2024

The Ubiquity of Time in Latent-Cause Inference

Dan-Mircea Mirea

Yeon Soon Shin

Sarah DuBrow

Yael Niv

The ubiquity of time in latent-cause inference

Dan-Mircea Mirea[1]*, Yeon Soon Shin[2]*, Sarah DuBrow[3], Yael Niv[1,4]
* both authors contributed equally
[1]Department of Psychology, Princeton University
[2]Department of Psychology, Bryn Mawr College
[3]Department of Psychology, University of Oregon
[4]Princeton Neuroscience Institute, Princeton University

Abstract

Humans have an outstanding ability to generalize from past experiences, which requires parsing continuously experienced events into discrete, coherent units, and relating them to similar past experiences. Time is a key element in this process; however, how temporal information is used in generalization remains unclear. Latent-cause inference provides a Bayesian framework for clustering experiences, by building a world model in which related experiences are generated by a shared cause. Here we examine how temporal information is used in latent-cause inference, using a novel task in which participants see 'microbe' stimuli and explicitly report the latent cause ('strain') they infer for each microbe. We show that humans incorporate time in their inference of latent causes, such that recently inferred latent causes are more likely to be inferred again. In particular, a 'persistent' model, in which the latent cause inferred for one observation has a fixed probability of continuing to cause the next observation, explains the data significantly better than two other time-sensitive models, although extensive individual differences exist. We show that our task and this model have good psychometric properties, highlighting their potential use for quantifying individual differences in computational psychiatry or in neuroimaging studies.

Introduction

The ability to generalize from relevant past experiences plays a crucial role in human learning and memory. Rather than learning information *de novo*, individuals often leverage previously learned knowledge in novel situations. The latent-cause inference framework provides a rational basis for such generalization, while also accommodating situations that are completely new and cannot build on past experience (Anderson, 1991; Courville et al., 2005; Gershman et al., 2010; Radulescu et al., 2021). In this nonparametric Bayesian framework, related experiences are clustered together as they are believed to share a common cause, while dissimilar experiences are segmented into distinct latent causes, with the overall number of latent causes unbounded (Franklin et al., 2020; Shin & DuBrow, 2021).

Organizing and segmenting experiences into coherent units is useful in making adaptive decisions that draw from past relevant experiences (Shadlen & Shohamy, 2016) and predictions about what will come next (Pettijohn & Radvansky, 2016; Rinck & Weber, 2003; Speer & Zacks, 2005; Zwaan, 1996). Work in the episodic memory literature has examined extensively how temporally continuous streams of experiences are segmented into discrete units, a process known as "event segmentation" (Clewett et al., 2019). This literature, including Sarah DuBrow's early work, shows how detecting and transitioning between events influences the encoding and retrieval of memories (DuBrow & Davachi, 2013, 2014, 2016; Ezzyat & Davachi, 2011; Heusser

et al., 2018; Radvansky & Zacks, 2011; Rouhani et al., 2020). Signaled temporal gaps (e.g., “a moment” or “a while” within a text) predicts event boundaries (Ezzyat & Davachi, 2011), and temporal information is better preserved within an event than across event boundaries (DuBrow & Davachi, 2013, 2014, 2016). Neuroimaging studies show that event structure is represented at different timescales along the cortical hierarchy, with the hippocampus responding to the boundaries of events (Baldassano et al., 2017; Ben-Yakov & Henson, 2018; Lee & Chen, 2022; Ritchey et al., 2015).

In organizing experiences, a key element is time, which weaves experiences into a stable continuum (Howard & Kahana, 2002; Jayakumar et al., 2023; Yu et al., 2021), and is a ubiquitous generalization clue – unless an event transition has occurred, the current experience is likely to be similar to the recent past. While the literature provides ample evidence for the significance of time in memory, the mechanisms by which individuals use time and similarity to recognize past events that are relevant to current experiences are less well understood.

Latent-cause inference offers a statistically principled (i.e., Bayesian) way to optimally generalize from past experience by inferring shared latent (hidden) causes for similar observations. By inferring the latent cause of the current observation, one can draw on knowledge from previous events that were presumably generated by the same latent cause. However, the standard Bayesian model of latent-cause inference, which relies on a Chinese Restaurant Process (CRP) prior, does not use temporal information (Aldous, 1985; Anderson, 1991). In this model, the prior probability of a latent cause is determined by its previous ‘popularity’ – the assumption is that a more prolific cause, i.e. one that has generated observations more often, is more likely to cause the next experience. This form of prior belief intentionally ignores the order of previous events (that is, the model is time-invariant by design), as this makes the model computationally tractable. However, the time-invariance assumption does not accord with causal structures in the world, which are most often temporally contiguous.

Recent studies have begun to investigate the role of temporal information in inferring latent causes from an ongoing stream of information, incorporating the assumption that recently encountered latent causes have a higher chance of generating the next observation (Blei & Frazier, 2011; Fox et al., 2011; Lloyd & Leslie, 2013). Indeed, empirical evidence suggests that learning behaviors in humans (Éltető et al., 2022) and rodents (Lloyd & Leslie, 2013; Song et al., 2022) are best captured by models in which the prior probability of a previously active latent cause decays over time (Blei & Frazier, 2011). A simpler model, which gives an extra boost to the most recently inferred latent cause (Fox et al., 2011), has been shown to account for stable event perception (Franklin et al., 2020; Gershman et al., 2014). Similarly, a model that assumes that temporal contexts persist over time captured learning behaviors in rodents (Lloyd & Leslie, 2013). This previous work suggests that temporal information is critical to latent cause inference.However, the evidence is indirect, through learning tasks that do not specifically examine latent-cause inference.

Here, we directly probe how temporal information contributes to generalization behavior in humans, using a novel task developed by Sarah DuBrow and Yeon Soon Shin to study latent-cause inference. In this “microbes task,” participants assign a stream of abstract visual stimuli, presented as “microbes,” to underlying similarity groups (“strains”), thereby explicitly probing the inference of latent causes from observations. Different from a categorization task,

the number of strains is not determined in advance, and the temporal sequence of stimuli is informative: consecutive stimuli tend to come from the same strain, though occasionally a microbe from an old strain will appear in the sequence, and a new strain can emerge at any moment. Using behavioral data from a large sample, we compare different latent-cause inference models: the standard Chinese Restaurant Process prior that does not rely on temporal information (Aldous, 1985; Anderson, 1991), and the above-mentioned three models that integrate temporal information in the inference process. We then test the reliability of model parameters to further explore the possibility of utilizing the inference models to capture individual differences that can map onto psychological constructs, such as *mental health conditions*.

Methods

Participants. Participants (*N* = 1928) were recruited on Prolific Academic in five waves between March 2021 and May 2022 for a larger study investigating how latent-cause inference parameters map onto individual differences in mental-health symptoms. This study was approved by the Institutional Review Board of Princeton University (protocol #11968), and all participants provided informed consent. All participants were in the United States, with a mean age of 35.0 years (*SD* = 13.3, range 18-84). 927 participants identified as women (56.4%), 668 identified as men (40.7%), and 48 identified as non-binary (2.9%). 285 participants did not provide demographic information, whereas 3 participants were likely dishonest in their age reports (reporting ages of 3, 7 and 332 years old, respectively). Our larger experiment included 192 self-report items measuring psychiatric symptoms that were subject to exploratory factor analysis to reduce dimensionality and multicollinearity. Based on previous work with a similar-sized questionnaire battery (Gillan et al., 2016), we estimated we needed high-quality psychiatric data from at least 1400 participants. We continued recruitment in waves until this criterion was met and slightly surpassed (total *N* who completed and passed attention checks on questionnaires = 1637).

The experiment involved two sessions. Participants were excluded from the study if they did not complete both sessions, or if they failed more than two out of 10 attention checks embedded in the self-report symptom questionnaires in the first session (e.g., failing to answer ‘I disagree a lot’ or ‘I disagree a little’ to the item ‘When something good happens, it makes me think about all the times I traveled to the moon’). Participants were also excluded if they made more than 10 errors in total on any of the forced-choice training trials, or if they failed to respond on more than 4 trials in the training block, 5 trials in the practice block, or 10 trials on the 2 main blocks in session 1 (see below for details). These values were determined based on the empirical distribution of missed trials for each block. Excluded participants (*N* = 691) were not invited to the second session, and their data were not analyzed further. The remaining participants (*N* = 1237) were invited back 1-2 days later, and recruitment was kept open for ~2 weeks with regular reminders for those who did not respond right away (maximum delay between sessions = 18 days; *N* = 1073 returned for the second session). The second session began with a series of comprehension questions regarding the task instructions. If a participant responded incorrectly to any of the questions, they were asked to try again. If they still had a mistake on their third try,

their session was terminated and they were excluded from the experiment (these participants were counted as “not returning for session 2”). In addition, participants who failed to complete all portions of the microbes task in session 2 (*N* = 91) or who missed more than 5 trials in more than one of the main blocks (*N* =10) were excluded from further analysis.

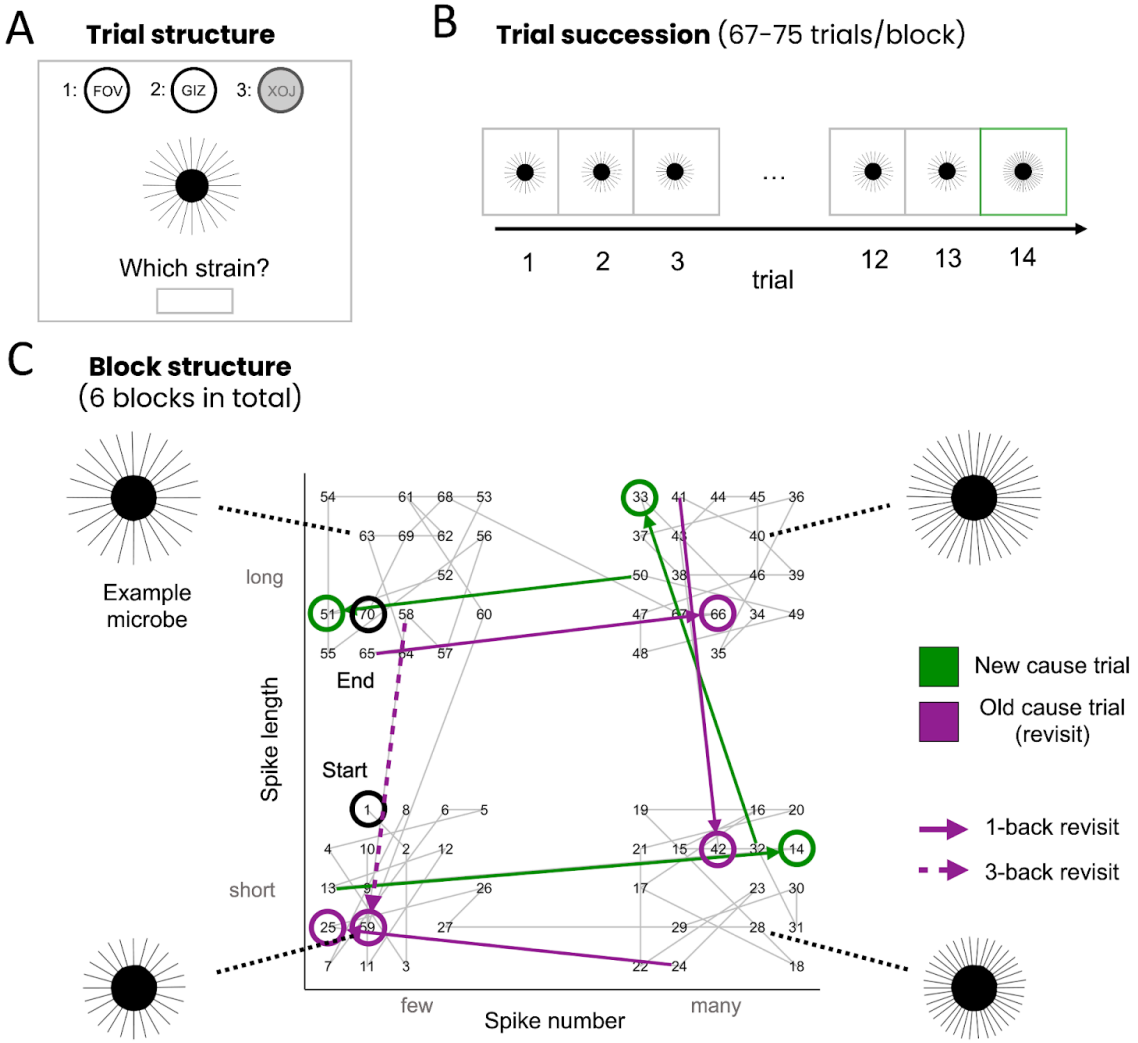

Figure 1. Structure of the ‘microbes’ task. *A*. Schematic of the trial structure. On each trial, participants saw a microbe and had to type in a number indicating the strain they thought the microbe belonged to. Each strain was denoted by a 3-letter label. The grayed out strain is the one that has not been inferred yet (a “novel” strain). ***B*.** Schematic of a temporal sequence of trials. ***C*.** Schematic of the structure of a single block. Numbers are the same as in *B*, denoting the sequence of trials and corresponding microbes. Axes denote the spike number and length of each microbe in the sequence (in subjectively equidistant steps, based on a distance-norming experiment not reported here). Four example microbes are shown linked with dotted black lines to their corresponding number in the sequence. Each block consisted of four different ground-truth latent causes (‘strains’). Microbes on most consecutive trials came from the same latent cause, but there would occasionally be a jump to a new latent cause (green arrows and

circles), or a jump back to a previous latent cause, called ‘revisits’ (purple arrows and circles). Revisits could be 1-back (full arrow in the schematic), 2-back (not present in this block) or 3-back (dashed arrow), and there were always one or two revisits from each ground-truth latent cause except the first one. The specific block shown corresponds to the second main block in session 2.

This left a final *N* = 972 participants (mean age 35.6 years (*SD* = 13.67; range 18-84); 553 women (56.9%), 390 men (40.1%), 28 non-binary people (2.9%); one participant did not provide demographic information) whose data were analyzed and are reported here.

Experimental design. The experiment was run in two sessions. In the first session, participants completed four blocks of the microbes task (a training block, a practice block, and two standard blocks) alongside a 192-item battery designed to assess self-reported mental health symptoms (which will not be reported or analyzed here) and a demographic questionnaire. The training involved 10 forced-choice trials, each of which was repeated until the participant made the correct response. The practice block involved 66 trials of the task, with feedback denoting the correct response after each trial. The main blocks were similar in structure to the practice block except that participants were not given any feedback. The second session began with a series of comprehension checks about the task instructions. It then included the same training and practice blocks, together with 6 standard blocks of the microbes task. At the end, participants completed a visuo-spatial working memory task (the symmetry span task; Kane et al., 2004).

The ‘microbes task.’ Participants completed a task in which they assigned abstract visual stimuli to either old or new latent causes (Figure 1). The stimuli, introduced in the cover story as ‘microbes,’ had spikes coming out of a core. These spikes varied along two dimensions: the number of spikes (dimension 1) and the length of the spikes (dimension 2). Participants were asked to classify the microbes into ‘strains’ (latent causes) based on their perceptual similarity. They were told that the stimuli were photos of microbes taken at consecutive time points and that, at any given time, one microbe strain is dominant; however, microbes mutate sometimes to generate a new strain, which quickly starts to dominate but does not take over completely, so they could still sometimes see exemplars of old strains.

In both experimental sessions, the task started with a short 10-trial forced-choice training phase in which microbes were presented with their correct strain, and a practice block where participants received feedback (i.e. were told which strain was correct) after choosing as strain for each of 66 microbes. Thereafter, in the main phase (2 blocks in session 1 and 6 blocks in session 2), no feedback was provided. Only behavior in the main blocks was analyzed and used for modeling. Each main block had between 67 and 75 trials with trial-unique microbes coming from 4 ground-truth latent causes, corresponding to four corners of the 2-dimensional feature space (Figure 1C). The exact arrangement of latent causes and stimulus sequence differed between blocks. Microbes on successive trials generally belonged to the same cause, with the exception of 3 new-cause jumps and 3-5 old-cause revisits per block (1-2 from each cause). Revisit trials were generated so that, across blocks, revisits from each cause to all previously dominant causes were observed. Each revisit trial could be followed by 1-2 additional trials in the revisited cause, and was then necessarily followed by a ‘post-revisit’ trial back to the cause

that the revisit was performed from. Notably, the microbes presented in the revisit trials were novel, while sharing features with other microbes in the revisited cause. Participants were not informed of this underlying structure, and could classify stimuli into as many causes per block as they wished. Each strain was denoted by a randomly generated three-letter “label” (e.g., BAF); to make these pronounceable by the participants, each label had a consonant-vowel-consonant structure. Participants could choose, on each trial, any of the previously inferred strains or a new strain by typing in the number associated with the label. All previous strains, their labels and associated numbers were listed on the top part of the experiment screen. Participants had 5s to write their response; if no response was entered, the trial would end and the next one would start. No feedback was given on their choices (participants received feedback on the number of missed trials at the end of each block).

Computational models. We modeled task behavior using a Bayesian model of latent-cause inference with a prior over latent causes and a likelihood for each latent cause. In all models, the prior was a variation of the Chinese Restaurant Process (CRP; also called an infinite capacity mixture model or a Dirichlet Process Mixture; Li et al., 2019). In the simplest model, which we termed ‘standard CRP’, the prior probability of the next observation coming from each of the existing latent causes (corresponding to strains in the task) was proportional to the number of observations already assigned to those causes (Equation 1). The probability of a new latent cause was proportional to a concentration (or ‘new latent cause’) parameter α:

$$p(c_t = k) = \begin{cases} \frac{N_k}{\sum_{j=1}^{K} N_j + \alpha}, & k \leq K \\ \\ \frac{\alpha}{\sum_{j=1}^{K} N_j + \alpha}, & k = K + 1 \end{cases} \quad (1)$$

where c_t is the latent cause (strain) of the observation on trial t, K is the number of latent causes (strains) inferred by the participant so far (at most, one per previous observation; note that we take the participants’ responses as proxies for inferred latent causes), and N_k is the number of times latent cause k has previously been inferred. As such, in this model, a more prolific latent cause (i.e. one that has generated more observations) is more likely *a priori* to cause the next observation.

The second model, which we call the ‘decay model’ (Blei & Frazier, 2011), used the standard CRP (Equation 1) with the modification that the counts N_k were decayed exponentially on each trial, with the rate of decay governed by a parameter λ:

$$N_{k,t} = exp(-\lambda) N_{k,t-1} \quad (2)$$

In this model, therefore, a more *recently* prolific latent cause is more likely to cause the next observation. The normalization factor in the denominator of the CRP was adjusted accordingly so the probabilities of inferring each latent cause summed to 1.

In the third model, which we term the 'sticky model' (Fox et al., 2011), the observation count for the most recently inferred latent cause (c_{t-1}) was 'boosted' by a 'sticky' parameter β:

$$p(c_t = k) = \begin{cases} \frac{N_k + \beta}{\sum_{j=1}^{K} N_j + \alpha + \beta}, & k = c_{t-1} \\ \frac{N_k}{\sum_{j=1}^{K} N_j + \alpha + \beta}, & k \neq c_{t-1}, k \leq K \\ \frac{\alpha}{\sum_{j=1}^{K} N_j + \alpha + \beta}, & k = K + 1 \end{cases} \quad (3)$$

In the fourth and final model, which we call the 'persistent model' (very similar to Lloyd & Leslie, 2013), the latent cause inferred on one trial had a fixed 'persistence probability' η of continuing to cause the next trial, with the remainder of the probability mass distributed over all latent causes according to the standard CRP:

$$p(c_t = k) = \begin{cases} \eta + (1 - \eta) \frac{N_k}{\sum_{j=1}^{K} N_j + \alpha}, & k = c_{t-1} \\ (1 - \eta) \frac{N_k}{\sum_{j=1}^{K} N_j + \alpha}, & k \neq c_{t-1}, k \leq K \\ (1 - \eta) \frac{\alpha}{\sum_{j=1}^{K} N_j + \alpha}, & k = K + 1 \end{cases} \quad (4)$$

The likelihood for the observation (microbe) to belong to each latent cause (strain) was computed as a product of the likelihoods for each dimension (length and number of spikes), assuming a Gaussian similarity function in each dimension. The mean of each Gaussian was set to the average of previous observations the participant had assigned to that strain, with a fixed variance (representing how variable/wide a strain is assumed to be) set as a 'size parameter' σ:

$$p(s_t|c_t = k) \propto exp(\frac{-(s_{m,t} - \mu_{m,k})^2}{2\sigma^2}) \times exp(\frac{-(s_{n,t} - \mu_{n,k})^2}{2\sigma^2}) \quad (5)$$

where s_t is the stimulus on trial t with stimulus values $s_{m,t}$ and $s_{n,t}$ along the two dimensions (m = length of spikes, n = number of spikes), and $\mu_{m,k}$ and $\mu_{n,k}$ are the average of previous stimulus values for latent cause k along each of the two dimensions.

Model fitting and analyses. The models were fit to the data with a Bayesian approach using the Stan programming language (Stan Development Team, 2021), compiled using the *cmdstanpy* package (version 1.0.0) in Python (version 3.9.7). Participant-level parameters were assumed to be drawn from a either a beta distribution (the η parameter) or a gamma distribution (the remaining parameters). Each model was first fit using a hierarchical (pooled) approach on a random subset of 200 participants to derive the hyperparameters of these distributions for the entire sample. The median of the posterior distribution was taken as the best fit parameter value. These fit hyperparameters were then used to estimate individual parameters for each participant using only that participant's data. The log likelihood of each participants' data was then computed in Python using the best fit parameters and used to calculate the Bayesian Information Criterion (BIC; Neath & Cavanaugh, 2012) for each model for each participant.

To ensure the fidelity of model fits at the individual level, the psychometric properties of the winning model were assessed. Split-half reliability was computed as a measure of internal consistency, by correlating parameters separately fit to data from the even blocks (blocks 2, 4, 6) and the odd blocks (blocks 1, 3, 5) of the task. Test-retest reliability was computed by correlating parameters fit to data from the 2 blocks from the first session with parameters fit to data from the first 2 blocks from the second session. Finally, parameter recovery was performed by simulating task responses under the winning model. For each set of individual parameters (corresponding to one participant), responses on a full run of the task were generated. The simulated data were then fit per simulated participant with the same hierarchical hyperparameters. These recovered parameter values were then correlated with the original values used to generate the data ("ground truth").

Statistical analyses. All analyses were performed in R (version 4.3.1) written in Jupyter Notebook (version 6.4.12). In bar plots, data are shown as medians +/- the standard error of the median (SE_{Median}). This was computed by multiplying the standard error of the mean (standard deviation divided by the square root of the sample size) by 1.2533 ($\sqrt{\pi/2}$), an analytically derived constant used to convert the standard error of the mean to the standard error of the median (Williams, 2001). Differences between medians were tested using paired Wilcoxon signed-rank tests, with rank-biserial correlations as measures of effect sizes. Relationships between proportions of responses and the number of the ground-truth cause (Figure 2C,D) were tested using linear mixed-effects models with random intercepts for participants. Parameter reliability and recoverability were computed as Spearman's rank correlation coefficient.

Results

Analysis of task behavior. We divided participant responses into three types, corresponding to the three trial types in the task: ‘stay’ (assigning the observation the same latent cause (strain) as on the previous trial, i.e. staying in the same cause), ‘revisit’ (inferring a jump back to an old latent cause), and ‘new’ (starting a new latent cause). In general, participants behaved congruently to the trial type (Figure 2A), staying in the same latent cause on stay trials (median count 312.5/361, $SE_{\text{Median}} = 2.2$), revisiting an old cause on revisit trials (median count 34/50, $SE_{\text{Median}} = 0.42$), and starting a new cause on new trials (median count 12/18, $SE_{\text{Median}} = 0.12$). This indicates that participants largely learned the task well and were sensitive to the features of the particular stimulus on each trial, suggesting the likelihood component of our models is likely to play an important role.

To assess whether the CRP prior is playing a role as well, we looked at predictions made by the CRP prior. First, when choosing which latent cause to assign the microbe to, causes with more previous observations (i.e. more prolific causes) should be more likely to be chosen. Indeed, on average, when choosing existing causes (‘stay’ or ‘revisit’ responses), chosen causes had significantly more prior observations than non-chosen causes (median of the per-participant mean observation counts for chosen causes = 8.39; median of the per-participant mean observation counts for other causes = 6.68; $p < 0.001$, paired Wilcoxon signed-rank test; rank-biserial correlation (an effect size measure for this test) $r = 0.94$; Figure 2B). Second, the CRP prior has a decreased tendency to start a new latent cause as the task progresses. This is because as observation counts for existing causes increase, the denominator in equation 1 (the normalization factor) increases, and the probability of a new latent cause becomes increasingly small. Note that the ground truth generative model of microbes in the task did not have this property – a new strain started every 12-22 trials. Nevertheless, the model predicts that the probability of creating a new cause on true new-cause trials should decrease throughout a block. As predicted, participants were progressively less likely to create a new cause between the first and third new-cause trials in a block (median proportion new-cause responses = 0.83 in ground-truth latent cause 1; 0.66 in cause 2; 0.50 in cause 3; linear mixed-effects model with random intercepts: $\beta = -0.16$, $SE = 0.0045$, $p < 0.001$; Figure 2C).

Finally, we were interested to see whether there was an effect of time on participant’s choices (e.g., in alignment with the decay model). In our task, when the participant had been seeing the fourth ground-truth cause (strain), a revisit trial could jump back 1, 2 or 3 latent causes to the third, second, and first ground-truth cause, respectively. Importantly, the 1-back and 3-back jumps are spatially matched, differing from the current cause only across one dimension (Figure 1). This design allows us to compare the choices between these two options. For this, we computed the percentage of correct responses on revisit trials for the fourth latent cause (i.e. percentage of 1-back responses on 1-back revisit trials etc.). As predicted by the decay model, the percentage of correct responses decreased as the revisit was farther in time (mean proportion matched responses = 0.26 for 1-back revisits; 0.18 for 2-back revisits; 0.14 for 3-back revisits (the median proportion for 2-back and 3-back revisits was 0 so we chose to measure and visualize means here); linear mixed-effects model with random intercepts: $\beta = -0.063$, $SE = 0.0043$, $p < 0.001$; Figure 2D). This suggests that latent causes that were active in the more distant past were less likely to be reused.

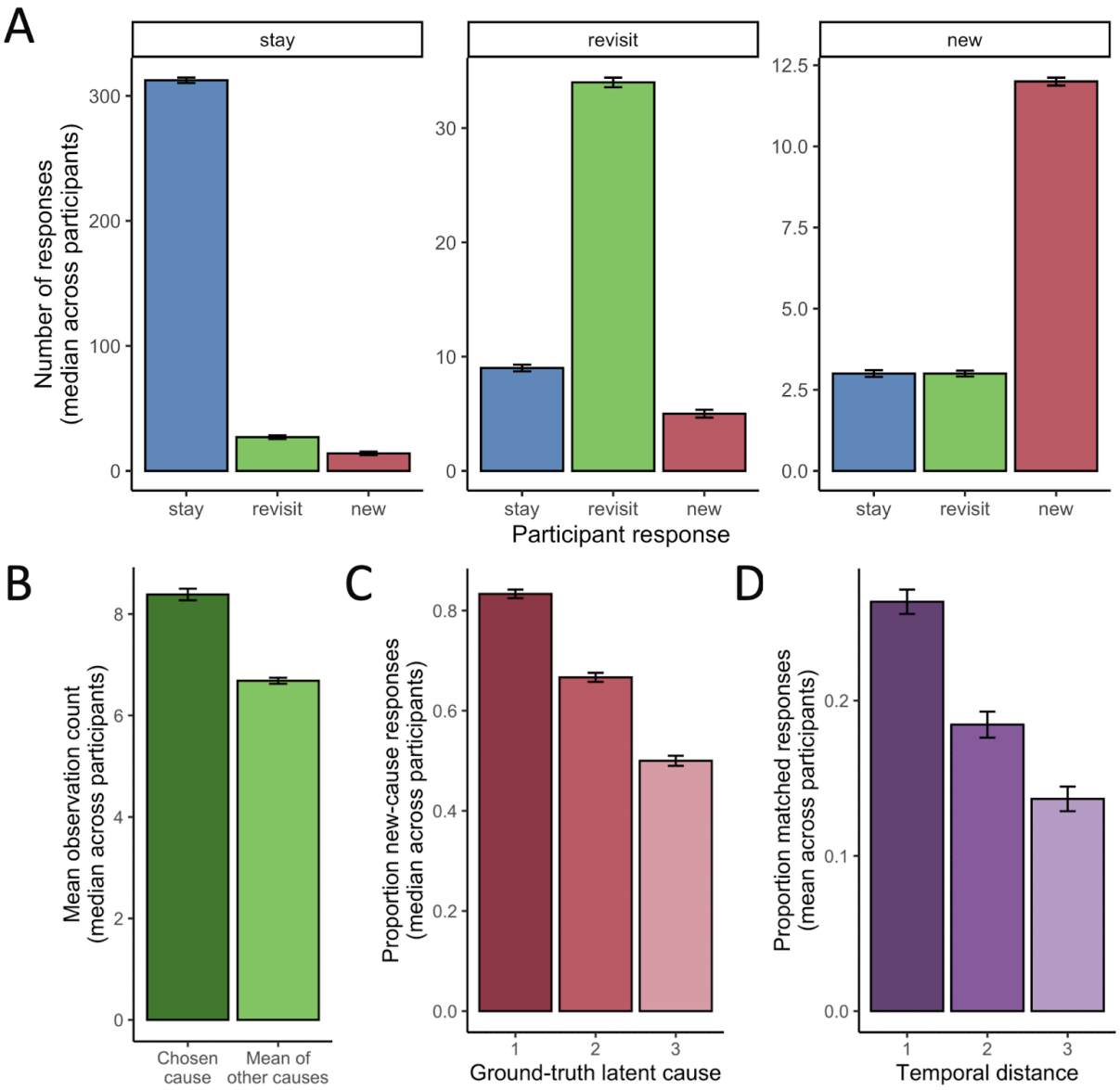

Figure 2. Task behavior is consistent with a time-sensitive CRP model. ***A.*** Number of responses of each type for each trial type. The three panels represent ground-truth trial types, whereas the x-axis and color show the participant response types. Note the differences in Y axis scaling. ***B.*** Previous observation count for chosen latent causes versus the mean of other (previously inferred) latent causes. For each participant we averaged all counts to one statistic. ***C.*** Proportion of 'new' responses across first, second and third ground-truth new-cause trials in a block. ***D.*** Proportion of responses to revisit trials in the fourth ground-truth cause that matched the temporal distance of the trial. Temporal distance refers to how many causes back the old cause is (i.e. 1 = 1-back). Plots show median values across participants and the error bars show the standard error of the median, with the exception of *D,* which shows mean values and standard errors of the mean.

Model comparison. Participants' raw behavior showed effects consistent with a time-sensitive CRP prior. To more formally test whether a time component improves the explanatory power of the latent-cause inference model, we performed statistical model comparison. We tested four different models - a 'standard' model with a standard CRP prior and three models with an added time component. These had either an exponential decay with time on the observations counts ('decay' model), an additive boost on the most recently inferred latent cause ('sticky' model), or a fixed probability of staying in the same latent cause and a remainder probability of drawing from the CRP ('persistent' model). For each participant, we computed the BIC score of each model (reflecting the likelihood of the participants' choices given the model, corrected for the number of free parameters in the model) and compared the distributions of the BIC scores.

We found that all time-sensitive models provided significantly better fits to the data (i.e., had lower BIC scores) compared to the standard model (median BIC for the standard model = 755.3; median BIC for the other models = 510.9-525.6; persistent vs. standard: $p < 0.001$, paired Wilcoxon signed-rank test (to account for heavy-tailed distributions), rank-biserial correlation (an effect size measure for this test) $r = -0.98$; decay vs. standard: $p < 0.001$, $r = -0.98$; sticky vs. standard: $p < 0.001$, $r = -1$; Figure 3A). Indeed, all participants but one were fit best by a time-sensitive model (Figure 3B). When comparing the three time-sensitive models, the persistent model fit the data better than both the sticky and the decay model, having significantly lower BIC values (persistent vs. sticky: $p < 0.001$, paired Wilcoxon signed-rank test, rank-biserial correlation $r = -0.62$; persistent vs. decay: $p < 0.001$, $r = -0.29$; Figure 3A), although the differences in the median BIC scores were small (persistent median BIC = 510.9, SE_{Median} = 13.2; sticky median BIC = 521.4, SE_{Median} = 11.9; decay median BIC = 525.6, SE_{Median} = 12.5). In fact, the persistent model was the winning model for only ~48% of the participants (465/972), indicating extensive individual differences (Figure 3B). To unpack these individual differences, we examined the difference in BIC scores between the persistent model and each of the decay and sticky models.

The distribution of BIC differences between the persistent and decay models was slightly more concentrated to the left of 0 but relatively wide (median difference = -25.5, SD = 135.6; Figure 3C), indicating that some participants were much better fit by the persistent model, whereas others were better fit by the decay model. In comparison, the BIC differences between the persistent and sticky models were smaller in magnitude and largely negative (median difference = -8.47, SD = 131.8; Figure 3D), indicating that for most participants, the persistent model was slightly better. This is not surprising given the similarity between the persistent and sticky models - both assign an especially high probability (a "boost") to the most recent latent cause. For both comparisons, however, there was a large positive tail indicating several participants with a much worse fit (higher BIC) for the persistent model. Finally, the BIC differences between the sticky and decay models had a negative median (median difference = -20.5, SD = 97.71; Figure 3E), suggesting more participants were fit better by the sticky model. However, a substantial proportion of the participants (393, or 40.4%) were better fit by the decay model, indicating large individual differences.

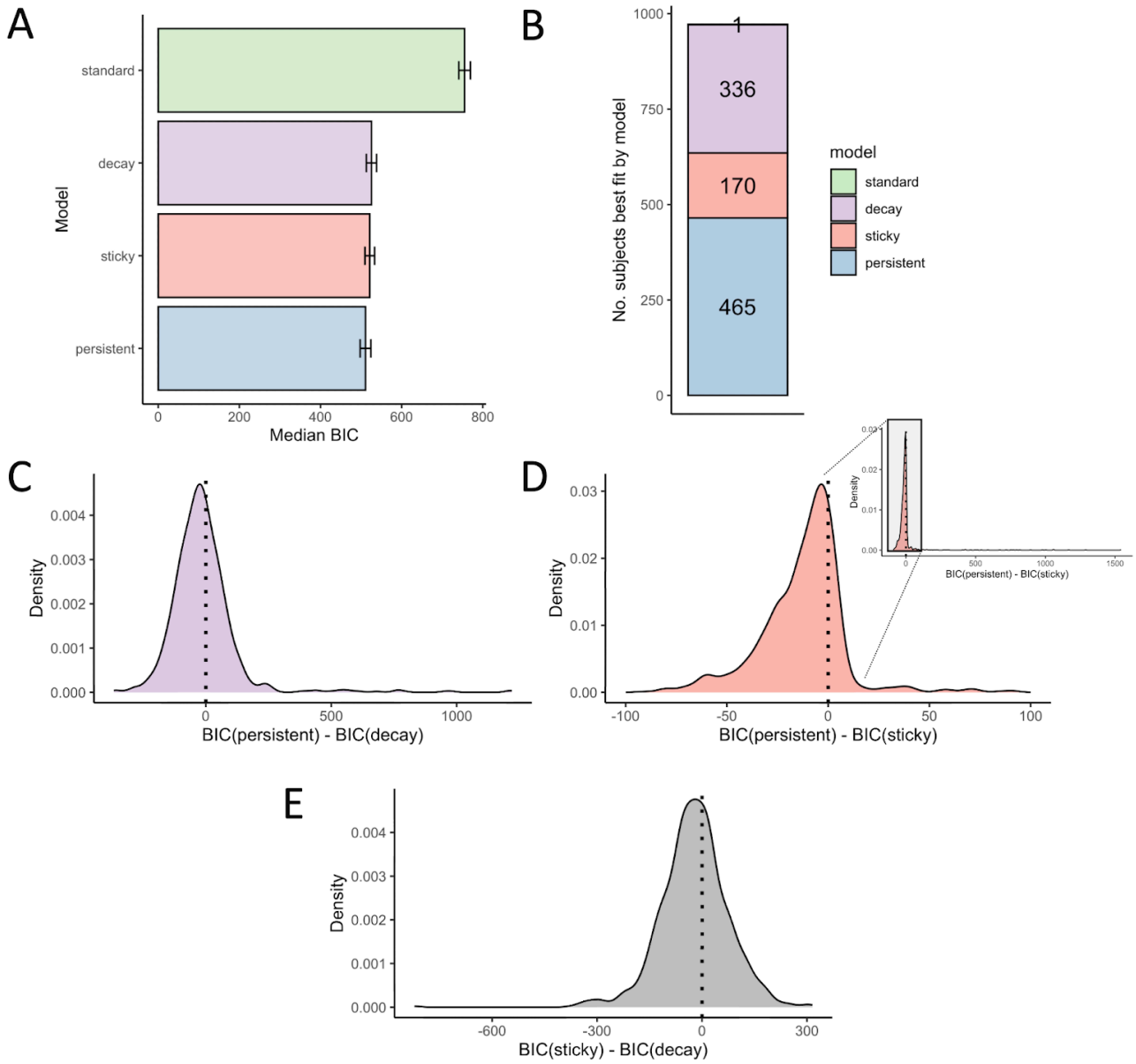

Figure 3. Model comparison shows time-sensitive CRP models provide a better fit for the data, with significant individual differences. ***A.*** Comparison of median BIC across participants for the four different models. Error bars show the standard error of the median. Lower BIC scores indicate a better fit of the data to the model predictions. ***B.*** Number of participants best fit by each of the models. Only one participant was fit best by the simple model that did not have a temporal component. ***C.*** Distribution of per-participant BIC differences between the persistent and decay models. ***D.*** Distribution of per-participant BIC differences between the persistent and sticky models. ***E.*** Distribution of per-participant BIC differences between the sticky and decay models. Dotted line: a difference of 0, corresponding to equal fit of both models. Density to the left of zero corresponds to the first model (persistent model in C and D, decay model in E) showing a better fit (lower BIC, hence negative difference). The top-right inset in *D* shows the full distribution (the main plot was truncated at a BIC difference of 100 for visualization purposes).

Reliability and recoverability of model parameters.

To verify the adequacy of using this task and the time-sensitive CRP models as individual-difference measures of latent-cause inference, we computed the split-half and test-retest reliability of each model's parameters. We did this first for the 'winning' persistent model. Split-half reliability was high, with Spearman coefficients equaling or exceeding 0.75 for all parameters (α: Spearman's ρ = 0.87, η: Spearman's ρ = 0.81, σ: Spearman's ρ = 0.75; Figure 4A). Test-retest reliability was lower, which can be expected given that the model was fit to only two blocks, and thus parameter fits were noisier. Nevertheless, reliability was larger than 0.5 for all parameters (α: Spearman's ρ = 0.66, η: Spearman's ρ = 0.56, σ: Spearman's ρ = 0.55; Figure 4B).

To ensure that the parameter estimates were stable and recoverable, we additionally performed a parameter recovery analysis. Here, we simulated task behavior with the parameter estimates for each participant, and then fit the persistent CRP model again to the simulated behavior to estimate the parameters and compare them to the known 'ground truth' parameters. The model had excellent parameter recoverability, with coefficients > 0.95 for all parameters (alpha ρ = 0.97, eta ρ = 0.96, sigma ρ = 0.96; Figure 4C).

We conducted similar reliability and recoverability analyses for the other two time-sensitive models – the decay and sticky models – as these nevertheless best explained the behavior of a minority of participants. We found that these two models had very similar split-half reliability, test-retest reliability and parameter recoverability to the persistent model (Table 1). In summary, all time-sensitive CRP models had adequate internal consistency, test-retest reliability, and parameter recoverability. This supports the use of these models and their parameters as individual-difference measures of the computational process of latent-cause inference.

Table 1. Reliability and recoverability of time-sensitive model parameters. Each value is a Spearman's correlation coefficient ρ.

Model	Parameter	Split-half reliability	Test-retest reliability	Parameter recoverability
Persistent model	Concentration α	0.87	0.66	0.97
	Persistence η	0.81	0.56	0.96
	Cluster size σ	0.75	0.55	0.96
Decay model	Concentration α	0.84	0.62	0.95
	Decay rate λ	0.88	0.63	0.98

	Cluster size σ	0.76	0.56	0.97
Sticky model	Concentration α	0.87	0.65	0.95
	Sticky boost β	0.83	0.56	0.95
	Cluster size σ	0.74	0.54	0.95

Discussion

Temporal information is crucial for forming stable representations of experiences. In this study, we investigated whether time matters in human latent-cause inference, using a novel task designed to quantitatively capture the inference process. Our findings show that humans are sensitive to temporal information when making inferences about the hidden causes that have generated individual experiences. Using non-parametric Bayesian models, we show that participants' inferences were best explained by models that incorporated time in estimating the probability of latent causes.

More specifically, our results show three types of behavioral profiles. The most prominent type persistently prioritizes the most recently inferred cause, while estimating the probabilities of other causes based on factors outside of time (i.e., the persistent model). The persistent model showed reliable estimates of model parameters across time and within sessions, suggesting that the model parameters can capture meaningful individual differences that can map onto psychological constructs (e.g., mental health symptoms). Additional behavioral patterns emerged in a subset of individuals: decaying the probability of the causes as a function of temporal recency (the decay model), and giving an advantage to the most recently inferred cause with a diminishing effect of the boost as the number of stimuli increases (the sticky model).

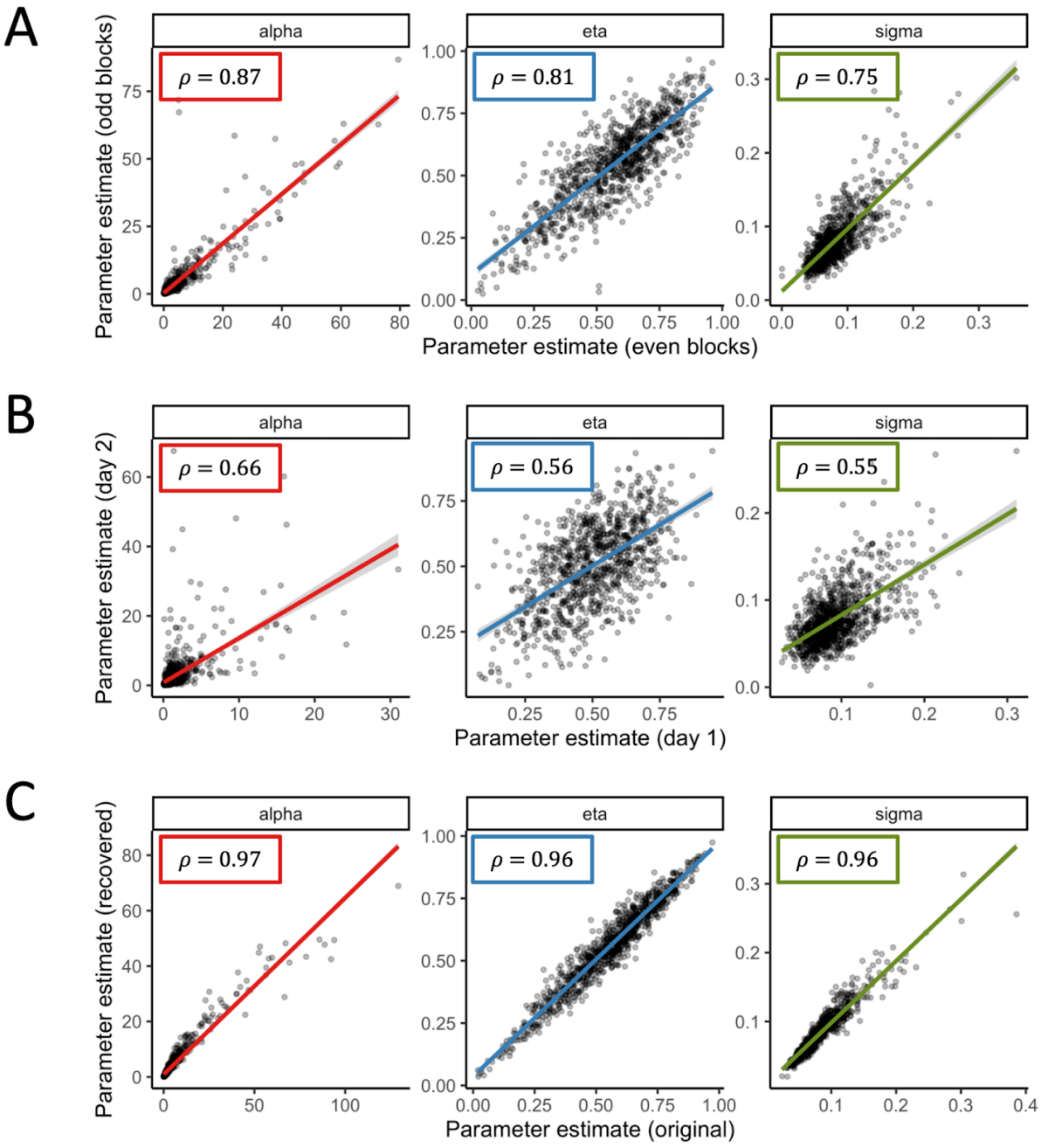

Figure 4. The persistent model, which best explained participants' choices, showed good split-half reliability, test-retest reliability, and parameter recovery. *A*. Split-half reliability analysis, showing correlations between parameters estimated separately from even and odd blocks of the task. ***B*.** Test-retest reliability analysis, showing correlations between parameters estimated on the two blocks in session 1 ('day 1') and parameters estimated on the first two blocks from session 2 ('day 2'). ***C*.** Parameter recovery analysis, showing correlations between fitted parameter values ('original') and values recovered from simulating task behavior with the fit parameters ('recovered'). Each parameter is shown on a column and with a corresponding color.

Our results are consistent with previous models of event segmentation where a new experience is monitored for its similarity to previous experiences on multiple dimensions, including time (Event-Indexing Model; (Zwaan et al., 1995)Zwaan et al., 1995). Similar to the “situations” inferred in event comprehension, our model infers the causal structure that generate experiences. Our results are also consistent with previous findings in causal inference that show temporal information plays a role in inferring causal structure (Éltető et al., 2022; Song et al., 2022) and forming memory traces accordingly (Franklin et al., 2020; Gershman et al., 2014). Building on these previous findings, we formally compared different ways by which temporal information is utilized in inferring latent causes. Of special interest were cases in which the likelihood for the most recently inferred cause is low (i.e., the observed stimulus is perceptually dissimilar from other stimuli generated by the current cause) – in these cases, how does time guide our inference of the latent cause?

One possibility is that the most recently inferred cause is prioritized in inference (higher prior probability), which enhances the stability of causal inference in light of noisy observations. This type of prioritization prevents frequent transitions between events – i.e., event boundaries, (Speer & Zacks, 2005) and “flushing” the memories from working memory (Swallow et al., 2009). In episodic memory, drawing event boundaries decreases the recall of memories that belong to previous events, even after re-visiting the same type of event (Radvansky et al., 2011). Our modeling results align with evidence from the episodic memory literature that demonstrates the importance of “staying” in the most recently inferred latent cause. Specifically, the model that persistently prioritizes the most recently inferred cause (i.e., persistent CRP prior) performed better than a model that diminished the recent-cause boost as the number of experiences increased (the sticky model; Fox et al., 2011). This suggests that the stability of causal inference is prioritized throughout the inference process.

A second possibility is that rather than treating all of the non-recent latent causes as equally “old”, the probability of each latent cause decreases with the number of observations since it was last active. In this case, more recently encountered latent causes would be more preferentially used to explain the current experience. Such a pattern would align with temporal contiguity effects where the association between two items decays with temporal distance (Boakes & Costa, 2014) and the probability of recalling an item from memory decays with temporal distance from the last item that was retrieved (Howard & Kahana, 2002). Indeed, this decay model provided the best explanation for approximately ⅓ of our participants. Notably, our model suggests that temporal decay is at the level of latent causes. That is, even though the “revisit” experiences were always perceptually novel, and thus there was no temporal recency for the exact stimulus, the old latent cause that generated other experiences that share features with the current stimulus was more likely to be used when the latent cause was more recently encountered. Interestingly, reaction times on our task showed that when individuals erroneously reported that a ‘revisit’ stimulus from an old cause comes from the current latent cause, they were slower to “stay” when the stimulus was in fact generated from a more recently active old cause. Using reaction times as a more sensitive measure of the inference process in future studies can perhaps help better differentiate individual differences in how people use past

information in their inference process. However, this would entail extending the latent cause models we used here to account for reaction times as a function of the inference process.

While our study investigates the significance of time in inference behavior, it is difficult to determine how latent causes are recognized from the past given this study alone. Neuroimaging studies can complement our modeling approach, addressing how the brain supports the inference of previously-encountered latent causes, and how it creates completely new latent causes when necessary. Sarah DuBrow originally developed the “microbes task” with an aim to examine the underlying neural processes of recognizing an old latent cause versus inferring a new one, going beyond detecting transitions and “event segmentation” and asking: how do I classify the new event when it is detected? For this, in a companion fMRI study now underway, we are using the task she originally designed, where “new” and “revisit” trials are matched for perceptual differences, to allow resolving the neural mechanisms of inference in these cases. In particular, Sarah DuBrow hypothesized that the orbito-frontal cortex (OFC), which has been shown to represent the posterior distribution of latent causes (Chan et al., 2016), might guide the hippocampus to either retrieve old latent causes or infer a new latent cause, while the hippocampus would in turn trigger updating the event representation in the OFC after an event concludes, similarly to what has been observed during learning in rodents (Guise & Shapiro, 2017; Srinivasan et al., 2023). If the posterior probabilities of existing latent causes are low at the onset of an event, akin to the prefrontal cortex triggering the hippocampal pattern separation at the start of learning (Guise & Shapiro, 2017), the prefrontal cortex may initiate the hippocampus to create a new memory trace, instead of updating an existing one. This will complement the event offset patterns, observed in the hippocampus in humans(Ben-Yakov & Henson, 2018; Lee & Chen, 2022) and rats (Srinivasan et al., 2023), which may update the posterior distribution over latent causes in the OFC.

We conceptualize latent cause inference as a process that is fundamental to generalization. Since no two experiences are exactly alike, learning relies on generalization, and as such, is intimately linked to memory processes as well. We thus hypothesize that alterations in latent cause inference can potentially lie at the heart of some mental health concerns, e.g., overgeneralization from past experiences in some anxiety disorders, and undergeneralization, or otherwise incoherent latent cause inference in schizophrenia and other psychotic disorders (Cisler et al., 2024). The model that best aligned with participants’ behavior – with the persistent CRP prior – had parameters that were internally consistent, reliable across days, and almost perfectly recoverable. These psychometric properties of the persistent model show that model parameters can reliably capture the process of inferring latent causes and suggest that the model parameters can be used as individual difference measures, to be correlated to different mental health symptoms.

In summary, the current findings suggest that the inference of latent causes relies on temporal information, by prioritizing the most recently inferred cause to explain novel experiences and decaying the probability of old causes over time. These results have implications for how humans organize memories and make decisions drawing from relevant past experiences, thereby providing fundamental insight into the ubiquitous impact of time information on learning and memory processes.

Data availability

All data as well as the modeling and analysis code used in this paper can be found on Github: https://github.com/danmirea/time-sensitive-LCI/.

Author contributions

Conceptualization: S.D., Y.S.S., Y.N.; Data Curation: D.M.M.; Formal Analysis: D.M.M.; Funding Acquisition: Y.N., S.D, Y.S.S.; Investigation: D.M.M., Y.S.S.; Methodology: Y.S.S., S.D., D.M.M., Y.N.; Project Administration: Y.N.; Software: D.M.M., Y.S.S.; Supervision: Y.N.; Validation: D.M.M.; Visualization: D.M.M.; Writing - Original Draft Preparation: D.M.M., Y.S.S.; Writing - Review & Editing: Y.N., D.M.M., Y.S.S.

Acknowledgements

This work is dedicated to Dr. Sarah DuBrow, who passed away in February 2022 and without whom none of it would have happened. For Yeon Soon Shin, Sarah was a science hero long before they met, and it was a great honor to be able to design this task with her, experiencing Sarah's eyes for meticulously controlled experiment designs Yeon Soon had admired so much. Yeon Soon still cannot pass by the conference room on the first floor of the Princeton Neuroscience Institute without thinking about those days when they were drawing up the microbes task on a board. Yael Niv and Sarah had worked together for several years to come up with ways to probe the internal, likely hierarchical structure of memories and how it interacts with new learning. From these weekly meetings—where they came up with an experimental design on one week, only for Sarah to come back the next week with a detailed explanation of why it would not work, or not be novel, or not answer the research question—that the microbes task was finally born. The intellectual depth of the discussions, and Sarah's persistence in search of a way to investigate the exact process she was interested in, were unique and cherished. Dan-Mircea Mirea only had the pleasure of working with Sarah on Zoom, during the pandemic, but even so he learned so much from Sarah and remembers those weekly meetings fondly. Sarah, thank you for being the most wonderful scientist and friend - we miss you dearly!

We would also like to thank Paul Kazelis for helping with collecting data for stimuli norming, and all the participants in our study for making this research possible. We are also grateful to Sashank Pisupati for introducing us to the persistent model, and to members of the Niv lab for helpful discussions.

Funding information

This research was supported by NIMH grant R21MH120798.

References

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de*

Saint-Flour XIII — 1983 (Vol. 1117, pp. 1–198). Springer, Berlin, Heidelberg.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409–429. https://doi.org/10.1037/0033-295X.98.3.409

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721. https://doi.org/10.1016/j.neuron.2017.06.041

Ben-Yakov, A., & Henson, R. N. A. (2018). The hippocampal film editor: Sensitivity and specificity to event boundaries in continuous experience. *The Journal of Neuroscience*, *38*(47), 10057–10068. https://doi.org/10.1523/JNEUROSCI.0524-18.2018

Blei, D. M., & Frazier, P. I. (2011). Distance dependent Chinese restaurant processes. *Journal of Machine Learning Research*, *12*, 2383–2410.

Boakes, R. A., & Costa, D. S. J. (2014). Temporal contiguity in associative learning: Interference and decay from an historical perspective. *Journal of Experimental Psychology. Animal Learning and Cognition*, *40*(4), 381–400. https://doi.org/10.1037/xan0000040

Chan, S. C. Y., Niv, Y., & Norman, K. A. (2016). A probability distribution over latent causes, in the orbitofrontal cortex. *The Journal of Neuroscience*, *36*(30), 7817–7828.

Cisler, J. M., Dunsmoor, J. E., Fonzo, G. A., & Nemeroff, C. B. (2024). Latent-state and model-based learning in PTSD. *Trends in Neurosciences*, *47*(2), 150–162. https://doi.org/10.1016/j.tins.2023.12.002

Clewett, D., DuBrow, S., & Davachi, L. (2019). Transcending time in the brain: How event memories are constructed from experience. *Hippocampus*, *29*(3), 162–183. https://doi.org/10.1002/hipo.23074

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2005). Similarity and discrimination in classical conditioning: A latent variable account. *Advances in Neural Information Processing Systems 17*, 313–320. http://papers.nips.cc/paper/2711-similarity-and-discrimination-in-classical-conditioning-a-l

atent-variable-account

DuBrow, S., & Davachi, L. (2013). The influence of context boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General*, *142*(4), 1277–1286. https://doi.org/10.1037/a0034024

DuBrow, S., & Davachi, L. (2014). Temporal memory is shaped by encoding stability and intervening item reactivation. *The Journal of Neuroscience*, *34*(42), 13998–14005. https://doi.org/10.1523/JNEUROSCI.2535-14.2014

DuBrow, S., & Davachi, L. (2016). Temporal binding within and across events. *Neurobiology of Learning and Memory*, *134*, 107–114. https://doi.org/10.1016/j.nlm.2016.07.011

Éltető, N., Nemeth, D., Janacsek, K., & Dayan, P. (2022). Tracking human skill learning with a hierarchical Bayesian sequence model. *PLOS Computational Biology*, *18*(11), e1009866. https://doi.org/10.1371/journal.pcbi.1009866

Ezzyat, Y., & Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological Science*, *22*(2), 243–252. https://doi.org/10.1177/0956797610393742

Fox, E. B., Sudderth, E. B., Jordan, M. I., & Willsky, A. S. (2011). A sticky HDP-HMM with application to speaker diarization. *The Annals of Applied Statistics*, *5*(2A), 1020–1056. https://doi.org/10.1214/10-AOAS395

Franklin, N. T., Norman, K. A., Ranganath, C., Zacks, J. M., & Gershman, S. J. (2020). Structured Event Memory: A neuro-symbolic model of event cognition. *Psychological Review*, *127*(3), 327–361. https://doi.org/10.1037/rev0000177

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*(1), 197–209. https://doi.org/10.1037/a0017808

Gershman, S. J., Radulescu, A., Norman, K. A., & Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLoS Computational Biology*, *10*(11), e1003939.

Gillan, C. M., Kosinski, M., Whelan, R., Phelps, E. A., & Daw, N. D. (2016). Characterizing a

psychiatric symptom dimension related to deficits in goal-directed control. *eLife*, *5*, e11305. https://doi.org/10.7554/eLife.11305

Guise, K. G., & Shapiro, M. L. (2017). Medial prefrontal cortex reduces memory interference by modifying hippocampal encoding. *Neuron*, *94*(1), 183-192.e8. https://doi.org/10.1016/j.neuron.2017.03.011

Heusser, A. C., Ezzyat, Y., Shiff, I., & Davachi, L. (2018). Perceptual boundaries cause mnemonic trade-offs between local boundary processing and across-trial associative binding. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(7), 1075–1090.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*(3), 269–299.

Jayakumar, M., Balusu, C., & Aly, M. (2023). Attentional fluctuations and the temporal organization of memory. *Cognition*, *235*, 105408. https://doi.org/10.1016/j.cognition.2023.105408

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology. General*, *133*(2), 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Lee, H., & Chen, J. (2022). A generalized cortical activity pattern at internally generated mental context boundaries during unguided narrative recall. *eLife*, *11*, e73693. https://doi.org/10.7554/eLife.73693

Li, Y., Schofield, E., & Gönen, M. (2019). A tutorial on Dirichlet process mixture modeling. *Journal of Mathematical Psychology*, *91*, 128–144. https://doi.org/10.1016/j.jmp.2019.04.004

Lloyd, K., & Leslie, D. S. (2013). Context-dependent decision-making: A simple Bayesian model. *Journal of The Royal Society Interface*, *10*(82), 20130069.

https://doi.org/10.1098/rsif.2013.0069

Neath, A. A., & Cavanaugh, J. E. (2012). The Bayesian information criterion: Background, derivation, and applications. *WIREs Computational Statistics*, *4*(2), 199–203. https://doi.org/10.1002/wics.199

Pettijohn, K. A., & Radvansky, G. A. (2016). Narrative event boundaries, reading times, and expectation. *Memory & Cognition*, *44*(7), 1064–1075. https://doi.org/10.3758/s13421-016-0619-6

Radulescu, A., Shin, Y. S., & Niv, Y. (2021). Human Representation Learning. *Annual Review of Neuroscience*, *44*(1), 253–273. https://doi.org/10.1146/annurev-neuro-092920-120559

Radvansky, G. A., Krawietz, S. A., & Tamplin, A. K. (2011). Walking through doorways causes forgetting: Further explorations. *Quarterly Journal of Experimental Psychology*, *64*(8), 1632–1645. https://doi.org/10.1080/17470218.2011.571267

Radvansky, G. A., & Zacks, J. M. (2011). Event perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(6), 608–620.

Rinck, M., & Weber, U. (2003). Who when where: An experimental test of the event-indexing model. *Memory & Cognition*, *31*(8), 1284–1292.

Ritchey, M., Libby, L. A., & Ranganath, C. (2015). Cortico-hippocampal systems involved in memory and cognition: The PMAT framework. *Progress in Brain Research*, *219*, 45–64. https://doi.org/10.1016/bs.pbr.2015.04.001

Rouhani, N., Norman, K. A., Niv, Y., & Bornstein, A. M. (2020). Reward prediction errors create event boundaries in memory. *Cognition*, *203*, 104269. https://doi.org/10.1016/j.cognition.2020.104269

Shadlen, M. N., & Shohamy, D. (2016). Decision making and sequential sampling from memory. *Neuron*, *90*(5), 927–939.

Shin, Y. S., & DuBrow, S. (2021). Structuring Memory Through Inference-Based Event Segmentation. *Topics in Cognitive Science*, *13*(1), 106–127.

https://doi.org/10.1111/tops.12505

Song, M., Jones, C. E., Monfils, M.-H., & Niv, Y. (2022, May 10). *Explaining the effectiveness of fear extinction through latent-cause inference*. arXiv.Org. https://arxiv.org/abs/2205.04670v1

Speer, N. K., & Zacks, J. M. (2005). Temporal changes as event boundaries: Processing and memory consequences of narrative time shifts. *Journal of Memory and Language*, *53*(1), 125–140.

Srinivasan, A., Riceberg, J. S., Goodman, M. R., Srinivasan, A., Guise, K. G., & Shapiro, M. L. (2023). Goal Choices Modify Frontotemporal Memory Representations. *Journal of Neuroscience*, *43*(18), 3353–3364. https://doi.org/10.1523/JNEUROSCI.1939-22.2023

Stan Development Team. (2021). *Stan Modeling Language Users Guide and Reference Manual* (Version 2.27) [Computer software]. https://mc-stan.org

Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General*, *138*(2), 236–257.

Williams, D. (2001). *Weighing the Odds: A Course in Probability and Statistics*. Cambridge University Press.

Yu, L. Q., Wilson, R. C., & Nassar, M. R. (2021). Adaptive learning is structure learning in time. *Neuroscience and Biobehavioral Reviews*, *128*, 270–281. https://doi.org/10.1016/j.neubiorev.2021.06.024

Zwaan, R. A. (1996). Processing narrative time shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*(5), 1196–1207.

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, *6*(5), 292–297.