayed in 1A and B, and substantial fear would be seen. On the second (and subsequent) sessions the situation would be similar to that in 1B once Rep C was created, and fear substantial. If context C was already familiar upon the first post-extinction exposure to it, Rep C would be activated at the start of the session and fear would be similar to that shown for the second session.

If short sessions in B were used to extinguish BaconX (Panel III), Rep A was active during all extinction sessions and overall behavior was similar to that in Panel I. One had essentially fooled the hippocampus into "thinking" it was in context A during extinction.

An initial trauma leading to clinically significant fear might occur at various PSIs, and virtual reality therapy situations might vary in the degree to which they mimic the actual conditioning situation. Figure 14a explores the course of exposure therapy using the above strategy at a variety of PSIs and contextual attribute overlaps. A complication of this approach is that session durations must be long enough so that B_{Rep} reaches B_{xBtwn} but not so long that a representation of the exposure therapy context gets created, and these maximum and minimum values vary as a function of PSI and degree of similarity of the conditioning and therapy contexts. Overall, successful therapy would require therapy contexts that emulate the conditioning context quite faithfully, and the PSI of conditioning could not have been too terribly short (below 150 intervals in the present instance).

Remediating unextinguishability caused by short-PSI conditioning

We have seen that when BaconX is conditioned at too-short PSIs, the fear it learns can be extremely resistant to extinction. In principle, this could to some extent be remedied by a return visit to the conditioning context to allow updating of the representation, after which exposure therapy could be carried out in the clinic as described in the previous section; but in most cases, return visits would be impractical or impossible.

However, we have seen in our discussion of updating (see Figure 7) that if B_{add} were lower than we had supposed, it might be possible to add attributes of a therapy context to the actual

conditioned representation, and this could possibly restore a capacity for between-session extinction in the clinic. Since there is some evidence that B_{add} values might in fact be surprisingly low, at least in mice (Zinn et al., 2020), we have explored how this might affect our simulations of exposure therapy. Simulations of BaconX "exposure therapy" when B_{add} is low are shown in Figure 14b. When compared to therapy with B_{add} at our standard high value, as shown in Figure 14a, it is immediately apparent that there is a much greater range of conditions under which exposure therapy would be successful at low than at high B_{add} values. However, there are two points that should be noted: (1) When B_{add} is high, so that updating does not occur in the therapy context, as in Figure 14a, effective therapy session durations depend on PSI and the similarity of the therapy context to the context, but all sessions can be of the same length. However, when B_{add} is low enough that updating in the therapy context can occur, effective session lengths vary from one session to another. Initially they must be long enough so that B_{Rep} reaches B_{Rep} but not so long that a new representation gets created. However, since Z_{rec} is higher on the next session, durations often have to be reduced to avoid B_{Rep} falling below B_{new} by the end of the session. However, as more and more of the therapy context attributes get added to the conditioned representation, sessions can become longer without mismatch between recalled and current attributes triggering formation of a new representation, and this allows more efficient extinction. (2) Unless the B_{Rep} value at which maximal fear expression occurs (B_{ff}) is reduced below the level used in most of our simulations, it will often be the case that early conditioning sessions will cause an *increase* of fear as the result of updating (as in many of the sessions of Figure 14b); however, this effect is overcome as therapy sessions continue.

4 | DISCUSSION

The present paper presented a neurocomputational model of context fear extinction in which the creation and activation of hippocampal contextual representations is similar to that in previous Marr-inspired models of hippocampal function (e.g., O'Reilly & McClelland, 1994; Treves & Rolls, 1994) but in which control of representation creation by the hippocampus, updating of contextual information associated with such representations and also conditionability within the amygdala as well as formation of inhibitory associations required for extinction is controlled by estimates of how sure an individual (or the model) is as to the validity of its currently active contextual representation. The metric used for degree of certainty is the Bayesian weight of evidence (B_{Rep}), which is computed, presumably extrahippocampally, by comparing what the individual has so-far observed about its current context and what it remembers about the context of the currently active representation. The model is an extension of a previous model BACON (Krasne et al., 2015) that did not deal with extinction. We have added to the model the assumption that extinction is due to Hebbian potentiation within the amygdala of synapses carrying hippocampal representation information to fear-inhibiting neurons. Particularly crucial to the model are the assumptions that (a) the inhibitory learning presumed to underlie extinction should consolidate only if subjects are sure that they really are in the place they "think" they are and (b) newly observed information about a context can be added to what is remembered about it, even after its initial encoding, if subjects are sufficiently sure about their identification of the current location.

As pointed out throughout Section 3, the model makes a number of testable predictions. We have summarized most of them in Table 2 and will discuss them further below. As indicted in the table, some of the model's predictions have been verified, and as far as we know, none disconfirmed. However, a number remain to be tested.

Below we review and discuss the main implications of the model, alternatives to a few of its assumptions, and some implications possibly relevant to the treatment of fear disorders.

4.1 | Implications/predictions of the model

4.1.1 | Extinction that will endure beyond the session in which it occurred requires confidence as to where one is

A crucial assumption we made, which followed from the spirit of the general approach used here, was that between-session extinction should occur only if an animal is confident that the context in which nonreinforcement is occurring is really the same context where conditioning itself had occurred. This led to the prediction that conditioning which occurs at very short PSIs would be difficult or impossible to extinguish (Table 2, G). This could be highly relevant to real-life situations like roadside bombs in which traumatic events occur without individuals having had time to fully apprise their surroundings. It also led to the prediction that between-session extinction should occur only if the extinction session was sufficiently long, and the length would have to be greater, the shorter the PSI of original conditioning (Table 2, H). As we saw in Figure 6, results consistent with these predictions have recently been demonstrated experimentally (Zinn et al., 2020).

It should be noted that in these figures almost maximal fear conditioning occurs at PSIs well short of those that allow between-session extinction. In the model, this followed from the supposition that animals would "rather be safe than sorry": Context fear conditioning would be successful even though so little had been observed about the situation in which a fear-producing event had occurred that an individual might not be able to identify that situation with certainty in the future (see Krasne et al., 2015), whereas extinction would, as said above, require much greater certainty. The relevant parameters of the model (Table 1) were set accordingly. Experimentally, Leake, Zinn, Corbit, and Vissel (2017) found that biochemical events thought to reflect contextual representation formation (arc expression) continued to increase as a function of PSI well beyond the time required for

TABLE 2 Predictions

	Prediction	Evidentiary status
A	Extinction of context fear is due to learned inhibition of fear-causing neurons	Little direct evidence, but presumed from cued fear studies
B	Extinction-mediating inhibition should be applied to neurons downstream of those whose excitation was increased by conditioning	Inhibition thought to be applied at amygdala intercalated cells (reviewed in Pare and Duvarci (2012)). Herry et al. (2008) provide evidence that extinction potentiates responses of projection neurons ("extinction" neurons) presumed to cause downstream inhibition
C	Strong context fear should generalize more than weak	?
D	Extinguished context fear should return in other similar contexts, and more so when the original fear was strong	?
E	Extinction of context fear should generalize better if carried out in the context of conditioning than in contexts to which fear has generalized	Huckleberry, Ferguson, and Drew (2016); Radulovic, Kammermeier, and Spiess (1998); and Zinn et al. (2020) provide evidence that extinction generalizes better from the conditioning context to other contexts than the reverse
F	Extinction produced by short extinction sessions in unfamiliar contexts that are similar to ones in which context fear was established may generalize better than that extinction produced by longer sessions (because extinction context gets interpreted as the conditioning context, itself)	? Note that in current practice, therapy sessions are usually relatively long
G	Extinction of fear conditioned at short PSIs may fail to consolidate because sparse knowledge of the conditioning context precludes the certainty as to location needed for between-session extinction	Zinn et al. (2020)
H	Extinction sessions that are too short to allow certainty that animal is in the conditioning context should result in poor between-session extinction	Zinn et al. (2020)
I	Postconditioning exposure to the conditioning context can allow acquisition of further information about the conditioning context ("updating") thereby restoring short-PEI extinguishability deficits	Zinn et al. (2020)
J	If updating can occur when animals are less than certain as to their whereabouts, postconditioning exposure to safe contexts similar to feared ones could cause the similar context to become feared and fear of the conditioned context to be lost (context fear "reversal")	Zinn et al. (2020) suggest that at least in mice B_{add} may be rather low, facilitating updating but potentially allowing scrambling of associations as in our Figure 9

plateau-levels of fear conditioning. This was later accompanied by a variety of findings by Zinn et al. (2020), including the data of our Figure $6A_2$ and B_2, all of which support this conclusion. It has also been suggested that fear stemming from conditioning that occurred very early in life may be mediated by what we now call the nonhippocampal "procedural learning" system and that such fear may be difficult to extinguish (Jacobs & Nadel, 1985). From our perspective, this might be expected if the contextual attributes that evoke fear are not encoded in a way that would allow assessment of the certainty that a current context is in fact the one in which conditioning originally occurred.

BaconX was designed to treat between-session extinction as a consolidated form of within-session extinction because this seemed a plausible assumption. However, there has been some discussion in the literature that within and between-session extinction might be independent processes (e.g., Almeida-Corrêa et al., 2015; Plendl & Wotjak, 2010; Toth et al., 2012). Therefore, in an Appendix and Figure 15, we have sketched a modification of BaconX, which we will refer to here as BaconX* in which this is the case. For suitable parameter choices, BaconX* makes exactly the same predictions as BaconX. However, since the circuits driving inhibition in within- and between-session extinction in BaconX* are separate, it would be possible for them to have different pharmacological properties, which the papers cited above indicate might be the case. We should note, however, that even with the assumptions that we have used in BaconX itself, within- and between-session extinction can appear to be rather independent. One can have no between-session extinction if the inhibition responsible for within-session extinction is considerable, but at session's end B_{Rep} is below B_{xBtwn}. And one can get between-session extinction without any visible within-session extinction having preceded it, because within-session inhibition starts increasing during a session as soon as fear begins to rise, but it takes time to grow to a point where there is enough inhibition to counter fear excitation. Therefore, at the end of a short session there can be considerable inhibition available to consolidate if B_{Rep} is greater than Bx_{Btwn}, even though expressed fear has shown no sign of diminishing during the session.

4.1.2 | Knowledge of a feared situation can be acquired even after fear has been conditioned to it and can improve its extinguishability

According to the BaconX model, newly observed contextual attributes can become associated with a context's representation ("updating") whenever BaconX has sufficiently high confidence as to its location (i.e., $B_{Rep} > B_{add}$). Therefore updating can occur even after conditioning has occurred, and if it does, it will become more likely that during extinction in a feared situation BaconX will have enough confidence as to its location to allow such extinction as occurs to become consolidated. Ability for postconditioning updating to restore a capacity for extinction following low PSI conditioning has recently been shown experimentally (Zinn et al., 2020——noted in our Table 2, I).

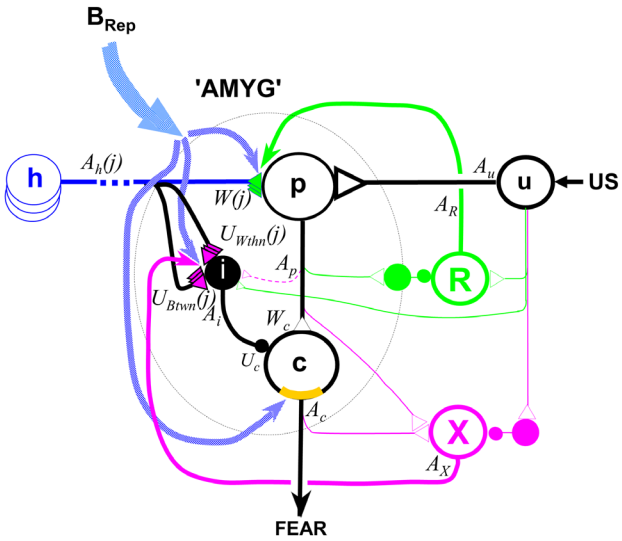

FIGURE 15 BaconX*. A version of BaconX in which within and between-session extinction are independent. See Appendix (Supplementary Matrials) for full explanation

4.1.3 | Erroneous updating of a conditioned contextual representation that is impoverished due to low PSI conditioning can occur in a context that is similar and can lead to fear being expressed in the safe rather than the dangerous context ("fear reversal")

We had originally assumed that an animal would have to be very confident as to its whereabouts before it should update a representation, because otherwise attributes not belonging to a situation might sometimes become associated with its representation, which could lead to maladaptive consequences. However, this assumption, together with the assumption that between-session extinction should require high confidence as to one's whereabouts, means that after very low PSI conditioning, extinction would *never* become possible. We therefore explored the effect of lowering B_{add}. We found that, as expected, it allowed updating that restored a capacity for extinction after conditioning with PSIs too short to be remediated by updating at higher B_{add} values. However, lowering B_{add} could also allow updating that resulted in fear of a dangerous situation to be lost while fear of a safe one was gained (Figure 9 and Table 2, J). Moreover, much to our surprise, there are recent data suggesting that this can actually happen (Zinn et al., 2020). So B_{add} might in fact be lower than one might expect. It may well be that from an adaptive perspective it is worth risking some inappropriate updating to make it possible to extinguish and undergo discrimination learning after conditioning that occurred at PSIs that would otherwise not allow this.

4.1.4 | Extinction of context fear, especially strong fear, is relatively context-specific

It is well known that extinction of cued fear is quite context-specific. Extinction to a discrete CS executed in one context does not transfer well to other contexts (the phenomenon of "renewal," Bouton, 2004). This is presumably adaptive because a cue that predicts danger in one situation may be innocuous in another. As we have seen in Figure 10 and Table 2, D, BaconX exhibits an analogous property in that extinguished context fear tends to return in contexts different from but similar to the conditioned context in which it was extinguished.

Although some extinguished fear returns as a result of context change with both cue and context conditioning, the mechanisms, as portrayed here for context fear and in a related model, (FRAT, Krasne, Fanselow, & Zelikowsky, 2011) for cued fear, are a little different. In both cases, inhibition becomes associated to the cue and context representational elements that are present during extinction. In cued fear Renewal experiments, when the extinguished CS is tested in a context different from that in which it had been extinguished, representational elements of the context in which extinction occurred and to which inhibition is associated are removed while those of the cue itself remain, so inhibition is reduced and Renewal occurs. However, according to the present model, when a subject that was context fear conditioned and then extinguished in context A is moved to a different but similar context B, the representational elements of B that are common to A have both fear excitation *and* inhibition associated with them, but fear returns because, due to the way the amygdala was designed (see relevant Section 3), fear excitation, if it was strong, generalizes more than does inhibition, and so some fear returns (as illustrated in Figure 10).

Although the present model is concerned with extinction of contextual fear, it also has implications for cued fear in so far as cued fear is influenced by context. For example, in renewal experiments we might expect that an animal that was cue conditioned and extinguished in context A and responses to the CS were then tested in a not too different context B would show more renewal if the CS were presented later, rather than earlier in the session. This is because the representation of A, which would inhibit fear caused by the CS, would probably be active until a representation of context B, which would not inhibit fear, got created, and this would not happen until

there had been enough time to sample sufficient contextual attributes to "realize" that the current context was not A. It should be possible to generate other related predictions.

4.1.5 | Extinction of primary context fear generalizes more than extinction of generalized fear

Design of an amygdala circuit that caused less generalization of extinction than of fear itself predicted certain asymmetries in the generalization of extinction. Extinction carried out within the conditioning context should generalize better to similar contexts than the reverse, and extinction of generalized fear should itself generalize poorly to other contexts to which fear had also generalized (Table 2, E). There is laboratory evidence of the first relationship (Huckleberry et al., 2016; Radulovic et al., 1998; Zinn et al., 2020), and the latter one, though as far as we know not yet evaluated in animal experiments, may be one reason for some failures of human exposure therapy.

4.1.6 | Plausible generalization of context fear and its extinction requires that the inhibition responsible for extinction be applied downstream of the locus of conditioning itself

As explained in Section 3, it seemed to us that to be optimally adaptive, extinction of contextual fear should generalize less strongly than fear itself. We then unexpectedly found that, given the biophysically plausible assumptions about interactions of excitation and inhibition that we had started with, this property of fear extinction required that the inhibition responsible for extinction be applied downstream from the amygdala neurons that were responsible for driving the fear (Figure 2b). There is considerable experimental evidence that extinction-causing inhibition is in fact applied downstream of the amygdala nuclei where fear is initiated (see Pare & Duvarci, 2012 for a review). It thus seems possible that we have stumbled onto the, or a reason, that the biological circuit is organized in this way.

4.2 | Role of PFC

Extinction-causing inhibition in BaconX is produced by a circuit that is entirely within the amygdala and does not directly involve the PFC, despite the common view that extinction-causing inhibition reaches the amygdala via the PFC. We have constructed the model in this way because it appears that the PFC is not required for development of within-session extinction (Milad & Quirk, 2012), nor does it appear to be required for suppression of fear whose extinction has already become consolidated (Do-Monte, Manzano-Nieves, Quiñones-Laracuente, Ramos-Medina, & Quirk, 2015, but see Marek et al., 2018). It appears to us that where PFC does play a role is in the *establishment* of between-session extinction (Milad & Quirk, 2012). Between-session extinction in BaconX depends on there being a high level of B_{Rep} at the end of an extinction session. It is often thought that metacognitive processing, of which assessments of certainty such as B_{Rep} would be an example, requires PFC (e.g., Bang & Fleming, 2018; Bartel, Marko, Rameses, Lamm, & Riečanský, 2020; Gherman & Philiastides, 2018; Qiu et al., 2018). We therefore conjecture that this might be why it would play a crucial role in the consolidation of extinction. However, we note that if disruptions of normal PFC functioning can indeed affect B_{Rep}, then according to BaconX, the effect of such disruptions on expression of previously established extinction would potentially be complex, because B_{Rep} affects both expression of any uninhibited fear and growth of inhibition of non-reinforced fear within a session.

4.3 | Extinction as a situational change

Our model postulates that the neurons that drive the inhibitory neurons that are responsible for suppressing fear in extinction are excited, via Hebb synapses, by the same set of contextual representation cells that drive fear itself. However, recently Lacagnina et al. (2019) (see also Tronson et al. (2009)) have provided evidence that different representation cells may be active in the conditioned and extinguished animal and that the extinction-causing representation inhibits the fear-causing one. Although BaconX assumes otherwise, these results are not fundamentally at odds with the approach taken here. It would be perfectly plausible to imagine that the unconditioned stimulus that is responsible for conditioning of fear acts as one of the attributes of the context that excite the fear-causing representation. If so, an animal that has judged (by some process that would have to be determined) that this attribute is now absent might create a new representation of the otherwise same context, but lacking that attribute. If this were to happen, the model as-is would predict that the extinction representation and the conditioning representation would mutually suppress one another, as they do in the Lacagnina et al.'s experiments. It would be especially interesting to compare a model that is structured in this way to the present model during simulations of repeated, alternating periods of conditioning and extinction. This is something we hope to do in the near future.

4.4 | Ideas for exposure therapy suggested by BaconX

4.4.1 | Generalization of extinction developed during exposure therapy might be improved if creation of an exposure therapy representation could be prevented

As shown in Figure 13, BaconX predicts that generalization of extinction which was executed in a context intended to mimic a virtual reality exposure therapy situation could be greatly enhanced if extinction sessions were kept short so that no representation of the "therapy" situation could develop. This would be the case because extinction-

causing inhibition would become associated with the conditioning-context representation itself rather than a representation of the "therapy session" context.

In experiments on rodents, it is plausible to attempt to prevent formation of a therapy situation representation by keeping sessions too short for B_{Rep} to go below B_{new} so that the representation of the actual conditioning situation would be the best representation available. However, with human patients undergoing exposure therapy in real life, it may be difficult to "trick" the hippocampus in that way. Nevertheless, it does seem imaginable that virtual reality situations could be devised that are similar enough to the real feared situation to evoke fear but that differ in ways that make the patient always insufficiently confident that the therapy situation is not actually the real one so that B_{Rep} never would never fall below B_{new}. Another possible approach is pharmacological. There is evidence that ACh may be important for the formation of new representations and that anticholinergic drugs may interfere with new encoding (Hasselmo & McGaughy, 2004; Schon et al., 2005). This has led to experiments in both animals and humans (Craske, Fanselow, Treanor, & Bystritksy, 2019; Zelikowsky et al., 2013) which suggest that, as our model would predict, generalization of extinction may be enhanced by anticholinergic drugs. We would predict that when such drugs were operative the most appropriate *pre-existing* representation (that of the conditioning context) would be activated both in the therapy context and in a similar real-world context, and this would improve generalization of extinction.

4.4.2 | During exposure therapy, extinction failures that are due to impoverished memory of the conditioning context might be reparable by updating the conditioning context representation with therapy situation attributes

When BaconX is conditioned at too-short PSIs, the fear it has learned can be extremely resistant to long-term extinction because not enough was learned about the conditioning context so that BaconX can ever become sure enough as to the identity of its current location to allow consolidation of whatever extinction has occurred (i.e., B_{Rep} can never reach B_{xBtwn}). This sort of problem would be expected to be exacerbated in a virtual reality therapy situation where the emulation of the conditioning context would be far short of perfect. However, because degree of confidence in active representation validity that is required for updating might be much lower than one would expect (as discussed in relation to Figures 8 and 9), it is possible that exposure therapy extinction deficits that are due to low-PSI conditioning could be repaired by associating *therapy situation* attributes with the representation of the conditioned context itself during therapy. We simulated this approach in Figure 14b. The approach required adjusting the degree of similarity of the therapy and conditioning contexts and the duration of the therapy session in a way that was in accordance with the PSI of conditioning, since, as seen in Figure 14, exposure therapy duration must be appropriately matched to the PSI of conditioning. This would of course be exceedingly difficult during real therapy. Nevertheless, we imagine that exploration of approaches motivated by this idea might have translational value.

5 | CONCLUSION

The construction and study of BACON showed that many properties of context fear acquisition could be made more comprehensible by a model in which an extra-hippocampally computed estimate of the validity of its currently active hippocampal representation controls representation formation in the hippocampus and association to hippocampal representations in the amygdala. Here, we have shown that a similar approach has the potential to greatly help us understand the properties of context fear extinction and suggests ways of perhaps translating that understanding to the design of fear disorder therapies.

ACKNOWLEDGMENTS

This work was supported by NIMH grant RO1-MH62122 to MSF and by the Boyarsky Family Trust, Tony and Vivian Howland-Rose, Doug Battersby, and David King and Australian Research Council Discovery project grant number 200102445 to B. V.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Franklin B. Krasne https://orcid.org/0000-0003-2554-1789

REFERENCES

Almeida-Corrêa, S., Moulin, T. C., Carneiro, C. F., Gonçalves, M. M., Junqueira, L. S., & Amaral, O. B. (2015). Calcineurin inhibition blocks within-, but not between-session fear extinction in mice. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *22*(3), 159–169. https://doi.org/10.1101/lm.037770.114

Bae, S. E., Holmes, N. M., & Westbrook, R. F. (2015). False context fear memory in rats. *Learning & Memory*, *22*(10), 519–525. https://doi.org/10.1101/lm.039065.115

Bang, D., & Fleming, S. M. (2018). Distinct encoding of decision confidence in human medial prefrontal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(23), 6082–6087. https://doi.org/10.1073/pnas.1800795115

Bartel, G., Marko, M., Rameses, I., Lamm, C., & Riečanský, I. (2020). Left prefrontal cortex supports the recognition of meaningful patterns in ambiguous stimuli. *Frontiers in Neuroscience*, *14*, 152. https://doi.org/10.3389/fnins.2020.00152

Bernier, B. E., Lacagnina, A. F., Ayoub, A., Shue, F., Zemelman, B. V., Krasne, F. B., & Drew, M. R. (2017). Dentate gyrus contributes to retrieval as well as encoding: Evidence from context fear conditioning, recall, and extinction. *The Journal of Neuroscience*, *37*(26), 6359–6371. https://doi.org/10.1523/JNEUROSCI.3029-16.2017

Blair, H. T., Schafe, G. E., Bauer, E. P., Rodrigues, S. M., & Ledoux, J. E. (2001). Synaptic plasticity in the lateral amygdala: A cellular hypothesis of fear conditioning. *Learning & Memory*, *8*, 229–242. https://doi.org/10.1101/lm.30901

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning & Memory*, *11*, 485–494. https://doi.org/10.1101/lm.78804

Brewin, C. R., Gregory, J. D., Lipton, M., & Burgess, N. (2010). Intrusive images in psychological disorders: Characteristics, neural mechanisms, and treatment implications. *Psychological Review*, *117*, 210–232. https://doi.org/10.1037/a0018113

Bucci, D. J., Saddoris, M. P., & Burwell, R. D. (2002). Contextual fear discrimination is impaired by damage to the postrhinal or perirhinal cortex. *Behavioral Neuroscience*, *116*(3), 479–488.

Castellucci, V., & Kandel, E. R. (1976). Presynaptic facilitation as a mechanism for behavioral sensitization in aplysia. *Science*, *194*(4270), 1176–1178. https://doi.org/10.1126/science.11560

Craske, M. G., Fanselow, M., Treanor, M., & Bystritksy, A. (2019). Cholinergic modulation of exposure disrupts hippocampal processes and augments extinction: Proof-of-concept study with social anxiety disorder. *Biological Psychiatry*, *86*(9), 703–711. https://doi.org/10.1016/j.biopsych.2019.04.012

Do-Monte, F. H., Manzano-Nieves, G., Quiñones-Laracuente, K., Ramos-Medina, L., & Quirk, G. J. (2015). Revisiting the role of infralimbic cortex in fear extinction with optogenetics. *The Journal of Neuroscience*, *35*(8), 3607–3615. https://doi.org/10.1523/JNEUROSCI.3137-14.2015

Doron, N. N., & Ledoux, J. E. (1999). Organization of projections to the lateral amygdala from auditory and visual areas of the thalamus in the rat. *The Journal of Comparative Neurology*, *412*, 383–409.

Fanselow, M. S. (1982). The post-shock activity burst. *Animal Learning & Behavior*, *10*, 448–454.

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

Fanselow, M. S. (2000). Contextual fear, gestalt memories, and the hippocampus. *Behavioural Brain Research*, *110*(1–2), 73–81. https://doi.org/10.1016/s0166-4328(99)00186-2

Fanselow, M. S., & Gale, G. D. (2000). Amygdala. In G. Fink (Ed.), *Encyclopedia of stress* (Vol. 1, pp. 178–182). San Diego, CA: Academic Press.

Fournier, D. I., Eddy, M. C., DeAngeli, N. E., Huszár, R., & Bucci, D. J. (2019). Retrosplenial cortex damage produces retrograde and anterograde context amnesia using strong fear conditioning procedures. *Behavioural Brain Research*, *369*, 111920. https://doi.org/10.1016/j.bbr.2019.111920

Gherman, S., & Philiastides, M. G. (2018). Human VMPFC encodes early signatures of confidence in perceptual decisions. *eLife*, *7*, e38293. https://doi.org/10.7554/eLife.38293

Giustino, T. F., & Maren, S. (2015). The role of the medial prefrontal cortex in the conditioning and extinction of fear. *Frontiers in Behavioral Neuroscience*, *9*, 298. https://doi.org/10.3389/fnbeh.2015.00298

Hasselmo, M. E., & McGaughy, J. (2004). High acetylcholine levels set circuit dynamics for attention and encoding and low acetylcholine levels set dynamics for consolidation. *Progress in Brain Research*, *145*, 207–231. https://doi.org/10.1016/S0079-6123(03)45015-2

Helmstetter, F. J. (1992). The amygdala is essential for the expression of conditional hypoalgesia. *Behavioral Neuroscience*, *106*(3), 518–528. https://doi.org/10.1037//0735-7044.106.3.518

Herry, C., Ciocchi, S., Senn, V., Demmou, L., Müller, C., & Lüthi, A. (2008). Switching on and off fear by distinct neuronal circuits. *Nature*, *454*(7204), 600–606. https://doi.org/10.1038/nature07166

Huckleberry, K. A., Ferguson, L. B., & Drew, M. R. (2016). Behavioral mechanisms of context fear generalization in mice. *Learning & Memory*, *23*(12), 703–709. https://doi.org/10.1101/lm.042374.116

Jacobs, W. J., & Nadel, L. (1985). Stress-induced recovery of fears and phobias. *Psychological Review*, *92*(4), 512–531.

Kandel, E. R., Schwartz, J. H., & Jessell, T. M. (2000). *Principles of neural science* (4th ed.). New York, NY: McGraw-Hill 1414 pp.

Kass, R. E., & Rafter, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Kim, J. J., & Fanselow, M. S. (1992). Modality-specific retrograde amnesia of fear. *Science*, *256*(5057), 675–677. https://doi.org/10.1126/science.1585183

Kim, J. J., Rison, R. A., & Fanselow, M. S. (1993). Effects of amygdala, hippocampus, and periaqueductal gray lesions on short- and long-term contextual fear. *Behavioral Neuroscience*, *107*(6), 1093–1098. https://doi.org/10.1037//0735-7044.107.6.1093

Krasne, F. B., Cushman, J. D., & Fanselow, M. S. (2015). A Bayesian context fear learning algorithm/automaton. *Frontiers in Behavioral Neuroscience*, *9*, 112. https://doi.org/10.3389/fnbeh.2015.00112

Krasne, F. B., Fanselow, M. S., & Zelikowsky, M. (2011). Design of a neurally plausible model of fear learning. *Frontiers in Behavioral Neuroscience*, *5*, 4. https://doi.org/10.3389/fnbeh.2011.00041

Lacagnina, A. F., Brockway, E. T., Crovetti, C. R., Shue, F., McCarty, M. J., Sattler, K. P., … Drew, M. R. (2019). Distinct hippocampal engrams control extinction and relapse of fear memory. *Nature Neuroscience*, *22*(5), 753–761. https://doi.org/10.1038/s41593-019-0361-z

Leake, J., Zinn, R., Corbit, L., & Vissel, B. (2017). Dissociation between complete hippocampal context memory formation and context fear acquisition. *Learning & Memory*, *24*(4), 153–157. https://doi.org/10.1101/lm.044578.116

Lester, L. S., & Fanselow, M. S. (1986). Naloxone's enhancement of freezing: Modulation of perceived intensity or memory processes?. *Physiological Psychology*, *14*(1-2), 5–10.

Li, P., & Zhuo, M. (1998). Silent glutamatergic synapses and nociception in mammalian spinal cord. *Nature*, *393*(6686), 695–698. https://doi.org/10.1038/31496

Liberzon, I., & Abelson, J. L. (2016). Context processing and the neurobiology of post-traumatic stress disorder. *Neuron*, *92*, 14–30. https://doi.org/10.1016/j.neuron.2016.09.039

Lingawi, N. W., Andrew, E., Laurent, V., Killcross, S., Westbrook, R. F., & Holmes, N. M. (2018). The conditions that regulate formation of a false fear memory in rats. *Neurobiology of Learning and Memory*, *156*, 53–59. https://doi.org/10.1016/j.nlm.2018.10.009

MacDermott, A. B., Role, L. W., & Siegelbaum, S. A. (1999). Presynaptic ionotropic receptors and the control of transmitter release. *Annual Review of Neuroscience*, *22*, 443–485. https://doi.org/10.1146/annurev.neuro.22.1.443

Marek, R., Jin, J., Goode, T. D., Giustino, T. F., Wang, Q., Acca, G. M., … Sah, P. (2018). Hippocampus-driven feed-forward inhibition of the prefrontal cortex mediates relapse of extinguished fear. *Nature Neuroscience*, *21*(3), 384–392. https://doi.org/10.1038/s41593-018-0073-9

Maren, S., Phan, K. L., & Liberzon, I. (2013). The contextual brain: Implications for fearconditioning, extinction and psychopathology. *Nature Reviews. Neuroscience*, *14*, 417–428. https://doi.org/10.1038/nrn3492

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *262*(841), 23–81. https://doi.org/10.1098/rstb.1971.0078

McDonald, A. J. (1998). Cortical pathways to the mammalian amygdala. *Progress in Neurobiology*, *55*, 257–332. https://doi.org/10.1016/s0301-0082(98)00003-3

Milad, M. R., & Quirk, G. J. (2012). Fear extinction as a model for translational neuroscience: Ten years of progress. *Annual Review of Psychology*, *63*, 129–151. https://doi.org/10.1146/annurev.psych.121208.131631

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus*, *4*, 661–682. https://doi.org/10.1002/hipo.450040605

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*, 311–345. https://doi.org/10.1037/0033-295X.108.2.311

Orsini, C. A., Kim, J. H., Knapska, E., & Maren, S. (2011). Hippocampal and prefrontal projections to the basal amygdala mediate contextual

regulation of fear after extinction. Version 2. *The Journal of Neuroscience*, *31*(47), 17269–17277. https://doi.org/10.1523/JNEUROSCI.4095-11.2011

Pare, D., & Duvarci, S. (2012). Amygdala microcircuits mediating fear expression and extinction. *Current Opinion in Neurobiology*, *22*(4), 717–723. https://doi.org/10.1016/j.conb.2012.02.014

Phillips, R. G., & LeDoux, J. E. (1992). Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. *Behavioral Neuroscience*, *106*(2), 274–285. https://doi.org/10.1037//0735-7044.106.2.274

Plendl, W., & Wotjak, C. T. (2010). Dissociation of within- and between-session extinction of conditioned fear. *The Journal of Neuroscience*, *30*(14), 4990–4998. https://doi.org/10.1523/JNEUROSCI.6038-09.2010

Qiu, L., Su, J., Ni, Y., Bai, Y., Zhang, X., Li, X., & Wan, X. (2018). The neural system of metacognition accompanying decision-making in the prefrontal cortex. *PLoS Biology*, *16*(4), e2004037. https://doi.org/10.1371/journal.pbio.2004037

Radulovic, J., Kammermeier, J., & Spiess, J. (1998). Generalization of fear responses in C57BL/6N mice subjected to one-trial foreground contextual fear conditioning. *Behavioural Brain Research*, *95*(2), 179–189. https://doi.org/10.1016/s0166-4328(98)00039-4

Rescorla, R. A. (2001). Retraining of extinguished Pavlovian stimuli. *Journal of Experimental Psychology. Animal Behavior Processes*, *27*(2), 115–124.

Rudy, J. W., Barrientos, R. M., & O'Reilly, R. C. (2002). Hippocampal formation supports conditioning to memory of a context. *Behavioral Neuroscience*, *116*(4), 530–538. http://dx.doi.org/10.1037/0735-7044.116.4.530.

Rudy, J. W., Huff, N. C., & Matus-Amat, P. (2004). Understanding contextual fear conditioning: insights from a two-process model. *Neuroscience & Biobehavioral Reviews*, *28*(7), 675–685. http://dx.doi.org/10.1016/j.neubiorev.2004.09.004.

Rudy, J. W., & O'Reilly, R. C. (1999). Contextual fear conditioning, conjunctive representations, pattern completion, and the hippocampus. *Behavioral Neuroscience*, *113*(5), 867–880. https://doi.org/10.1037//0735-7044.113.5.867

Schon, K., Atri, A., Hasselmo, M. E., Tricarico, M. D., LoPresti, M. L., & Stern, C. E. (2005). Scopolamine reduces persistent activity related to long-term encoding in the parahippocampal gyrus during delayed matching in humans. *The Journal of Neuroscience*, *25*(40), 9112–9123. https://doi.org/10.1523/JNEUROSCI.1982-05.2005

Stote, D. L., & Fanselow, M. S. (2004). NMDA receptor modulation of incidental learning in Pavlovian context conditioning. *Behavioral Neuroscience*, *118*(1), 253–257. https://doi.org/10.1037/0735-7044.118.1.253

Toth, I., Dietz, M., Peterlik, D., Huber, S. E., Fendt, M., Neumann, I. D., & Slattery, D. A. (2012). Pharmacological interference with metabotropic glutamate receptor subtype 7 but not subtype 5 differentially affects within- and between-session extinction of Pavlovian conditioned fear. *Neuropharmacology*, *62*(4), 1619–1626. https://doi.org/10.1016/j.neuropharm.2011.10.021

Treves, A., & Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*, 374–391. https://doi.org/10.1002/hipo.450040319

Tronson, N. C., Schrick, C., Guzman, Y. F., Huh, K. H., Srivastava, D. P., Penzes, P., … Radulovic, J. (2009). Segregated populations of hippocampal principal CA1 neurons mediating conditioning and extinction of contextual fear. *The Journal of Neuroscience*, *29*(11), 3387–3394. https://doi.org/10.1523/JNEUROSCI.5619-08.2009

Wiltgen, B. J., Sanders, M. J., Anagnostaras, S. G., Sage, J. R., & Fanselow, M. S. (2006). Context fear learning in the absence of the hippocampus. *The Journal of Neuroscience*, *26*(20), 5484–5491. https://doi.org/10.1523/JNEUROSCI.2685-05.2006

Young, S. L., Bohenek, D. L., & Fanselow, M. S. (1994). NMDA processes mediate anterograde amnesia of contextual fear conditioning induced by hippocampal damage: Immunization against amnesia by context preexposure. *Behavioral Neuroscience*, *108*(1), 19–29. https://doi.org/10.1037//0735-7044.108.1.19

Zelikowsky, M., Hast, T. A., Bennett, R. Z., Merjanian, M., Nocera, N. A., Ponnusamy, R., & Fanselow, M. S. (2013). Cholinergic blockade frees fear extinction from its contextual dependency. *Biological Psychiatry*, *73*(4), 345–352. https://doi.org/10.1016/j.biopsych.2012.08.006

Zelikowsky, M., Hersman, S., Chawla, M. K., Barnes, C. A., & Fanselow, M. S. (2014). Neuronal ensembles in amygdala, hippocampus, and prefrontal cortex track differential components of contextual fear. *The Journal of Neuroscience*, *34*(25), 8462–8466. https://doi.org/10.1523/JNEUROSCI.3624-13.2014

Zinn, R., Leake, J., Krasne, F. B., Corbit, L. H., Fanselow, M. S., & Vissel, B. (2020). Maladaptive properties of context-impoverished memories. *Current Biology*, *30*, 1–12. https://doi.org/10.1016/j.cub.2020.04.040

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Krasne FB, Zinn R, Vissel B, Fanselow MS. Extinction and discrimination in a Bayesian model of context fear conditioning (BaconX). *Hippocampus*. 2021;31:790–814. https://doi.org/10.1002/hipo.23298