HUMAN INDUCTIVE BIASES FOR AVERSIVE CONTINUAL LEARNING — A HIERARCHICAL BAYESIAN NONPARAMETRIC MODEL

Sashank Pisupati [1,2,3*] **Isabel M. Berwian** [1,2*] **Jamie Chiu** [1,2] **Yongjing Ren** [1,2] **Yael Niv** [1,2]

[1] Princeton Neuroscience Institute, Princeton U [2] Dept. of Psychology, Princeton U, [3] Limbic Ltd

ABSTRACT

Humans and animals often display remarkable continual learning abilities, adapting quickly to changing environments while retaining, reusing, and accumulating old knowledge over a lifetime. Unfortunately, in environments with adverse outcomes, the inductive biases supporting such forms of learning can turn maladaptive, yielding persistent negative beliefs that are hard to extinguish, such as those prevalent in anxiety disorders. Here, we present and model human behavioral data from a fear-conditioning task with changing latent contexts, in which participants had to predict whether visual stimuli would be followed by an aversive scream. We show that participants' learning in our task spans three different regimes — with old knowledge either being updated, discarded (forgotten) or retained and reused in new contexts (remembered) by different participants. The latter regime corresponds to (maladaptive) spontaneous recovery of fear. We demonstrate using simulations that these behavioral regimes can be captured by varying inductive biases in Bayesian non-parametric models of contextual learning. In particular, we show that the "remembering" regime can be produced by "persistent" variants of hierarchical Dirichlet process priors over contexts and negatively biased "deterministic" beta distribution priors over outcomes. Such inductive biases correspond well to widely observed "core beliefs" that may have adaptive value in some lifelong-learning environments, at the cost of being maladaptive in other environments and tasks such as ours. Our work offers a tractable window into human inductive biases for continual learning algorithms, and could potentially help identify individual differences in learning strategies relevant for response to psychotherapy.

1 INTRODUCTION

Humans and other biological learning systems often display a remarkable capacity for continually accumulating knowledge throughout their lifetime, which has served as inspiration for continual- and lifelong-learning algorithms at several mechanistic levels (Kudithipudi et al., 2022; Hadsell et al., 2020). Such algorithms are infused with naturalistic inductive biases about the non-stationary, modular and recurrent structure of the environment, and seek to match humans' and animals' ability to adapt quickly to changing environments without overwriting or forgetting old knowledge, but rather protecting and reusing it to generalise to new contexts or when familiar contexts resurface. The "contextual" or "latent-cause inference" framework has been a useful tool for translating these abilities into a common probabilistic language that captures many features of human and animal learning using Bayesian nonparametric models (Gershman et al., 2015; Heald et al., 2021).

However, the very inductive biases that endow humans and animals with such lifelong-learning abilities can turn maladaptive (Pisupati & Niv, 2022), especially in environments with adverse outcomes. For instance, certain forms of anxiety may reflect over-enthusiastic protection of outdated threat associations, making them stubbornly resistant to updating and prone to inappropriately resurfacing in the future even in the absence of threat ("spontaneous recovery of fear"; Gershman & Hartley, 2015; Zika et al., 2022; Goldway et al., 2023). Similar mechanisms that over-protect drug-related associations are thought to underlie relapse phenomena. Finally, mechanisms that enable generalisation between old and new contexts for the purpose of forward or backward transfer may overgeneralise negative experiences to neutral situations, giving rise to the widespread negative biases seen in post-traumatic stress disorder (PTSD) and depression (Norbury et al.; Cohen & Kahana, 2022).

[*] These authors contributed equally

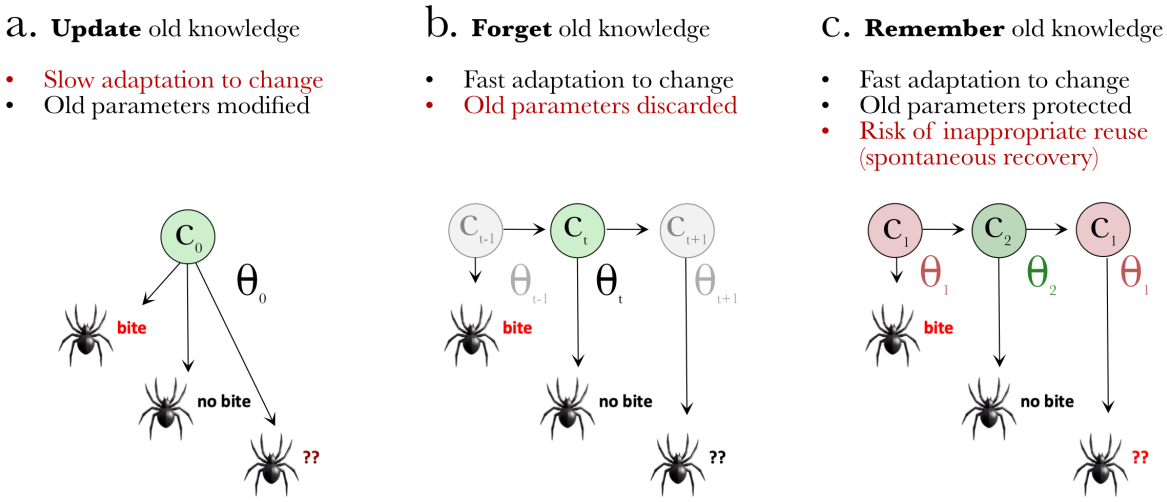

Figure 1: **Learning strategies and their failure modes:** Schematic illustrating three different learning strategies in the face of changing experience, the assumptions of latent structure they reflect and their consequent failure modes. **a. Updating**: When experiences are assumed to arise from a single *stationary* distribution, context or latent-cause C_0, then all experiences update the same set of parameter θ_0 (here, the likelihood of pain due to a bite), leading to slow adaptation to changing or conflicting experiences (e.g. experiences of spiders that don't bite after an experience with a spider that did bite will be slowly integrated together, leading to intermediate predictions for future encounters). **b. Forgetting**: When conflicting experiences are assumed to arise from a single *non-stationary* distribution, context or latent-cause C_t, then past experiences are forgotten and recent experiences more heavily impact the inferred current value of the parameter θ_t, leading to fast adaptation to change, at the cost of discarding past parameter values (e.g. the most recent experience of not being bitten is most influential on future predictions, with past experiences of being bitten being discarded). **c. Remembering**: When experiences can potentially arise from multiple distinct contexts or latent-causes C_i, then conflicting experiences are assumed to arise from different latent causes (in this example, spiders that don't bite are assumed to belong to latent cause C_2, distinct from C_1 to which the biting spiders belong). Separation into different latent causes enables fast adaptation to change while still allowing past parameter values to be "remembered" or retained and protected from updates (the memory of parameter θ_1 for spiders that do bite is unaffected by experiences of those that don't, which modify θ_2). Old parameter values can then be reused in future encounters, which may be adaptive (e.g. when encountering dangerous spiders) or maladaptive (e.g. inappropriately generalizing to safe spiders). This phenomenon of inappropriate reactivation of threat beliefs despite exposure to safe situations is often referred to as "spontaneous recovery of fear."

Humans and animals often display substantial variability in their learning strategies and consequent failure modes (Gershman & Hartley, 2015; Zika et al., 2022; Norbury et al.; Goldway et al., 2023). Individual variability in continual-learning mechanisms can interact with extended courses of psychotherapy for mental health conditions (such as exposure therapy), potentially giving rise to different forms of therapy failure (Pisupati et al., 2021). A precise characterisation of the space of inductive biases underlying individual variability could help design more efficient therapy plans tailored to individuals (Niv et al., 2021), and offer further inspiration to the design and safety evaluation of artificial algorithms for lifelong learning (Schulz & Dayan, 2020).

In this work, we present human behavioral data in an aversive conditioning task with changing latent contexts. We show that participants' behavior aligns with three different learning regimes, distinguished by whether old knowledge is "updated", "forgotten" or "remembered" (illustrated in Figure 1) – with the latter regime being adaptive in many continual learning settings, but turning maladaptive when it leads to the "spontaneous recovery" of outdated fear associations. We simulate data using Bayesian nonparametric models of "contextual" and "latent cause" inference and hand-tuned parameters to match human behavior in the three groups. We show that the three regimes of participants' behavior can be captured by differing assumptions about the stationarity and recurrence of contextual change. Finally, we show that the "remembering" regime is encouraged by persistent priors over contexts, and biased, deterministic priors over negative events. These priors naturally map on to clinically relevant "core beliefs," offering tractable targets for personalising psychotherapy and a potential window into human inductive biases for continual learning.

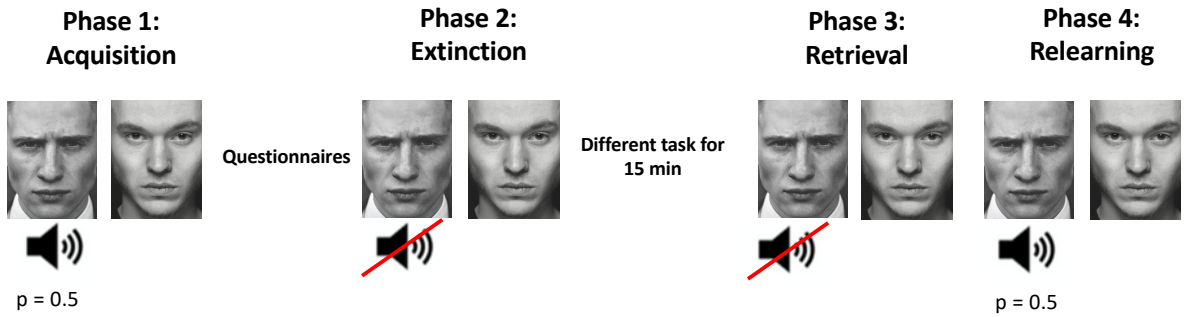

Figure 2: **Task design.** The behavioral fear conditioning task consisted of four phases. In each trial in all phases, participants saw one of two angry faces (depicted images are meant to illustrate the real stimuli). In the acquisition phase, one face (the CS+) was followed by a loud, aversive, but tolerable scream in 50% of the trials. The other face (the CS-) was never followed by a scream. Participants filled out two questionnaires before the second phase, the extinction phase, in which neither of the stimuli was followed by a scream. Participants completed a different behavioral task between the second and the third phase. Phase 3, the retrieval phase, was used to test for spontaneous recovery of the expectation of the US by presenting the stimuli without screams. Finally, phase 4, the relearning phase, was similar to phase 1.

2 METHODS

2.1 BEHAVORIAL TASK

The task design is illustrated in Figure 2. In each trial, participants saw one of two stimuli, and had to press the space bar to see the trial's outcome. Depending on the phase, one stimulus (conditioned stimulus, CS+) was followed on 50% of the trials by an aversive but tolerable scream (unconditioned stimulus, US; participants calibrated the volume to achieve this level before the task started). The other stimulus (CS-) was never paired with the aversive scream. Once every three trials, on average, participants were asked to rate how likely they expected each stimulus to be followed by a scream, on a scale from 0 to 100. Once every ten trials, on average, they were asked to rate how unpleasant, anxious, and fearful each stimulus made them feel.

In the "acquisition" phase, consisting of 26 trials, 50% of CS+ trials were followed by a scream (total: 8 screams; CS+ presented for 16 trials). After a break of approximately 3-5 minutes, in which participants filled out two questionnaires, came an "extinction" phase of 30 trials with no aversive screams. Next, participants completed a separate task for approximately 15 minutes. This was followed by a "retrieval" phase without aversive screams for 16 trials (to test for spontaneous recovery of expectation of the US), which were immediately followed by a "relearning" phase with 16 trials and aversive screams (4 screams; 10 presentations of the CS+).

As the task was administered online to remote participants without supervision of an experimenter, six attention checks were spread throughout the task. On these attention checks, participants were given low-volume auditory instructions to press a specific key (e.g., 'please press the letter f') during the fixation cross between trials. This procedure was implemented to ensure participants did not turn off their audio volume. Participants were also asked to indicate at the end of the task if they had changed their volume at any point. They were instructed that this information would not influence their payment.

2.2 PROCEDURE

Seventy-nine participants from Prolific were recruited to participate in an online behavioral experiment in the fall of 2022. This study was approved by the Institutional Review Board of Princeton University, and all participants provided written informed consent. The total study duration was approximately 50 minutes per participant. Participants received monetary compensation for their time (rate $13/hr). To be eligible for the experiment, participants had to reside in the United States, Canada, Australia, or New Zealand, be fluent in English, have access to headphones, and had not participated in a previous version of this behavioral experiment.

69 Participants completed the experiment. Data from 19 participants who failed more than one of the audio attention checks and six who indicated that they changed their audio volume during the study were excluded before the analyses.

2.3 GENERATIVE MODEL

In order to capture individual differences in learned structure that could give rise to the different observed learning regimes, we formalize latent structure learning on our task using Bayesian nonparametric models of contextual and latent-cause inference (Gershman et al., 2015; Gershman & Blei, 2012; Heald et al., 2021). Fundamentally, this class of models assumes that rather than directly learning associations between different observations (e.g. cues and outcomes), agents are learning about hidden (latent) "causes" that jointly give rise to cues and outcomes.

Let us assume that an agent encounters Bernoulli observations for each i-th feature $O_{i,t} = \mathcal{I}_i$. The latent-cause framework reformulates the learning problem from one of learning a single set of static or dynamic associative parameters Θ between the features, into one of partitioning experiences into an unknown number of latent causes $L_t = j$, each with its own unique set of parameters $\Theta_j = \phi_{i,j}$ that determine observation probabilities of cues and outcomes. Bayesian nonparametric methods allow us to impose infinite-capacity priors over these latent causes, such that agents can flexibly but judiciously add new latent causes when they become necessary to explain surprising observations.

For the purposes of modeling our task, we assume an additional layer of hierarchy – that the same latent causes (with distinct associative parameters) may be encountered with varying frequencies in different "contexts" $C_t = k$, such that each context has a unique distribution over latent causes, but the latent causes themselves are shared across contexts. The three different learning regimes we explore correspond to three different assumptions about contexts:

2.3.1 UPDATING REGIME

To model the updating regime, we assumed a single, stationary context across time, whose distribution over latent causes was given by a standard **Chinese Restaurant Process** (CRP) prior over latent causes, with a concentration parameter ν that determines how often new latent causes are created. This prior assumes an infinite number of "tables" corresponding to latent causes, each with N_j "customers" representing the number of past trials already generated by cause j. The generative model creates the observations on trial t by first choosing a latent cause $L_t = j$ according to:

$$p(L_t = j|\mathbf{L}_{1:t-1}) = \begin{cases} \frac{N_j}{t-1+\nu} & \text{if } j \text{ is an old latent cause} \\ \frac{\nu}{t-1+\nu} & \text{if } j \text{ is a new latent cause} \end{cases} \tag{1}$$

Observations $\mathbf{O}_t$ (e.g. cue, outcome, each a feature $O_{i,t}$) are then generated by independent Bernoulli processes with probabilities $\phi_{i,j}$ for each feature, conditioned on the latent cause $L_t = j$:

$$p(O_{i,t}|L_t = j) = \phi_{i,j}. \tag{2}$$

When a new latent cause j is initialized, the probabilities $\phi_{i,j}$ are drawn from Beta priors with two parameters a_{obs} and b_{obs}. The sum of these parameters controls how "deterministic" the $\phi_{i,j}$ will be (with smaller a_{obs} and b_{obs} leading to $\phi_{i,j}$ that are closer to 0 or 1, hence more deterministic), and their ratio controls how biased they are (with $a_{obs} > b_{obs}$ leading to $\phi_{i,j}$ that are closer to 1 than to 0).

2.3.2 FORGETTING REGIME

In the forgetting regime, we assume a single, non-stationary context over time whose distribution over latent causes continuously evolves over time. This is effectively achieved by using a variant of the **distance-dependent CRP** prior (Song et al., 2022) over latent causes, with an exponentially decreasing distance function whose decay rate τ determines the effective timescale of change:

$$p(L_t = j|\mathbf{L}_{1:t-1}) = \begin{cases} \frac{\Sigma_{t'<t} e^{-\tau(t-t')}\delta(L_{t'},j)}{\Sigma_{t'<t} e^{-\tau(t-t')}+\nu} & \text{if } j \text{ is an old latent cause} \\ \frac{\nu}{\Sigma_{t'<t} e^{-\tau(t-t')}+\nu} & \text{if } j \text{ is a new latent cause} \end{cases} \tag{3}$$

Where $\delta(L_{t'}, j) = 1$ if $L_{t'} = j$ and 0 otherwise. This model reduces to the updating regime when $\tau = 0$.

2.3.3 REMEMBERING REGIME

For the remembering regime, we assume switching dynamics between (an unknown number of) persistent & recurrent contexts, each with its own unique, stable distribution over latent causes. We use the same Bayesian nonparametric techniques as before to impose infinite-capacity priors over the context transition matrix, which allows agents to create new contexts as necessary. We use a **"persistent" variant of the Chinese Restaurant Process** prior over contexts to allow the previous context to remain active with a persistence probability η, with contexts being redrawn from the

Chinese Restaurant Process prior with probability $1-\eta$. Hence, the prior probability of a new trial belonging to context $C_t = k$ is given by:

$$p(C_t = k|C_{t-1} = k', \mathbf{C}_{1:t-2}) = \begin{cases} \eta + (1-\eta) * \frac{N_k}{t-1+\alpha} & \text{if } k = k' \text{ i.e. the previously active latent cause} \\ (1-\eta) * \frac{N_k}{t-1+\alpha} & \text{if } k \neq k' \text{ i.e. any other old latent cause} \\ (1-\eta) * \frac{\alpha}{t-1+\alpha} & \text{if } k \text{ is a new latent cause} \end{cases} \tag{4}$$

Where α is a concentration parameter that determines how frequently new contexts are created. This regime reduces to the updating regime when $\alpha = 0$

We impose a **Hierarchical Dirichlet Process** prior, similar to the one used in Heald et al. (2021) on the context-conditioned probability over latent causes. This allows contexts to share latent causes (drawn from a global distribution over latent causes), while still allowing each context to have a distinct distribution over latent causes (drawn from a Dirichlet process over the global distribution), as well as allowing for the creation of new latent causes.

In practice we use the Chinese Restaurant Franchise representation to sample directly from the local context-conditioned distribution of latent causes. This representation assumes an infinite number of "restaurants" corresponding to contexts $C = k$, each with an infinite number of "tables" $\zeta_k = m$, with each table serving a "dish" $L_\zeta = j$ corresponding to a unique latent cause. Each table has $N_{m,k}$ "customers" representing the number of past trials created by that table (that is, trials created by this latent cause in this context).

On trial t, in order to sample a dish or latent cause $L_t = j$ given a restaurant or context $C_t = k$, we first assign the trial a table $\zeta_{k,t} = m$, with a concentration parameter ν determining the probability of creating a new table:

$$p(\zeta_{k,t} = m|C_t = k) = \begin{cases} \frac{N_{m,k}}{t-1+\nu} & \text{if } m \text{ is an old table} \\ \frac{\nu}{t-1+\nu} & \text{if } m \text{ is a new table} \end{cases} \tag{5}$$

If a new table is created, we sample a dish or latent cause $L_\zeta = j$ for that table using the global distribution over dishes or latent causes, with N_j tables assigned to dish or latent cause $L = j$ across all restaurants or contexts, and a concentration parameter γ determining the probability of invoking a new dish or latent cause:

$$p(L_{\zeta,t} = j) = \begin{cases} \frac{N_j}{t-1+\gamma} & \text{if } j \text{ is an old latent cause} \\ \frac{\gamma}{t-1+\gamma} & \text{if } j \text{ is a new latent cause} \end{cases} \tag{6}$$

2.4 Inference simulations

The three generative models were used to infer, given observations (CSs and USs on each trial), a distribution over latent causes (and contexts), and therefore a prediction of whether a scream would appear. As exact inference is not tractable, we performed approximate inference using particle filtering (for non-hierarchical models) and particle learning (for the hierarchical variant, following Heald et al. (2021)) using sequential importance resampling of 200 particles. At every point in time, the algorithm inferred the posterior probability over possible partitions of trials into latent causes and contexts based on all observations made so far. We used the posterior distribution before US observation to generate a prediction of the probability of the US. We then recalculated the posterior distribution after observing whether a scream US was present in that trial, and used this distribution to update estimates of the observation probabilities given each latent cause (across contexts).

In simulating behavior on the task, the model observed the same stimulus order as participants. To simulate breaks participants took in between acquisition and extinction, and then again before retrieval, we presented the model with a different "dummy" stimulus for a 9-17 trials before the extinction phase and 34 trials before the retrieval phase corresponding to the length of the participants' break.

3 Results

3.1 Behavioral results

Data from 44 participants were included in the analyses. As evident from the average data in Figure 4A, by the end of acquisition, participants learned to differentiate between the CS+ and the CS- and correctly predicted that a scream followed the CS+, but not the CS-. Furthermore, participants decreased their expectations of the negative outcome during the extinction phase. Nevertheless, in the retrieval phase after the break, expectancy ratings increased

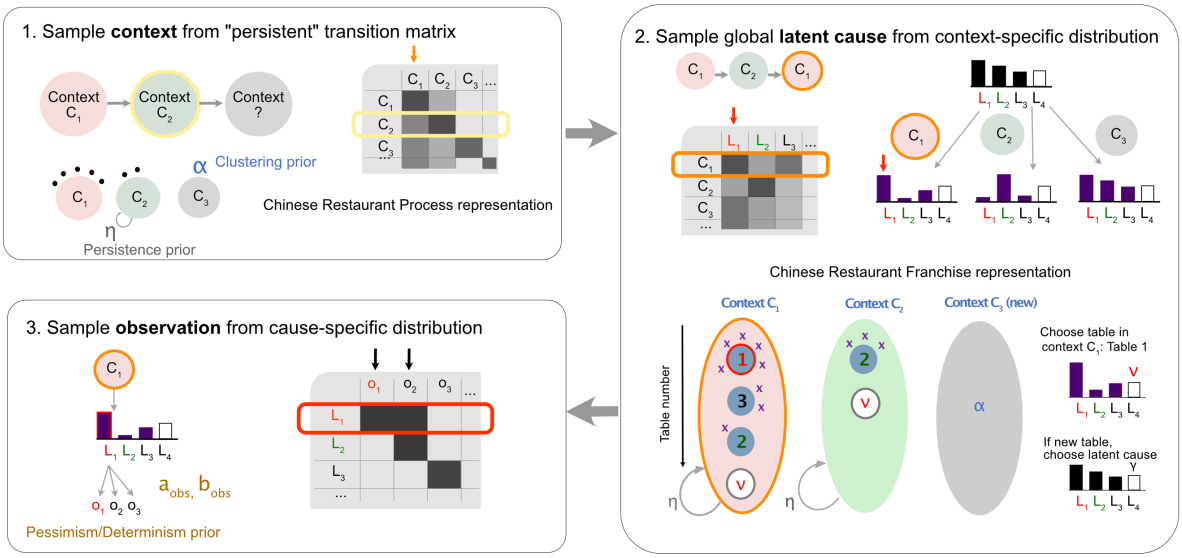

Figure 3: **Illustration of the generative model** The three levels of sampling in the generative model, corresponding to context, latent cause and observation. **1.** Upcoming contexts are sampled from a transition matrix conditioned on the current context (right) whose rows are given by a "persistent" version of the Chinese Restaurant Process (bottom, circles denote tables, and dots represent customers), with persistence η and clustering parameter α. **2.** Latent causes conditioned on contexts are sampled from a hierarchical Dirichlet process, which determines the context-conditioned prior distribution over latent causes (rows in matrix). This is displayed in the bottom of the figure using the analogy to a Chinese Restaurant Franchise. Ovals indicate "restaurants," which correspond to contexts in our model. They contain "tables" (circles inside the ovals) that serve different dishes (numbers inside circles). The dishes correspond to latent causes. Vertical bars in purple represent the context-specific table distribution, and vertical bars in black represent the global distribution of dishes or latent causes. Parameters α, ν, and γ correspond to the concentration parameters for the hierarchical Dirichlet processes over contexts, tables, and latent causes, respectively. Each purple 'x' indicates a customer (trial in the task) assigned to a specific table. The self-transition parameter η denotes the probability of staying in the same context for the next trial. Given the context chosen in step 1, a table is chosen based on the tables' popularity (previous trial assignments) within that context and the probability of creating a new table (proportional to ν). Finally, a dish (latent cause) must be determined if a new table is selected. This choice is based on the popularity of dishes across contexts and the probability of creating a new latent cause (proportional to γ). **3.** Observations conditioned on chosen latent causes are sampled from a Bernoulli process for each observation dimension, with a probability that is drawn from beta priors with parameters a_{obs} and b_{obs}.

(spontaneous recovery) for the CS+, and to a lesser extent, for the CS-. Finally, participants quickly relearned that the CS+ predicted the scream again in the relearning phase. This general pattern of results is in line with the literature (Purves et al., 2019). However, there were considerable individual differences in learning and spontaneous recovery.

To analyze these differences, we divided participants into three groups based on their behavior on the task. First, we excluded from further analysis five participants who did not expect the scream to follow the CS+ at least 10% more than the CS- in the acquisition phase, as this indicated that they did not learn the task contingencies. Remaining participants who did not increase their expectancy of scream for either the CS+ or the CS- by more than 15 % between the end of the extinction phase and the beginning of the retrieval phase were assigned to the 'Updating' group (N=16), while participants who increased their rating for the CS+, but not the CS-, by more than 15 %, were assigned to the 'Remembering' group (N=9) and participants who increased their ratings for both stimuli were assigned to the 'Forgetting' group (N=12). Two participants did not meet criteria for either of our groups and were not included in the following analyses.

These individual differences suggest that human learners differ in their learning strategies and in how they treat old knowledge when confronted with new contexts. The three strategies have features in line with the three different learning regimes described in Figure 1. The 'Updating' group shows slow extinction, no spontaneous recovery in the retrieval phase, and slower relearning. The 'Forgetting' group made similar predictions for the CS+ and the CS- in the retrieval phase, until they learned again, apparently from scratch, that the CS+ is paired with a scream in the relearning

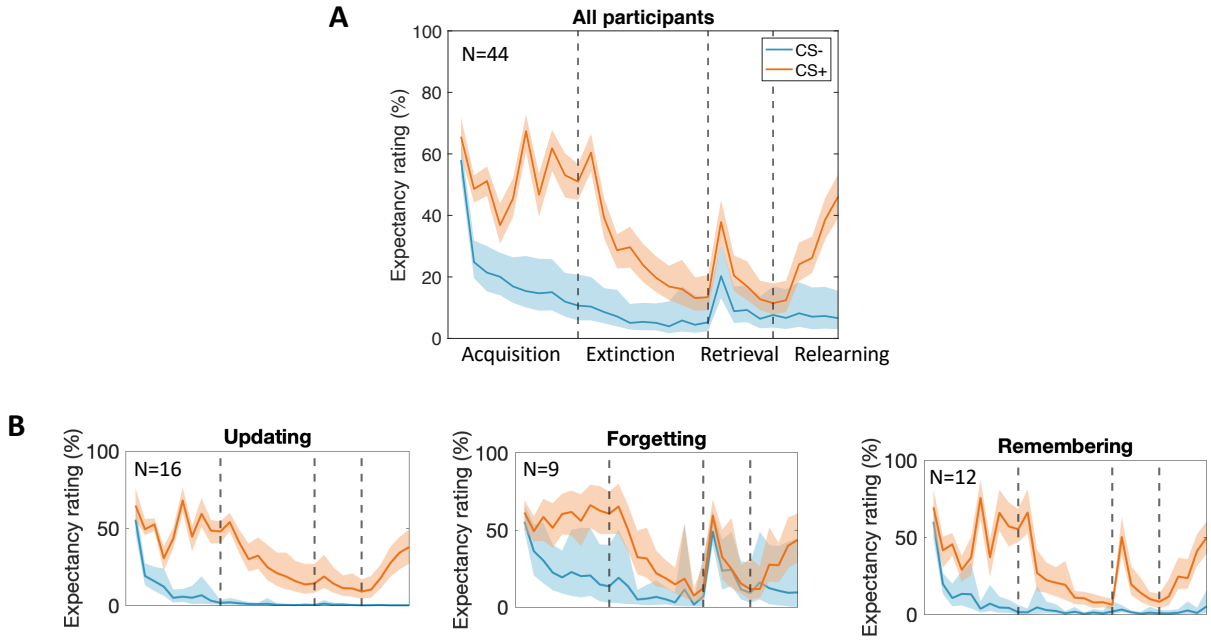

Figure 4: **Empirical results.** Every three trials, on average, participants were asked for expectancy ratings (how likely each stimulus would be followed by an aversive scream). Solid lines show mean expectancy ratings for the CS+ (orange) and the CS- (blue) for the four phases of acquisition, extinction, retrieval, and relearning over the course of the experiment. Dashed vertical lines indicate switches between phases. The first and second phase switches involved breaks of approximately 5 and 15 minutes, respectively. Shaded areas indicate 95% bootstrapped confidence intervals. N indicates the number of participants included in each group. **A)** Behavior of all participants. **B)** Behavior of the "Updating", "Forgetting", and "Remembering" groups.

phase. This suggests resetting of learning after the long break, and forgetting of any predictive difference between the stimuli. The 'Remembering' group, in contrast, showed fear (i.e., prediction of scream) of the CS+ but not the CS- in the retrieval phase. These participants extinguished their fear responses faster than did the 'Updating' group in the extinction phase, and relearned faster in the relearning phase, suggesting a new context was initiated in extinction, preserving the learning in the old context for future retrieval and use.

3.2 Inference and Simulations

We simulated behavior using different model regimes to observe the influence of the model structure on outcome predictions and latent cause and context inferences.

First, we simulated the model in the updating regime with a standard CRP prior. The parameter ν for creating new tables (equivalent to latent causes in this model) was set to 0.2, a relatively large number to allow for creation of more latent causes, and both Beta prior parameters were set to 0.001 to generate deterministic observation probabilities for the CSs and the US for each latent cause. Together, these parameters encourage the model to infer a different latent cause for each combination of observations. Simulations of outcome predictions under these settings (Figure 5A) showed slow adaption during the extinction phase, no spontaneous recovery during the retrieval phase, and very slow relearning in the last phase for the CS+. This behavioral pattern closely matched the behavior of the 'Updating' group in the empirical data. The model indeed inferred different latent causes for CS+ alone trials, CS+ with US trials, and CS- alone trials (Figure 5D). The slow updating resulted from the fact that the predictions of the US were based on the posterior over latent causes after CS onset, but before US observation. On CS+ trials, both CS+ only and CS+ and US latent causes were inferred, with their relative popularity so far determining their probability for the current trial. Thus the slow decrease in popularity of the CS+ and US latent cause led to slow extinction of the US expectation.

When we added a decay parameter $\tau = 0.15$ leading to fast forgetting of inferred latent causes, simulated behavior showed faster extinction than the 'Updating' regime, but increased predictions of outcomes after both the CS+ and the CS- observations in the retrieval phase (Figure 5 B). Indeed, predictions for the two stimuli were not differentiated at

all until the CS+ was paired again with the US in the relearning phase. This pattern of results closely matched the human data of the 'Forgetting' group. Inspection of the latent-cause structure inferred by this model shows that its behavior during the retrieval and the relearning phases resulted from completely new latent causes, as the originally learned latent cause assignments were forgotten during the second break (Figure 5E). The new latent causes were initialized at the prior, corresponding to p(scream)=0.5 for both the CS+ and the CS- at the beginning of the retrieval phase.

Finally, we used the hierarchical Dirichlet process setting $\alpha = 5$ to allow creation of new contexts, the context persistence parameter η to 0.8 (relatively high persistence of context), the parameter for creation of new tables ν to 0.5, and γ for creating new dishes (new latent causes) to 1. This setup leads to a balance that allows the creation of new latent causes in new context, but also sharing of existing latent causes across context. The smaller ν reduces the probability of two tables with the same dish in the same restaurant. We set the Beta prior parameters for CSs to 10^{-7} (relatively deterministic latent causes), and the Beta prior parameters for the US to 0.005 and 10^{-7} (deterministic and biased to expect that the US is more likely to be present than not). This model showed the fastest extinction, with spontaneous recovery of the US expectation only for the CS+ during the retrieval phase, and fast relearning in the last phase (Figure 5C). This pattern is reminiscent of human behavior in the 'Remembering' group. The striking difference between the 'Updating' and 'Remembering' regimes in the simulations emerged even though both simulations inferred the same underlying latent-cause structure (Figures 5D,F). The difference in behavior resulted from the additional inferred latent contexts in the 'Remembering' regime (Figure 5G). Different from the 'Updating' regime, where only one context was possible, in the 'Remembering' regime, later trials in the acquisition phase were assigned primarily to latent context 3 (which contained three latent causes, one for the CS+ alone trials, one for trials with CS+ and US, and one for CS- trials; Figure 5I). This latent context was not updated in the extinction phase, where context 1 (containing only latent causes for the CS+ alone and the CS- alone) was inferred (Figure 5G). Notably, after the long break (which reset the persistence of contexts 1-3), on the first CS+ trial of the retrieval phase, both contexts 1 and 3 were inferred (Figure 5H , red circle), leading to spontaneous recovery of the expectation of the US. After observing the absence of the US on that trial, the final inference was of context 1, which, after a few trials, dominated the inference even at stimulus onset due to persistence, leading to rapid decrease of the expectation the US. The latent cause for CS- trials, similar and shared in all contexts, never generated expectations of the US regardless of the inferred context. Further exploration of the model's behavior around these parameter regimes can be found in the supplementary material.

4 Discussion

Humans and animals often show a range of inductive biases for continual learning, but these biases may turn maladaptive in adverse environments. Here, we presented and modeled behavioral data from a human fear conditioning task that revealed three learning regimes – updating, forgetting and remembering of old knowledge. The latter regime, while desirable in many continual learning settings, is maladaptive when it leads to the spontaneous recovery of outdated fear associations. We modeled these data using Bayesian nonparametric models of "contextual" and "latent cause" inference, a framework that has successfully captured human behaviour in this and related domains using a common probabilistic vocabulary. We showed that inferring a stationary or non-stationary distribution of latent causes within a single context gives rise to "updating" or "forgetting" behaviour respectively, while inferring changing, persistent and recurrent contexts with shared latent causes gives rise to "remembering" of old knowledge, and consequently yields the risk of spontaneous recovery of fear due to its inappropriate reuse.

We also showed that prior beliefs about the structure of the world – namely persistent priors over contexts and deterministic, biased priors over negative events – encourage the creation of new "safe" latent causes in the absence of negative events, instead of updating old "dangerous" ones to no longer predict danger. These priors map onto several well known "core beliefs" and cognitive distortions that are thought to have clinical implications (Piray & Daw, 2021). For instance, deterministic and biased priors over negative events may reflect "black and white thinking" and pessimistic beliefs about how dangerous the world is. Similarly, priors about the persistence of contexts may be related to beliefs about the volatility of the environment.

Such beliefs in patients could potentially reflect past adverse experiences, and may impact individuals' tendencies to develop anxious beliefs and/or resist updating their beliefs during psychotherapy. These beliefs may offer tractable targets for quantifying individual differences and personalizing psychotherapy. Such a quantification could potentially predict who will show an enduring response to exposure therapy – patients who create fewer new latent causes in new contexts should have a lower risk of relapse after exposure therapy (see, e.g., spontaneous recovery after fear conditioning and extinction in Gershman & Hartley, 2015). The enhanced understanding of learning processes in exposure therapy that we gain through the model could also help to build more tailored interventions for individual patients.

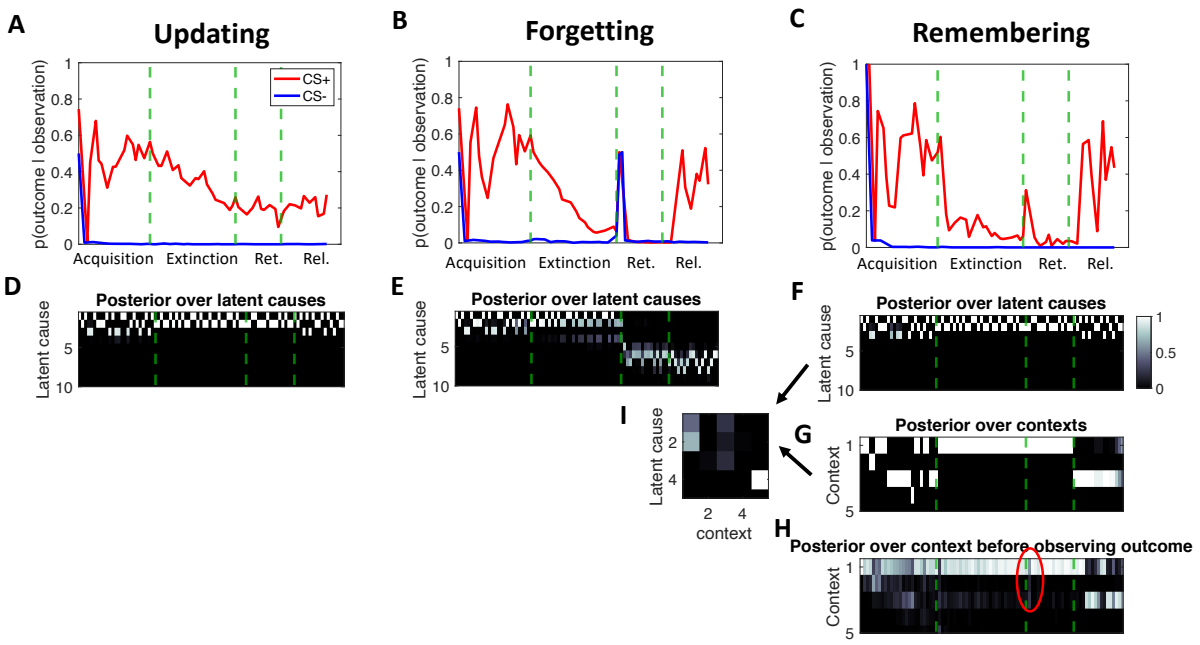

Figure 5: **Simulations of behavior for the 'Updating', 'Forgetting', and 'Remembering' regimes. A.** Data from the 'Updating' regime were simulated with a standard CRP, and deterministic priors over CSs and the US (see text). **B.** Data from the 'Forgetting' regime were simulated using the same CRP as in A, with decay of inferred latent causes on each trial. **C.** Data from the 'Remembering' regime were simulated using a hierarchical Dirichlet process with persistence at the context level, deterministic priors over CSs, and deterministic and biased priors over USs. A-C show simulated predictions of USs after observations of CS+ or CS- (similar to expectancy ratings in human behavior) for all four phases of the task. **D-F.** Posterior distribution over latent causes for the four phases, as inferred at the end of the experiment (i.e., given all the observations so far). Each column denotes a trial, lighter gray denotes higher posterior probability. **G.** Posterior distribution over contexts at the end of the task. **H** Posterior distributions over contexts as calculated on each trial when making a US prediction (that is, before observing whether a US did or did not appear on that trial, and given only previous observations). **I.** Latent-cause assignments per context for the 'Remembering' regime. Shaded boxes: lighter shading = higher probability. Ret. = Retrieval, Rel. = Relearning. Vertical dashed green lines in all plots indicate phase boundaries. Note that in D-H additional latent causes and contexts were inferred in the breaks to account for the "dummy" CS presented then (not shown; context 5 with latent cause 4 in I).

4.1 Limitations

One of the major limitations of our work so far is our reliance on qualitative fits between model simulations and behavioral data using hand-picked parameters. In future work, we plan to directly fit the model parameters to human data. Some key factors that limit the interpretation of our results are differences in the details of the behavioral task's design between humans and the model. These include the fact that humans rate the outcome expectancy only every few trials, the impoverished representation of the break in the model's inputs, and the fact that humans can potentially simulate additional experience during the break. As humans might have a negative sampling bias, such simulations in the break might change priors over outcomes in following blocks to be more negatively biased. These need to addressed in the future, and may even reveal features of interest that help narrow the differences between simulations and behavior.

4.2 Conclusion

We showed that human behaviour on an aversive conditioning task spans three regimes – updating, forgetting or remembering old knowledge. We accounted for these by assuming biased, deterministic and persistent priors in a hierarchical Bayesian nonparametric model. This fairly minimal, normative model of contextual learning accounts for sizeable variation in human behavior through variations in priors that can be mapped onto core beliefs and cognitive distortions. We are hopeful that this model will offer a useful window into adaptive and maladaptive human inductive

biases in continual learning settings, with potential utility for personalizing psychotherapy as well as human-level continual learning.

ACKNOWLEDGEMENT

SP and YN were supported by grant R01MH119511 from the National Institute for Mental Health. IMB's, JC's, YR's and YN's work on "Precision Psychiatry for treatment selection in depression" is supported by Wellcome Leap as part of the Multi-Channel Psych Program.

REFERENCES

Rivka T Cohen and Michael Jacob Kahana. A memory-based theory of emotional disorders. *Psychological Review*, 2022.

Samuel J Gershman and David M Blei. A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12, 2012.

Samuel J Gershman and Catherine A Hartley. Individual differences in learning predict the return of fear. *Learning & behavior*, 43(3):243–250, 2015.

Samuel J Gershman, Kenneth A Norman, and Yael Niv. Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, 5:43–50, 2015.

Noam Goldway, Eran Eldar, Gal Shoval, and Catherine A Hartley. Computational mechanisms of addiction and anxiety: a developmental perspective. *Biological Psychiatry*, 2023.

Raia Hadsell, Dushyant Rao, Andrei A Rusu, and Razvan Pascanu. Embracing change: Continual learning in deep neural networks. *Trends in cognitive sciences*, 24(12):1028–1040, 2020.

James B Heald, Máté Lengyel, and Daniel M Wolpert. Contextual inference underlies the learning of sensorimotor repertoires. *Nature*, 600(7889):489–493, 2021.

Dhireesha Kudithipudi, Mario Aguilar-Simon, Jonathan Babb, Maxim Bazhenov, Douglas Blackiston, Josh Bongard, Andrew P Brna, Suraj Chakravarthi Raja, Nick Cheney, Jeff Clune, et al. Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, 4(3):196–210, 2022.

Yael Niv, Peter Hitchcock, Isabel M Berwian, and Gila Schoen. Toward precision cognitive-behavioral therapy via reinforcement learning theory. *Precision Psychiatry: Using Neuroscience Insights to Inform Personally Tailored, Measurement-Based Care*, pp. 199, 2021.

Agnes Norbury, Hannah Brinkman, Mary Kowalchyk, Elisa Monti, Robert H Pietrzak, Daniela Schiller, and Adriana Feder. Latent cause inference during extinction learning in trauma-exposed individuals with and without ptsd. *Psychological Medicine*, pp. 1–12.

Payam Piray and Nathaniel D Daw. A model for learning based on the joint estimation of stochasticity and volatility. *Nature communications*, 12(1):1–16, 2021.

Sashank Pisupati and Yael Niv. The challenges of lifelong learning in biological and artificial systems. *Trends in cognitive sciences*, 2022.

Sashank Pisupati, Angela Langdon, and Yael Niv. Two factors underlying maladaptive inference of causal structure can drive resistance to extinction in anxiety. *Biological Psychiatry*, 89(9):S283, 2021.

Kirstin L Purves, Elena Constantinou, Thomas McGregor, Kathryn J Lester, Tom J Barry, Michael Treanor, Michael Sun, Jürgen Margraf, Michelle G Craske, Gerome Breen, et al. Validating the use of a smartphone app for remote administration of a fear conditioning paradigm. *Behaviour research and therapy*, 123:103475, 2019.

Eric Schulz and Peter Dayan. Computational psychiatry for computers. *Iscience*, 23(12):101772, 2020.

Mingyu Song, Carolyn E Jones, Marie-H Monfils, and Yael Niv. Explaining the effectiveness of fear extinction through latent-cause inference. *arXiv preprint arXiv:2205.04670*, 2022.

Ondrej Zika, Katja Wiech, Andrea Reinecke, Michael Browning, and Nicolas W Schuck. Trait anxiety is associated with hidden state inference during aversive reversal learning. *bioRxiv*, 2022.