

Free recall is shaped by inference and scaffolded by event structure

Corresponding Author: Mr Ata Karagoz

This file contains all editorial decision letters in order by version, followed by all author rebuttals in order by version.

Version 0:

Decision Letter:

**** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors ****

Dear Mr Karagoz,

Thank you for your patience during the peer-review process. Your manuscript titled "Free recall is shaped by inference and scaffolded by event structure" has now been seen by 3 reviewers, whose comments are appended below. You will see that they find your work of some potential interest. However, they have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version that fully addresses these serious concerns.

We hope you will find the Reviewers' comments useful as you decide how to proceed. Should additional work allow you to address these criticisms, we would be happy to look at a substantially revised manuscript. If you choose to take up this option, please highlight all changes in the manuscript text file, and provide a detailed point-by-point reply to the reviewers.

Please bear in mind that we will be reluctant to approach the reviewers again in the absence of substantial revisions.

Editorially, we consider it crucial that additional empirical evidence is provided to strengthen the current conclusions. To this end, we ask that the revision includes data from an additional experiment or additional analyses to address current confounds in the design. Furthermore, please provide a formal model comparison against suitable candidate models to address the referees' concerns in this regard.

I am attaching a checklist that details critical reporting requirements for the revised manuscript. Please attend to each item and ensure your manuscript is fully compliant. We are requesting that your manuscript aligns with these requirements as this facilitates the evaluation of your manuscript, reducing delays in re-review and potential future acceptance. If your revised manuscript is not aligned with these requests on major issues, such as those concerning statistics, it may be returned to you for further revisions without re-review. Additional information can be found in our style and formatting guide <[a href="https://www.nature.com/documents/commpsychol-style-formatting-guide-accept.pdf"](https://www.nature.com/documents/commpsychol-style-formatting-guide-accept.pdf)>Communications Psychology formatting guide.

If the revision process takes significantly longer than five months, we will be happy to reconsider your paper at a later date, provided it still presents a significant contribution to the literature at that stage.

We would appreciate it if you could keep us informed about an estimated timescale for resubmission, to facilitate our planning.

We are committed to providing a fair and constructive peer-review process. Please do not hesitate to contact us if you wish to discuss the revision in more detail.

Please use the following link to submit your

- revised manuscript,
- point-by-point response to the referees' comments,
- cover letter (as a separate document),
- the Editorial Policy Checklist (see below),

- the Reporting Summary (see below), and
- the completed Editorial Request Table (attached):

Link Redacted

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

Thank you for the opportunity to review your work.

Best regards,

Troby Lui

Troby Lui, PhD
Associate Editor
Communications Psychology

REVIEWER EXPERTISE:

Reviewer #1: event segmentation theory, memory
Reviewer #2: memory, computational
Reviewer #3: memory, computational

REVIEWER REPORTS:

Reviewer #1 (Remarks to the Author):

Title: Free recall is shaped by inference and scaffolded by event structure
Authors: Karagoz, Kool, & Reagh

Summary

This manuscript reports a single experiment exploring how event boundaries defined by changes in task goals influence later memory. This was done by having people engage in a hidden rule word task, akin to the Wisconsin Card Sorting Task. Rule changes from one word to the next were identified as event boundaries. This was followed by a distractor task. After that, people were asked to recall as many words as they could remember. The results of the experiment revealed that rule shifts resulted in a decrease in accuracy and a slow-down in response time. Moreover, people did not show better memory for the event boundary items, and showed poorer than experienced memory for post-boundary words. People tended to cluster words from the same event together in their recalls. The data were well-accounted for by a reinforcement learning model, although the outcome deviated from prior work. The results are interpreted as showing that event boundaries can influence performance in this task, but that the ways they do so do not line up with Event Segmentation Theory.

Recommendation

This is an interesting and well-written manuscript that explores interesting some ideas. However, I do have some points of concern that are listed below.

Major Points

1. It would be helpful to consider a wider range of event cognition theories beyond EST, such as the Event Horizon Model and the Event Indexing Model. For these theories, causal structure plays a larger role than is outlined in EST. As it is, the current manuscript is limited in scope.
2. While there are many studies that involve passive observation of events, there are also several studies that look at interactive experiences. As one example, work on how walking through doorways causes forgetting is event cognition research in which people typically are navigating a spatial environment (e.g., Radvansky & Copeland, 2006). There are also studies in which people are actively playing video games, and the active interaction with event boundaries influences performance (e.g., Magliano, Radvansky, & Copeland, 2007). Thus, the current work is not as unique on this dimension as the manuscript implies.
3. By having rule changes every 6-8 trials, did this produce a situation in which people could predict that a rule change was coming? Is this problematic?
4. The manuscript says that people were treating rules shifts as event boundaries, but were they? How do we really know? We know that they changed their behavior, but were they conceptualizing this as an event boundary. Really, what is done here is to operationally define it theoretically as an event boundary, which is fine, but this does involve an inference.
5. The definition of post-boundary slowing is not really the hallmark of an event boundary. Some work has shown that this is

more likely due to a causal break when an event shift occurs, and people are not really sure why it is happening (Pettijohn & Radvansky, 2016). When narrative event boundaries are foreshadowed, there is no slow-down, but other indicators point to an updating of the event models.

6. The manuscript talks about event boundaries being defined by changes in perceptual input, which is not the case here. Perceptually, it is just another word, and just another feedback score. What defines the event boundary here, if that is what it is, is a conceptual shift.

7. The manuscript also discusses the current work as not being passive unlike previous work. This is only partially true. While it is the case that people are more actively involved with the task, it is more passive in that they are not initiating any actions that bring about an event boundary, as compared to, say, walking through a doorway.

Minor points

1. The recall rate (16.6%) seems really low. Is this a problem?
2. There are a LOT of abbreviations in his manuscript. This places a burden on the reader in terms of the amount of information they need to keep track of. Can the number be reduced?
3. When the phrase “on the other hand,...” is used, it should be preceded by the phrase “on the one hand,...” somewhere in this discourse.
4. Compared to other research in which the events occur within and across spatial-temporal frameworks (e.g., rooms, regions, jumps in time, etc.), involving entities (e.g., people and objects), that have properties (e.g., physical, emotional, goals, thoughts, etc.), and may have relationships to one another (e.g., spatial, functional, social, ownership, etc.). There may also be linking relations from one event to another, such as causal relationships. Compared to these types of events, the events in the current study are quite impoverished. This should be explicitly acknowledged at some point.

This has implications for how the manuscript is written. For example, the sentences “A major goal of the present study was to develop an interactive approach to studying the role of event structure in shaping episodic memory. In our paradigm, much like in many events we encounter in daily life, people must interact with their environment to understand which set of behaviors fit a given situation” implies that what was done in the current study is like these everyday events, which just isn’t true. When we are trying to decide what behaviors fit in a given situation, there is often a rich causal structure about why the world is changing, and why some behaviors are better under those circumstances than others. In the study, the changes are arbitrary and the behaviors in response to those changes involve an element of guessing with no causal reason for why that particular behavior would follow from that particular change other than it is different. This doesn’t mean that the current work is wrong, only that some of the generalizations derived from it, and the parallels drawn to everyday life, may go too far.

Reviewer #2 (Remarks to the Author):

This paper fills a specific gap in the literature on how inferred events (inferred task rule) influence subsequent free recall.

The task paradigm requires the subject to make binary decisions for a sequence of word stimuli. Importantly, the latent decision rule has to be inferred by trial and error. As the author pointed out, the task is akin to the Wisconsin card sorting task (WCST). Different from WCST, the stimuli here are word stimuli. Hence, it is easier to connect this study to other free recall / list learning memory paradigms, as they often use word stimuli as well.

The claim that recall is organized by the structure of (inferred) events is supported by the temporal contiguity effect at the event level (e.g. fig5a), which is very cool. It nicely complements the results from Lila Davachi’s group (as cited in this paper) on how events influence people’s episodic memory (e.g. temporal distance judgement, order memory, etc.) and results from the free recall paradigm by Mike Kahana et al.

Suggestions/comments/questions:

First, the results (fig2) demonstrated that inferred event boundaries (rule switches) are associated with increased reaction time for the post-boundary trial. It would be useful to have more discussion on the connection to the switch cost in the task switching / cognitive control literature – in this task paradigm, is the notion of task switch identical to event boundaries? Is switch cost a general characteristic beyond this task paradigm? In general, I think the task-switching literature and event cognition literature can be connected better, and this work seems to be a great place to make this contribution.

For every session, subjects respond to 56 words, and there are 8 latent rule switches. Fig3a shows that there is no primacy effect when looking at the entire 56 items. I’m curious if there is a primacy effect at the event level.

Fig3b shows the conditional-response probability curve – it clearly shows that recall is clustered by temporal distance. Was the forward symmetry effect significant here?

Fig4 shows that task switches seems to hinder memory encoding for that item (at the switch point). This makes sense.

For the conditional-response probability curves (fig3b and 5a) – I thought they should sum to 1. Sorry if I misunderstood something here. But if this is related to the normalizer, perhaps clarifying them in the method section would help.

For fig5c, on the x-axis, what is “Label”? Is it item position within an event?

The boundaries-as-anchor-points idea is very interesting. I'm not sure I understand how the analysis method work. The the following description:

“That is, we predicted that if participants transitioned to recalling items from a new event, they would be more likely to transition to words presented on an event boundary words than those in the middle of events. In line with this prediction, items at boundaries were used as anchors more often than non-boundary items...”

– how did you detect event-level transitions during recall? By detecting whether consecutive words came from different events? What if subjects go back and forth between 2 events (during recall) for multiple time points?
- and if you detected all event boundaries in all recall sequences, do you simply look at all post boundaries items when calculating probabilities?

A reinforcement learning (RL) model is used to capture human behavior during encoding when subjects have to make binary decisions based on the inferred rule. I'm a bit confused about why this is thought of as a RL problem. The feedback after each response is immediate (rather than delayed) and there is no explicit external reward. It seems to be a supervised problem. The feedback is also very supervised – it is a correct/incorrect feedback for a yes/no response. Can the positive/negative reward prediction error (RPE) simply be thought of as positive/negative PE?

The author showed that subsequent recall probability is positively related to negative RPE and negatively related to positive RPE. This is interesting, but I don't know what / how to think about it. Is there any speculation/explanation of why this is the case?

Btw I do think viewing free recall as a RL problem makes sense because, in the free recall task, there is no supervised signal on what to do in both the encoding phase and the recall phase for every time point. Here's a model of free recall that used RL for this reason:

Li, M., Jensen, K. T., Lu, Q., Zhang, Q., & Mattar, M. G. (2024). Modeling Multiplicity of Strategies in Free Recall with Neural Networks. Conference on Cognitive Computational Neuroscience.

... However, in your work, the sub-part being modeled seems pretty supervised.

“In our task, more negative RPEs tend to occur when participants place a high value on a particular rule, and then receive negative feedback when responding according to that rule. ... Meanwhile, particularly positive RPEs occur when people are not placing a high enough value on a particular hypothesis.”

- What do you mean by placing a high “value” on a rule? Do you mean probability/weight? The RL model you used assigns values for the two actions yes vs no.
- and I didn't find the equation describing how RPE is computed. How did you sum the RPEs over rules to get the RPE for a time point?

“In the current study we track the value of responding ‘yes’ or ‘no’ to a stimulus. The value of saying ‘yes’ is calculated as the weighted sum of features present in the stimulus:”

- Why is the value of yes vs. no modeled as the weighted sum of features of the word? How is feature f constructed?
- And if the model is tracking action values (the value of yes vs no) and chooses an action at time t , it seems that positive or negative RPE should signify a discrepancy between the predicted action (yes / no) vs. the feedback (correct / incorrect). In the main text, you said that positive /negative RPE tend to occur when not placing a high-enough weight on a rule / placing too much weight on a rule. Why is that?

In the discussion section, you said that:

“In this study, however, we saw no increase in memory for the pre-boundary or boundary items relative to the non-boundary items. This may result from a difference between interactively discovering an event boundary versus passively viewing it occur.”

There are other places in the paper where the notion of “interactively discovering an event boundary” is used. I think instead of saying “interactively discovering an event boundary”, it is better to say that rule inference disrupts memory encoding. This seems to be the explanation in the main paper. I also think it is a bit strange to describe subjects’ behavior as “interactively discovering an event boundary”, since at event boundaries, the prediction error profile is systematically different from non-boundaries time points, and it is commonly thought that prediction error is one way to trigger event boundaries.

“

One limitation of all free recall experiments, of course, is that participants tend to recall only a fraction of the studied list. This is a limitation in our current experimental paradigm as well.

”

Why is this a limitation?

I think one interesting future direction is to study how event inference impacts episodic memory encoding and subsequent recall with more naturalistic stimuli, like Antony et al (2024).

Antony, J., Lozano, A., Dhoat, P., Chen, J., & Bennion, K. (2024). Causal and chronological relationships predict memory organization for nonlinear narratives. *Journal of Cognitive Neuroscience*, 1–18. https://doi.org/10.1162/jocn_a_02216

Reviewer #3 (Remarks to the Author):

Please see attached

EDITORIAL POLICIES

We ask that you ensure your manuscript complies with our editorial policies and reporting requirements.

To that end, we require revised manuscripts to be accompanied by two completed items: a reporting summary that collects information on study design and procedure, and an editorial policy checklist that verifies compliance with all required editorial policies

- <https://www.nature.com/documents/nr-reporting-summary.zip>>Nature Research Reporting Summary
- <https://www.nature.com/documents/nr-editorial-policy-checklist.pdf>>Editorial Policy Checklist

All points on the policy checklist must be addressed. Your revised manuscript can only be sent back to the referees if these checklists are completed and uploaded with the revision.

Notes: If you have submitted a Stage 1 Registered Report, Review, Primer, Comment, or Perspective you do not need to submit these forms. If you have already submitted these forms, you may disregard this request.

*** TRANSPARENT PEER REVIEW:** Communications Psychology uses a transparent peer review system. This means that we publish the editorial decision letters including Reviewers' comments to the authors and the author rebuttal letters online as a supplementary peer review file. However, on author request, confidential information and data can be removed from the published reviewer reports and rebuttal letters prior to publication. If your manuscript has been previously reviewed at another journal, those Reviewers' comments would not form part of the published peer review file.

**** Visit Nature Research's author and referees' website at** <http://www.nature.com/authors>**>**www.nature.com/authors**** for information about policies, services and author benefits******

Communications Psychology is committed to improving transparency in authorship. As part of our efforts in this direction, we are now requesting that all authors identified as ‘corresponding author’ create and link their Open Researcher and Contributor Identifier (ORCID) with their account on the Manuscript Tracking System prior to acceptance. ORCID helps the scientific community achieve unambiguous attribution of all scholarly contributions. You can create and link your ORCID from the home page of the Manuscript Tracking System by clicking on ‘Modify my Springer Nature account’ and following the

instructions in the link below. Please also inform all co-authors that they can add their ORCID to their accounts and that they must do so prior to acceptance.

<https://www.springernature.com/gp/researchers/orcid/orcid-for-nature-research>

For more information please visit <http://www.springernature.com/orcid>

If you experience problems in linking your ORCID, please contact the [Platform Support Helpdesk](http://platformsupport.nature.com/).

Version 1:

Decision Letter:

**** Please ensure you delete the link to your author homepage in this e-mail if you wish to forward it to your coauthors ****

Dear Mr Karagoz,

Your manuscript titled "Free recall is shaped by inference and scaffolded by event structure" has now been seen by our reviewers, whose comments appear below. In light of their advice I am delighted to say that we are happy, in principle, to publish a suitably revised version in Communications Psychology.

We therefore invite you to revise your paper one last time to address the remaining concerns of our reviewers and a list of editorial requests. At the same time we ask that you edit your manuscript to comply with our format requirements and to maximise the accessibility and therefore the impact of your work.

EDITORIAL REQUESTS:

Please address Reviewer #2's remaining concerns regarding the RL interpretation and the reference list.

Please review our specific editorial comments and requests regarding your manuscript in the attached "Editorial Requests Table". Please outline your response to each request in the right hand column. Please upload the completed table with your manuscript files as a Related Manuscript file.

If you have any questions or concerns about any of our requests, please do not hesitate to contact me.

SUBMISSION INFORMATION:

In order to accept your paper, we require the files listed here <https://www.nature.com/documents/commsj-file-checklist.pdf>.

OPEN ACCESS:

Communications Psychology is a fully open access journal. Articles are made freely accessible on publication. For further information about article processing charges, open access funding, and advice and support from Nature Research, please visit <https://www.nature.com/commspsychol/open-access>

At acceptance, you will be provided with instructions for completing the open access licence agreement on behalf of all authors. This grants us the necessary permissions to publish your paper. Additionally, you will be asked to declare that all required third party permissions have been obtained, and to provide billing information in order to pay the article-processing charge (APC).

*** TRANSPARENT PEER REVIEW:** Communications Psychology uses a transparent peer review system. On author request, confidential information and data can be removed from the published reviewer reports and rebuttal letters prior to publication. If you are concerned about the release of confidential data, please let us know specifically what information you would like to have removed. Please note that we cannot incorporate redactions for any other reasons.

*** CODE AVAILABILITY:** All Communications Psychology manuscripts must include a section titled "Code Availability" at the end of the methods section. We require that the custom analysis code supporting your conclusions is made available in a publicly accessible repository at this stage; please choose a repository that generates a digital object identifier (DOI) for the code; the link to the repository and the DOI must be included in the Code Availability statement. Publication as Supplementary Information will not suffice.

* DATA AVAILABILITY:

All Communications Psychology manuscripts must include a section titled "Data Availability" at the end of the Methods section. More information on this policy, is available in the Editorial Requests Table and at <http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf>.

Please use the following link to submit the above items:

Link Redacted

** This url links to your confidential home page and associated information about manuscripts you may have submitted or be reviewing for us. If you wish to forward this email to co-authors, please delete the link to your homepage first **

We hope to hear from you within two weeks; please let us know if you need more time.

Best regards,

Troby Lui

Troby Lui, PhD
Associate Editor
Communications Psychology

REVIEWERS' COMMENTS:

Reviewer #1 (Remarks to the Author):

I am satisfied with the changes made in the manuscript. I have no further serious concerns.

Reviewer #2 (Remarks to the Author):

I'm sorry for the delayed response. I really appreciate the comprehensive revision. I think the clarity of various results/methods has been improved substantially.

The author(s) addressed most of my concerns. I'm still not 100% sure about the RL interpretation (e.g., the use of reward prediction error rather than simply prediction error). Mathematically, the RW can be viewed as a simplified case of learning an input-output mapping – it basically updates associative strength based on prediction errors. To me, a good fit with the RW-based model doesn't necessarily imply the task is treated as an RL task (by the brain). The interpretation of the results does not heavily draw upon RL theory too much, besides the error-correcting learning component. To me, using PE instead of RPE might be more parsimonious. That said, I understand the reasoning here. So, there is no need to address this point if the author feels very strongly that RL is involved here.

Another minor point: several newly added in-text citations are not in the reference list yet.

Reviewer #3 (Remarks to the Author):

The authors answered all my concerns. I appreciate their thorough and thoughtful responses. Best of luck!

** Visit Nature Research's author and referees' website at www.nature.com/authors for information about policies, services and author benefits**

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

We thank the editor and reviewers for their time and effort in evaluating our manuscript and for giving us helpful feedback. We addressed these comments and suggestions to the best of our abilities, and we believe that the changes made have substantially improved the paper from our initial submission. In addition to many changes to the text, we have run a large ($n = X$) replication of the study, reported in the Supplemental material. This study helps demonstrate the robustness of our approach and results. We have also included new analyses. One of these directly tests the effect of uncertainty on recall. The other is a parameter recovery analysis. We hope that we have addressed the reviewers' concerns and look forward to their evaluation. Please find point-by-point responses below. We have indented reviewer comments, reproduce text from the manuscript in quotation marks and changes to the text in blue.

We will note that upon revising the paper, we discovered a bug in our reinforcement learning model that altered the shape of the results slightly. It is still the case that the highest RPE is at the boundary trial. We now do not see an interaction between RPE magnitude and valence and instead see a general negative relationship between RPE magnitude and predicted recall.

Reviewer 1

This manuscript reports a single experiment exploring how event boundaries defined by changes in task goals influence later memory. This was done by having people engage in a hidden rule word task, akin to the Wisconsin Card Sorting Task. Rule changes from one word to the next were identified as event boundaries. This was followed by a distractor task. After that, people were asked to recall as many words as they could remember. The results of the experiment revealed that rule shifts resulted in a decrease in accuracy and a slow-down in response time. Moreover, people did not show better memory for the event boundary items, and showed poorer than experienced memory for post-boundary words. People tended to cluster words from the same event together in their recalls. The data were well-accounted for by a reinforcement learning model, although the outcome deviated from prior work. The results are interpreted as showing that event boundaries can influence performance in this task, but that the ways they do so do not line up with Event Segmentation Theory.

1. It would be helpful to consider a wider range of event cognition theories beyond EST, such as the Event Horizon Model and the Event Indexing Model. For these theories, causal structure plays a larger role than is outlined in EST. As it is, the current manuscript is limited in scope.

We appreciate the reviewer's suggestion to include other theories of event cognition to broaden the scope beyond EST. We chose to directly consider EST and alluded to the model of Shin and DuBrow (Shin and DuBrow 2020) here as there is no causal structure binding items together. In doing so, we make contact with a growing number of experiments using relatively simple sequences and testing temporal order memory (see DuBrow & Davachi, 2016, among many others). Given that these findings have been interpreted, and many such experiments designed, with EST in mind, we believe that this emphasis is appropriate in our manuscript. However, the reviewer is correct to point out that other frameworks exist, and should be noted. We have added

a short section to both the introduction and to the discussion to highlight both the fact that these other theories exist, and we note in the Discussion that we are not in a place to make direct statements about their applicability to our study due to the lack of explicit causal structure.

Introduction:

“Several theoretical frameworks have been proposed to explain the way people segment and represent continuous events. Among these are Event Segmentation Theory (EST; Zacks et al., 2007), the Event Indexing Model (Zwaan et al., 1995), and the Event Horizon Model (Radvansky, 2012). Here, we focus on EST given its prominent role in guiding recent studies of episodic memory...” (page 3)

Discussion:

“This allowed us to better understand how uncertainty reduction interacts with event model construction. One limitation of all free recall experiments, of course, is that participants tend to recall only a fraction of the studied list. This is a limitation in our current experimental paradigm as well. However, future studies can mitigate this limitation by using a recognition memory paradigm which would enable one to gauge memory signal for each item. Another limitation of the current work is that event boundaries are being inferred from changes in response time and recall behavior rather than directly indexed. In conjunction with the simplicity of the task, this might raise a question of whether “event boundaries” per se are being induced. However, this experimental logic has been applied to many contemporary studies over similar topics (DuBrow & Davachi, 2014; Heusser et al., 2018), and the behavioral effects we observe (e.g., slowed RTs after a shift) are consistent with this operationalization of event boundaries. A final limitation of the current work is that the experiment does not contain causal connections between events. In daily life, there are many causal reasons as to why a given scenario follows another, and there are frameworks of event cognition that put more weight on this (Radvansky 2012, Zwaan et al. 1995, Shin and DuBrow 2021). Future work can expand this experimental paradigm by embedding causal connections between the rule shifts to better compare the predictions of these frameworks.
(page 20)

2. While there are many studies that involve passive observation of events, there are also several studies that look at interactive experiences. As one example, work on how walking through doorways causes forgetting is event cognition research in which people typically are navigating a spatial environment (e.g., Radvansky & Copeland, 2006). There are also studies in which people are actively playing video games, and the active interaction with event boundaries influences performance (e.g., Magliano, Radvansky, & Copeland, 2007). Thus, the current work is not as unique on this dimension as the manuscript implies.

We thank the reviewer for raising this important issue and appreciate the opportunity to clarify our stance. Our goal was not to convey that no *interactive* studies have been done previously. Indeed, both the work that the reviewer cites as well as the work by Heusser and colleagues (2018) that we cite in the paper includes interaction with the stimulus. In the study by Heusser

and colleagues, participants answer questions about the degree to which they like the color combination of item and background color. In the Radvansky and Copeland study, participants perform a much more natural and complex interaction by moving to a table and placing down an object before picking up the next one. The complexity of interaction in our study is diminished relative to this study, but our primary goal was not to study interaction with items *per se*. Instead, we wanted to study event cognition in contexts in which participants make choices to interactively uncover something about the state of the world, that is, the responses they make *resolve uncertainty*. In both aforementioned studies, the interactions do nothing to resolve any uncertainty about the event context. In response to your comment, we have now made the goal of our study clearer in the Introduction:

“A present limitation of studies of event cognition is that these observations largely stem from passive viewing experiments (Swallow, Zacks, and Abrams 2009; Sargent et al. 2013; Bailey et al. 2017; [though see Radvansky and Copeland 2006](#)). However, in many real-world activities, people are not merely passive observers. Instead, they interact with their environment and use the information from these interactions to learn about the world. For instance, merely observing people play a hand of poker might be represented and remembered differently than if you were playing the hand yourself. [Such differences could arise due to the rapidly changing nature of goals as one plays the game leading to different attentional settings \(Cowan et al. 2024, Wang and Egner 2022\)](#). This reveals an important gap in our understanding of event segmentation and its effects on memory: to what extent do theories such as EST and its body of supporting evidence extend to dynamic, active environments [where individuals are making choices to resolve uncertainty about the events in question?](#)” (page 3)

3. By having rule changes every 6-8 trials, did this produce a situation in which people could predict that a rule change was coming? Is this problematic?

We thank the reviewer for highlighting this. We did not find any evidence that participants are predicting that a rule change is incoming. Even in the case where a given participant might develop some sense of “around 7 trials in, it will shift” they would not be able to predict the trial exactly and would still have a prediction error when performing the task.

4. The manuscript says that people were treating rules shifts as event boundaries, but were they? How do we really know? We know that they changed their behavior, but were they conceptualizing this as an event boundary. Really, what is done here is to operationally define it theoretically as an event boundary, which is fine, but this does involve an inference.

We thank the reviewer for bringing up this interesting insight. In previous work, there are two major approaches to identifying event boundaries. In one, participants are explicitly asked to perform segmentation where they believe one event ends and another begins. In the other, event boundaries are induced, and though the participant is not explicitly asked, the presence of event boundaries is inferred by behavioral markers such as slowing response times during transitions. Though this is not a *direct* measure, the operationalization of event boundaries based on these criteria is now standard in the literature. We follow this operationalization from previous literature on event boundaries and their effects on memory (Ezzyat and Davachi, DuBrow and Davachi, Heusser et al., Clewett et al., Radvansky et al., Wang and Egner).

5. The definition of post-boundary slowing is not really the hallmark of an event boundary. Some work has shown that this is more likely due to a causal break when an event shift occurs, and people are not really sure why it is happening (Pettijohn & Radvansky, 2016). When narrative event boundaries are foreshadowed, there is no slow-down, but other indicators point to an updating of the event models.

We thank the reviewer for this comment. “hallmark” was perhaps too strong a word, instead we should say that this post-boundary slowing is a piece of convergent evidence that we can interpret the rule switch detections as event boundaries. We have altered the text to reflect this:

“Another common behavioral indicator of event segmentation is post-boundary slowing (Zwaan and Radvansky, 1998; DuBrow & Davachi, 2014; Heusser et al. 2018; however, we note that post-boundary slowing could be linked to task-switch costs: Rogers & Monsell, 1995).” (page 12)

6. The manuscript talks about event boundaries being defined by changes in perceptual input, which is not the case here. Perceptually, it is just another word, and just another feedback score. What defines the event boundary here, if that is what it is, is a conceptual shift.

This is a fair point. However, while we cannot ascertain with complete certainty that participants are experiencing an event boundary per se, our design is not substantially different from many simplified tasks aimed at investigating the influences of event structure on memory (e.g., Heusser et al., 2018 among many others). We would argue that the boundary is only noticeable due to the perceptual input provided by the feedback score. If the feedback score remained unchanged throughout, there would be no ability to ascertain whether boundaries occurred. This, in conjunction with the fact that participants clearly show altered behavior after such an ostensible boundary has been perceived (e.g., slowing of response times, and robust effects on memory-related behaviors), leads us to feel confident that an important cognitive state transition has occurred. Whether this amounts to an event boundary specifically is perhaps debatable, but if our paradigm is not inducing event boundaries, then we would suggest that a great deal of other contemporary studies are also not showing the induction of event boundaries. We allude to this issue in the Discussion:

“Another limitation of the current work is that event boundaries are being inferred from changes in response time and recall behavior rather than directly indexed. In conjunction with the simplicity of the task, this might raise a question of whether “event boundaries” per se are being induced. However, this experimental logic has been applied to many contemporary studies over similar topics (DuBrow & Davachi, 2014; Heusser et al., 2018), and the behavioral effects we observe (e.g., slowed RTs after a shift) are consistent with this operationalization of event boundaries.” (page 20)

7. The manuscript also discusses the current work as not being passive unlike previous work. This is only partially true. While it is the case that people are more actively

involved with the task, it is more passive in that they are not initiating any actions that bring about an event boundary, as compared to, say, walking through a doorway.

Thank you for raising this point and nudging us to clarify our position. The distinction we are trying to make here is not necessarily between “active” and “passive,” but rather between one’s actions being intentional, strategic, and consequential versus not. In other words, being an active participant in an activity does not necessitate full control over outcomes in one’s environment. In this task, actions are explicitly the method by which participants uncover the latent rule states and so any event boundary happens as a consequence of them interacting with the world that has shifted in the background. We have attempted to make this clearer:

“This reveals an important gap in our understanding of event segmentation: to what extent do theories such as EST and its body of supporting evidence extend to dynamic, active environments [where individuals are making choices to resolve uncertainty about the events in question?](#)” (page 3)

8. The recall rate (16.6%) seems really low. Is this a problem?

We thank the reviewer for this point as it highlights a need for increased clarity in the manuscript. Though the recall rate percentage is low in our task, the number of items recalled is similar to previous studies. On average in our task participants are recalling about 9 words per run. In work examining the effects of encoding conditions on free recall during incidental encoding, the recall rate by percentage is between 37 and 44%, and yet the word lists in that task were 16 words, meaning that participants recalled an average of 6 words per run (Healey 2018). Likewise, in data from Heusser et al. 2018 participants recalled 31.3% of items (about 8 items out of a 24-item list). We believe that our reduction in percentage is due to the increased task demands but do not believe it presents an issue with respect to the analyses or findings.

We have added the following text making this point in the manuscript:

“On average, participants recalled 17.4% of the words they encountered ($\mu = 0.174, \sigma = 0.065$). [While this percentage is low compared to previous free recall studies, the length of the recall list length \(56\) is much longer, so a smaller percentage still reflects a sizable number of words recalled \(Healy 2018; Heusser et al. 2018\).](#)” (page 13)

9. There are a LOT of abbreviations in his manuscript. This places a burden on the reader in terms of the amount of information they need to keep track of. Can the number be reduced?

We appreciate the reviewer pointing this out. We have replaced instances of RL with reinforcement learning and CRP with conditional response probability. We opt to keep RPE and EST as they are field standard abbreviations and because they are used repeatedly throughout the manuscript. We have also added an analysis based on Kullback-Leibler (KL) divergence. We keep this as KL divergence as that is also field standard. We also keep our task abbreviated as the WRIT. If the reviewer feels that these abbreviations should also be removed, we are open to doing so.

10. When the phrase “on the other hand,...” is used, it should be preceded by the phrase “on the one hand,...” somewhere in this discourse.

We thank the reviewer for catching this, we have updated the text to instead read:

“.... This suggests that constructing event models in environments where one can interactively reduce uncertainty is not conducive to encoding episodic memories. [At the same time, we found that](#) event structure guided free recall, such that participants used events as a fundamental organizing property...” (page 18)

11. Compared to other research in which the events occur within and across spatial-temporal frameworks (e.g., rooms, regions, jumps in time, etc.), involving entities (e.g., people and objects), that have properties (e.g., physical, emotional, goals, thoughts, etc.), and may have relationships to one another (e.g., spatial, functional, social, ownership, etc.). There may also be linking relations from one event to another, such as causal relationships. Compared to these types of events, the events in the current study are quite impoverished. This should be explicitly acknowledged at some point.

We thank the reviewer for this point. There is often a tradeoff between adhering to the complexity of real-world experience and experimental control. In this work, we opted to develop an experimental paradigm to try and isolate a phenomenon of interest. The events as operationalized in our manuscript are in line with many previous experiments (see DuBrow & Davachi, 2014; Heusser et al., 2018). We believe that this does not need to explicitly acknowledged, but are happy to reconsider if the reviewer believes this to be an important point.

12. This has implications for how the manuscript is written. For example, the sentences “A major goal of the present study was to develop an interactive approach to studying the role of event structure in shaping episodic memory. In our paradigm, much like in many events we encounter in daily life, people must interact with their environment to understand which set of behaviors fit a given situation” implies that what was done in the current study is like these everyday events, which just isn’t true. When we are trying to decide what behaviors fit in a given situation, there is often a rich causal structure about why the world is changing, and why some behaviors are better under those circumstances than others. In the study, the changes are arbitrary and the behaviors in response to those changes involve an element of guessing with no causal reason for why that particular behavior would follow from that particular change other than it is different. This doesn’t mean that the current work is wrong, only that some of the generalizations derived from it, and the parallels drawn to everyday life, may go too far.

We appreciate the reviewer’s concerns. While we would agree that many events in daily life have rich causal structure, we would nonetheless specify that this work captures a dimension of natural behavior not previously well researched. That is: our study is primarily focused on understanding the role of active decision making in segmenting events, and the consequences of

this particular process on memory-guided behaviors. Though we follow the quoted statement with a caveat, we have expanded this to mitigate potential parallels that readers may infer. We have added the following explanation of the limitation to the discussion in response to point 1 and believe it also handles this point:

“This allowed us to better understand how uncertainty reduction interacts with event model construction. One limitation of all free recall experiments, of course, is that participants tend to recall only a fraction of the studied list. This is a limitation in our current experimental paradigm as well. However, future studies can mitigate this limitation by using a recognition memory paradigm which would enable one to gauge memory signal for each item. Another limitation of the current work is that event boundaries are being inferred from changes in response time and recall behavior rather than directly indexed. In conjunction with the simplicity of the task, this might raise a question of whether “event boundaries” per se are being induced. However, this experimental logic has been applied to many contemporary studies over similar topics (DuBrow & Davachi, 2014; Heusser et al., 2018), and the behavioral effects we observe (e.g., slowed RTs after a shift) are consistent with this operationalization of event boundaries. A final limitation of the current work is that the experiment does not contain causal connections between events. In daily life, there are many causal reasons as to why a given scenario follows another, and there are frameworks of event cognition that put more weight on this (Radvansky 2012, Zwaan et al. 1995, Shin and DuBrow 2021). Future work can expand this experimental paradigm by embedding causal connections between the rule shifts to better compare the predictions of these frameworks.” (page 20)

Reviewer 2

This paper fills a specific gap in the literature on how inferred events (inferred task rule) influence subsequent free recall.

The task paradigm requires the subject to make binary decisions for a sequence of word stimuli. Importantly, the latent decision rule has to be inferred by trial and error. As the author pointed out, the task is akin to the Wisconsin card sorting task (WCST). Different from WCST, the stimuli here are word stimuli. Hence, it is easier to connect this study to other free recall / list learning memory paradigms, as they often use word stimuli as well.

The claim that recall is organized by the structure of (inferred) events is supported by the temporal contiguity effect at the event level (e.g. fig5a), which is very cool. It nicely complements the results from Lila Davachi's group (as cited in this paper) on how events influence people's episodic memory (e.g. temporal distance judgement, order memory, etc.) and results from the free recall paradigm by Mike Kahana et al.

1. First, the results (fig2) demonstrated that inferred event boundaries (rule switches) are associated with increased reaction time for the post-boundary trial. It would be useful to have more discussion on the connection to the switch cost in the task switching / cognitive control literature – in this task paradigm, is the notion of task switch identical to event boundaries? Is switch cost a general characteristic beyond this task

paradigm? In general, I think the task-switching literature and event cognition literature can be connected better, and this work seems to be a great place to make this contribution.

We thank the reviewer for making this point. We have included more text in the results to better clarify our perceived connection between task switching literature and event boundaries.

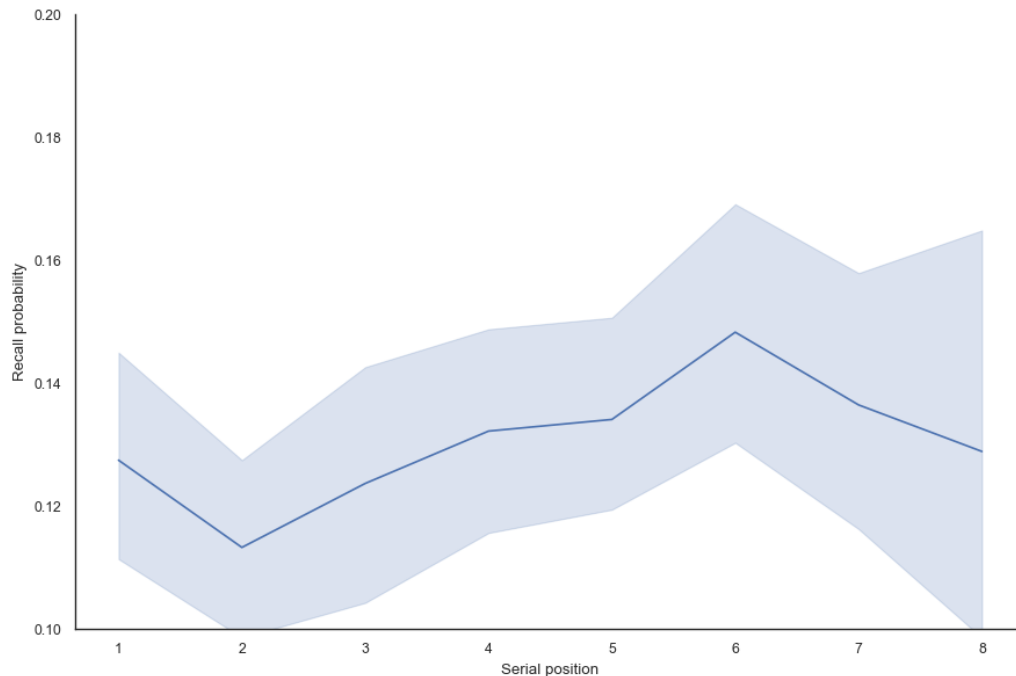
“Another common behavioral indicator of event segmentation is post-boundary slowing (Zwaan and Radvansky, 1998; DuBrow & Davachi, 2014; Heusser et al. 2018). However, we note that post-boundary slowing could be linked to task-switch costs (Rogers & Monsell, 1995). Consistent with this phenomenon, a linear mixed effects model showed that participants slowed down after the rule shift ($\beta_{post_boundary} = 0.038$, $CI_{95\%} = [0.012, 0.064]$, $p = 0.004$; Figure 2B).” (page 12)

We have also included further text in the discussion to better highlight the potential connections of our work with task-switching and cognitive control processes.

“This enabled us to discover effects of event boundaries on memory that run counter to predictions of EST. Furthermore, the gain in information from individual trials is slower and more easily modeled than in more naturalistic stimuli. This allowed us to better understand how uncertainty reduction interacts with event model construction, as well as better investigate reaction times as in the task-switching literature (Rogers and Monsell 1995).” (page 20)

2. For every session, subjects respond to 56 words, and there are 8 latent rule switches. Fig3a shows that there is no primacy effect when looking at the entire 56 items. I’m curious if there is a primacy effect at the event level.

We thank the reviewer for highlighting this interesting possibility. We examined this and found no evidence for an event-specific primacy effect. We have included the figure produced by the analysis below.



- Fig3b shows the conditional-response probability curve – it clearly shows that recall is clustered by temporal distance. Was the forward symmetry effect significant here?

We thank the reviewer for this question. Upon running a paired samples t-test we found no evidence for a forward asymmetry effect. We have inserted this result into the text:

“Recall performance also featured a high degree of temporal contiguity (Figure 3B), with participants tending to recall sequences of words they experienced close in time in the WRIT (Kahana, Diamond, and Aka, 2022). However, we did not find evidence for the standard forward asymmetry in recall (test of lag+1 vs lag-1, $t(72) = 1.72$, $Cohen's d = 0.29$, $p = 0.091$).” (page 13)

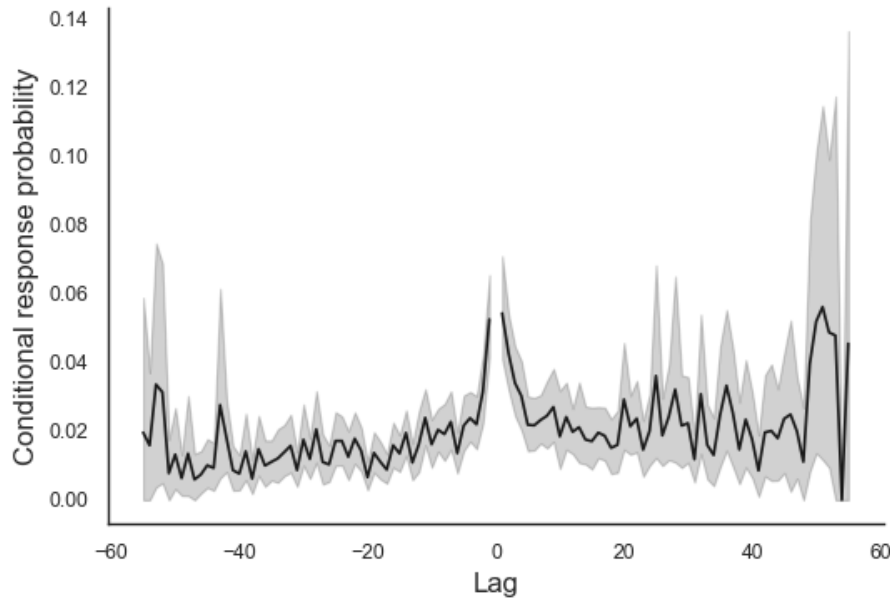
- For the conditional-response probability curves (fig3b and 5a) – I thought they should sum to 1. Sorry if I misunderstood something here. But if this is related to the normalizer, perhaps clarifying them in the method section would help.

We apologize for the lack of clarity regarding the CRP analysis. While the reviewer is correct in that the sum of the curve should be 1, we truncate our CRP curve to show only local transitions of lag size up to 8. As our word lists are longer than previous studies, we see long noisy tails out to jumps of size 55 (the maximum jump size). We have adjusted the figure caption to specify that we have truncated the lag to this range to focus on the contiguity effect:

“**B.** Conditional response probability curve showing evidence of canonical temporal contiguity effects in recall. The lags have been truncated to display only -8 to +8 to focus on the temporal

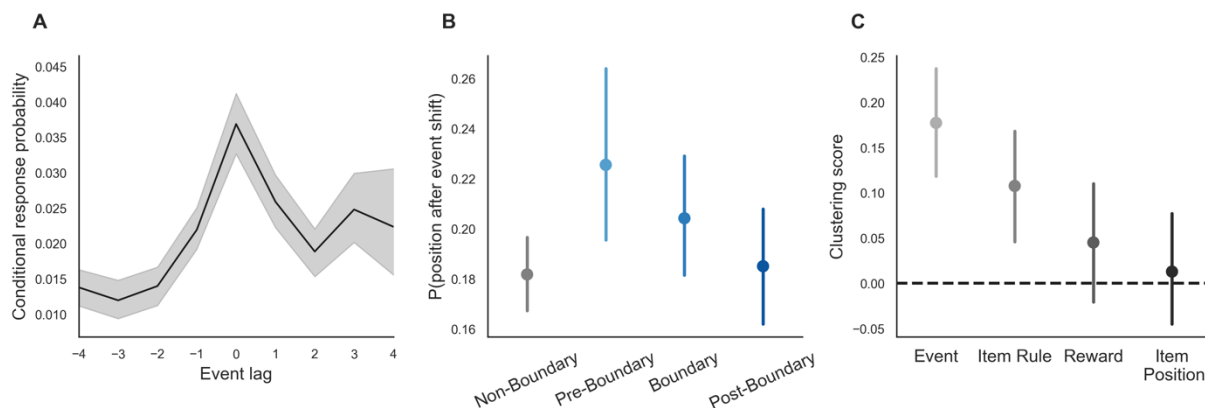
contiguity effect (full version in Supplemental results). Shading is 95% confidence interval.” (page 11)

We have also included the full panel in the supplement and reproduced it here for convenience:



5. For fig5c, on the x-axis, what is “Label”? Is it item position within an event?

We thank the reviewer for catching this error in generating the plot. Your assumption was correct. We have remade the figure and adjusted the label accordingly. We reproduce the updated panel below for convenience.



6. The boundaries-as-anchor-points idea is very interesting. I’m not sure I understand how the analysis method work. The following description:
 “That is, we predicted that if participants transitioned to recalling items from a new

event, they would be more likely to transition to words presented on an event boundary words than those in the middle of events. In line with this prediction, items at boundaries were used as anchors more often than non-boundary items...”
– how did you detect event-level transitions during recall? By detecting whether consecutive words came from different events? What if subjects go back and forth between 2 events (during recall) for multiple time points? and if you detected all event boundaries in all recall sequences, do you simply look at all post boundaries items when calculating probabilities?

We appreciate the reviewer’s request for greater clarity when describing this analysis. You have guessed exactly correctly that we detect event-level transitions in recall by looking at whether consecutive recalled words came from different events. We look at all post-event-shift items in this analysis.

If participants go back and forth between 2 events during recall for multiple timepoints, each of those shifts would be categorized in the boundaries-as-anchor-points analysis. If a participant jumps from event-boundary of event 1 to the middle of event 2 and back to middle of event 1 and then keeps going back and forth you would expect non-boundary proportions to be very high.

We have included the following text to the methods to better clarify the analytic approach:

“Transitions between events. Recent evidence points to event boundaries as “anchor points” in recall. When transitioning between events, participants are more likely to jump to boundaries than information that occurred in the middle of an event (Michelmann et al. 2023). To assess this, we calculated the likelihood of transitioning to specific positions when moving to a new event. For each participant, we took the position of items that were recalled after a shift in event relative to an event boundary. *We used this approach to identify all the instances during recall wherein two consecutive words belonged to different events. For example, if a participant’s first recalled item was from event 2 and they next recalled an item from event 4, we examined the relative location of the event 4 item with respect to its nearest boundary during encoding. This provided us with relative proportions that each participant used boundary items as anchor points.”* (page 10)

We have also included a call out from the results:

*“That is, we predicted that if participants transitioned to recalling items from a new event, they would be more likely to transition to words presented on an event boundary words than those in the middle of events. We detected event level transitions during recall by assessing whether participants recalled consecutive words from different events (See Methods: *Transitions between events*).”* (page 15)

7. A reinforcement learning (RL) model is used to capture human behavior during encoding when subjects have to make binary decisions based on the inferred rule. I’m a bit confused about why this is thought of as a RL problem. The feedback after each

response is immediate (rather than delayed) and there is no explicit external reward. It seems to be a supervised problem. The feedback is also very supervised – it is a correct/incorrect feedback for a yes/no response. Can the positive/negative reward prediction error (RPE) simply be thought of as positive/negative PE?

We thank the reviewer for this opportunity to increase clarity. Part of the reason we use RL is that it fits participant performance well in previous work that inspired our task (Niv et al., 2015). When examining model fits, we can see that this model performs well in predicting participant performance. A new parameter recovery analysis indicates that we robustly recover all parameters (**Supplement: Recovery Analysis**).

8. The author showed that subsequent recall probability is positively related to negative RPE and negatively related to positive RPE. This is interesting, but I don't know what / how to think about it. Is there any speculation/explanation of why this is the case? Btw I do think viewing free recall as a RL problem makes sense because, in the free recall task, there is no supervised signal on what to do in both the encoding phase and the recall phase for every time point. Here's a model of free recall that used RL for this reason:

Li, M., Jensen, K. T., Lu, Q., Zhang, Q., & Mattar, M. G. (2024). Modeling Multiplicity of Strategies in Free Recall with Neural Networks. Conference on Cognitive Computational Neuroscience.

... However, in your work, the sub-part being modeled seems pretty supervised.

We thank the reviewer both for the question and the interesting citation provided. First, it is our opinion that the RPE effects in our work stem from the specifics of the structure of our task. It is the case in our task that negative RPEs occur when a specific rule combination is overvalued. When looking at Figure 5A one can see that the largest negative RPE occurs at the boundary timepoint. This is because the participants at that point should have a stable value for the rule from the previous block and when forced into an error by our design receive no reward when expecting one. Positive RPEs mainly occur right after the event boundary as all the rules have a low level of value and so any reward will add value to the rule on that trial. We have enhanced the clarity of the explanation in the main text as follows:

“Upon further examining the RPE as a function of trials (Figure 6A), one can see that in most trials there is a small positive prediction error. This could be somewhat of an artifact of how the weight decay works in the model. On any given trial only 1 rule is active but values of 2 rules are always updated based on the choice and reward. This leads to a situation where the non-relevant feature alternates it constantly generates a small positive RPE. Interestingly, the model without this decay fits worse across all participants (See **Supplement: Model Comparison**).” (page 16)

9. “In our task, more negative RPEs tend to occur when participants place a high value on a particular rule, and then receive negative feedback when responding according to that rule. ... Meanwhile, particularly positive RPEs occur when people are not placing a high enough value on a particular hypothesis.”
- What do you mean by placing a high “value” on a rule? Do you mean

probability/weight? The RL model you used assigns values for the two actions yes vs no.

- and I didn't find the equation describing how RPE is computed. How did you sum the RPEs over rules to get the RPE for a time point?

We apologize for the lack of clarity regarding the model. First, the word value in your quote should have been replaced with weight. Second, we have included the following text to explain how RPE is calculated

“The value of ‘no’ is the sum of stored weights for the features that are not present in S . Upon receiving feedback, the model updates each weight using a Rescorla-Wagner update rule:

$$W^{new}(f) = W^{old}(f) + \eta[R_t - V(\text{choice})] \forall f \in \text{choice}$$

Here $[R_t - V(\text{choice})]$ is the reward prediction error (RPE) between the value of the choice (‘yes’ or ‘no’) and the reward outcome R_t . This is multiplied by the learning rate η which is a free parameter fit for each participant. This equation is used to update all the feature weights that were present in the stimulus if the choice was ‘yes’, or the weights that are not in the stimulus if the choice was ‘no’. We update the value of all the feature weights associated with the stimulus or its inverse as we cannot be sure of the exact feature the participant is responding to.” (page 7)

10. “In the current study we track the value of responding ‘yes’ or ‘no’ to a stimulus. The value of saying ‘yes’ is calculated as the weighted sum of features present in the stimulus:”

- Why is the value of yes vs. no modeled as the weighted sum of features of the word? How is feature f constructed?

- And if the model is tracking action values (the value of yes vs no) and chooses an action at time t , it seems that positive or negative RPE should signify a discrepancy between the predicted action (yes / no) vs. the feedback (correct / incorrect). In the main text, you said that positive /negative RPE tend to occur when not placing a high-enough weight on a rule / placing too much weight on a rule. Why is that?

We thank the reviewer for this question and opportunity to clarify the model.

First, we use the weighted sum of the features in the word as the participant cannot be sure which of the features they are receiving reward for. For instance, if the correct rule is small and manmade (such as a baseball) or small and natural (such as an apple) would both provide a reward of 1 if the participant responded yes.

The feature f is a set of binary values that indicate whether an item belongs to one of four categories. The value of one responding yes is modeled as the mixture of both the origin and size dimension. We have added clarification as to how the feature f is constructed:

“Here, the model calculates the value (V) of the ‘yes’ response as the sum of the stored weights for each feature present in the stimulus (S). The stimulus on each trial is modeled as a binary feature vector (f), where for each of the two stimulus dimensions (size and origin) the stimulus can have a value of 0 or 1. If items were “small”, they had a 0 in the first dimension, and if they

were “large” this was a 1. Meanwhile “natural” and “manmade” corresponded to 0 and 1 in the second dimension respectively. Therefore, an item like a baseball which is small and manmade would have the feature vector [0, 1].” (page 7)

11. In the discussion section, you said that:

“In this study, however, we saw no increase in memory for the pre-boundary or boundary items relative to the non-boundary items. This may result from a difference between interactively discovering an event boundary versus passively viewing it occur.”

There are other places in the paper where the notion of “interactively discovering an event boundary” is used. I think instead of saying “interactively discovering an event boundary”, it is better to say that rule inference disrupts memory encoding. This seems to be the explanation in the main paper. I also think it is a bit strange to describe subjects’ behavior as “interactively discovering an event boundary”, since at event boundaries, the prediction error profile is systematically different from non-boundaries time points, and it is commonly thought that prediction error is one way to trigger event boundaries.

We thank the reviewer for raising these opportunities to improve clarity. First, we have revised instances where the phrase in question was used, in order to more explicitly lay out the phenomenon in question. For example:

“This may result from a difference between [discovering that an event boundary has occurred through interaction with the environment](#), versus passively viewing it occur.” (page 18)

Reviewer 3

The authors addressed an interesting and timely question, namely, how hidden boundaries (unobserved changes of rules) influence memory. They found that memories for words were clustered based on events, as well as that positive and negative reward prediction errors had opposite effects on memory recall. Although the research is interesting and of potential merit, there are methodological issues that question the robustness of the findings.

1. The authors should replicate their findings in an additional sample. Ideally, with some extension of the findings (see comment 2).

We thank the reviewer for this suggestion and have collected a replication sample of the experiment we report in the main text. The results of the replication sample are reported in the Supplement under **Replication Sample**. All the major results of the manuscript are replicated in the larger replication sample.

2. One issue with the current design is that, while the authors argue that the results stem from the inference of a rule change, RPE and inference are confounded in the current design. Thus, it is unclear whether the memory effects seen result from the influence of RPE on memory or the ensuing inference of the rule change. One potential direction, related to point 1, could be to replicate and extend the finding while de-confounding RPE

and inference. This could be done in different ways, but one idea is to change the task such that under one rule, participants need to indicate, e.g., bigger or smaller than a backpack. Then, they could potentially be wrong about the feature (e.g. if some objects are about the same size), and that would elicit RPE, but the rule wouldn't change (and it'll change at other times). If I'm not mistaken, this situation might exist already in the study (e.g., the mountain in Figure 1). If so, analyzing these instances would be helpful, but some replication would still be needed, ideally with a direct experimental manipulation decoupling RPE and inference. Alternatively, finding a way to invoke inference without explicit reward could also decouple RPE and inference.

We appreciate that LCI and RPE are confounded here as RPE is the mechanism we used to inform participants that they should shift rules and begin performing inference. We have added text to the discussion to highlight this confound and temper mechanistic interpretations.

“Previous work incorporating reward into event cognition has found that unsignaled surprise heightens memory and forms event boundaries (Rouhani et al. 2018, Rouhani and Niv, 2021). [Here, we found the inverse, that large magnitude RPEs dampen memory.](#) This may be due to the structure of our task, where the largest magnitude negative RPEs occur at the event boundary by design [and the largest positive RPEs occur directly afterwards.](#) [Though it should be noted that the effect of higher magnitude positive RPEs predicting memory decrements replicates, negative RPEs show a different trend \(Supplemental Results: Replication\).](#) Nonetheless, this reveals that memory is not unilaterally affected by surprise, a distinction not made by theories of event cognition.

[Theories of latent cause inference predict increased uncertainty after rule shifts, at the same time when our reinforcement learning model encodes particularly strong negative RPEs. Indeed, the RPEs signal the need for a revaluation of the internal task representation. This suggests that our model, even though not explicitly formulated in terms of latent cause inference, accounts for uncertainty reduction over latent causes. Consistent with this, KL divergence \(a measure of uncertainty over rules\) strongly predicted memory performance, in line with predictions from latent cause inference-based theories of event cognition \(Shin and DuBrow; 2021\). Future experiments could aim to better disentangle latent cause inference and reinforcement learning to enable different modeling approaches, as previous work has highlighted the strong connections between latent cause inference and event cognition \(Franklin et al., 2020; Shin and DuBrow, 2021; Varga, Morton and Preston, 2022; Bein and Niv, 2023; Lu et al., 2024; Beukers et al., 2024\).” \(page 19\)](#)

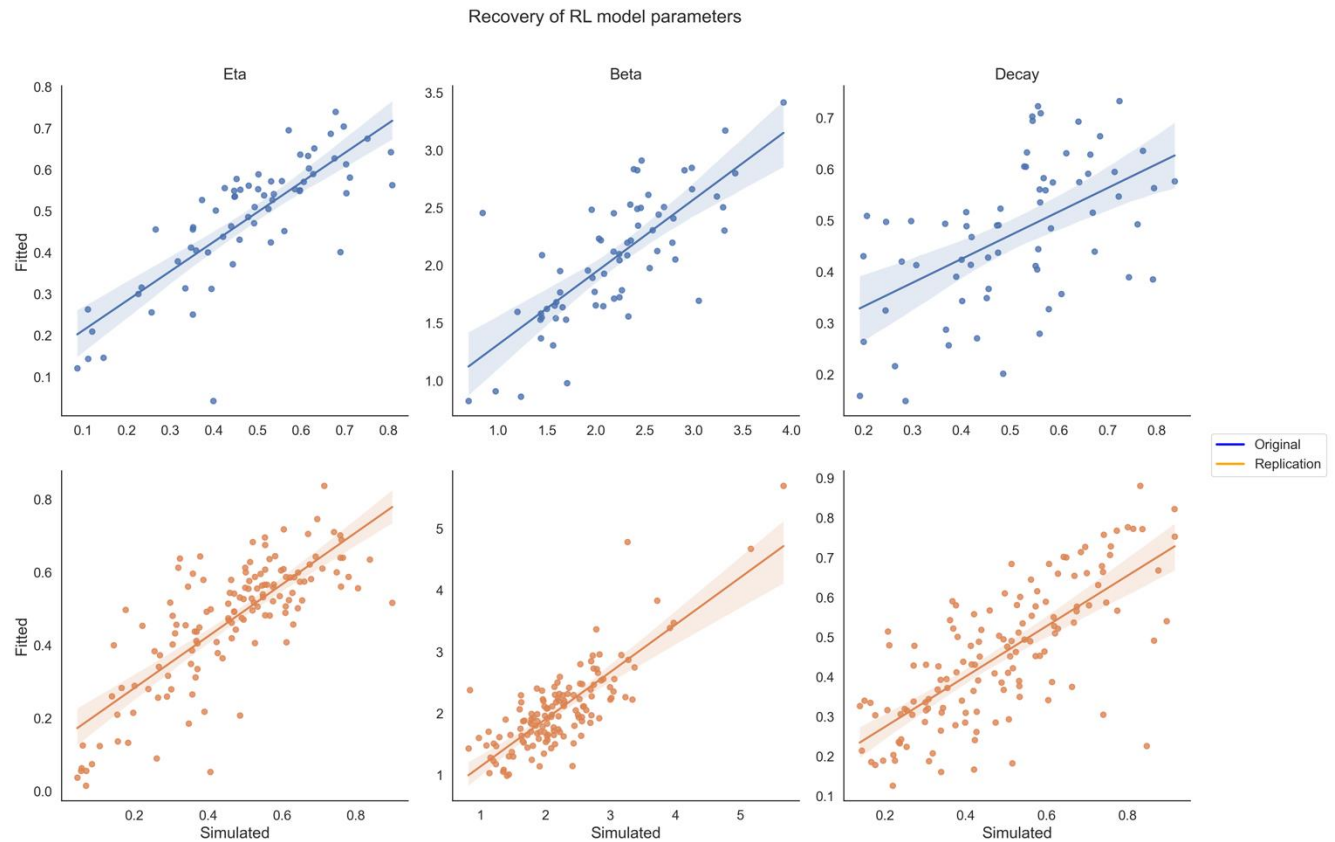
Also, while we believe that decoupling these mechanisms is informative, we would argue that extending with a separate experiment is not trivial and is beyond the scope of the current manuscript. To be clear, this is an interesting avenue for a future study that our group is interested in pursuing, but the fundamental issue is that our study and response contingencies are not ideally set up to fit an LCI model. We hope that the above additions to the Discussion are satisfactory, and if the reviewer feels that further steps are needed, we are happy to continue the conversation.

- Did the authors perform any model comparison or use other methods (e.g., parameter recovery) to ensure the model adequately fits the data? Please report these. This, together with a replication, is important especially since that the authors do not replicate previous RPE effects on memory recall.

We thank the reviewer for highlighting this point. We have included parameter recovery results in the supplement (**Parameter Recovery**). As well as alluded to it in the main methods:

“Thus, the model had 3 free parameters (η, β, d), which we fit using a maximum a-posteriori approach for each participant separately, with priors set based on previous work (Gershman 2016). A parameter recovery analysis (see Supplement: Parameter Recovery) indicated that this model was robustly identifiable from behavior.” (page 8)

We reproduce the primary results of the parameter recovery here for convenience:



Supplemental Figure 7: Recovery analysis. The top row depicts the correlations between the simulated and fitted parameters for the original dataset in the manuscript. The bottom row depicts the same for the replication dataset.

Experiment	parameter	n	rho	CI95%	p
Original	Eta	66	0.819852	[0.72, 0.89]	<0.001

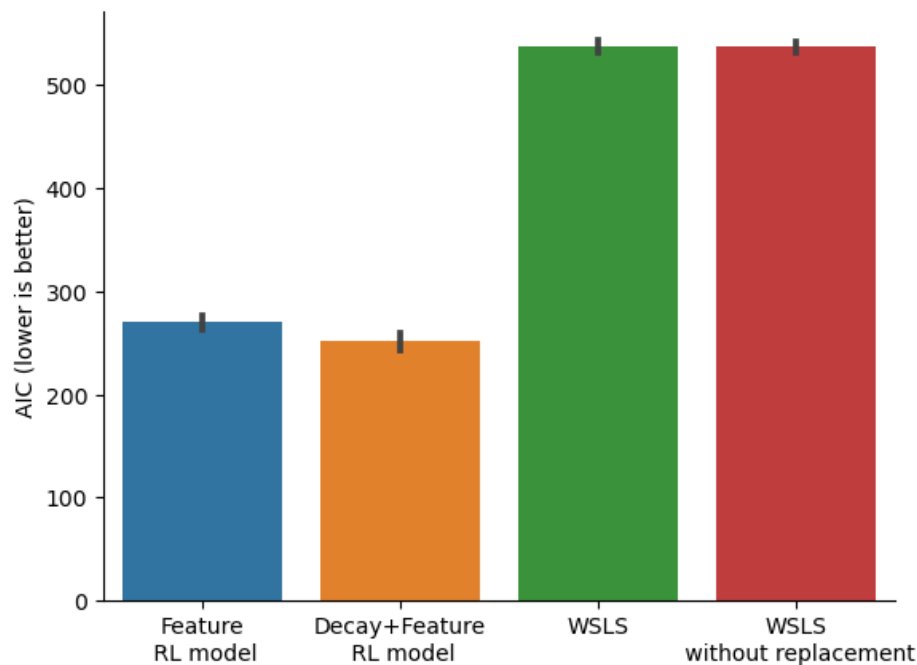
Original	Beta	66	0.756268	[0.63, 0.84]	<0.001
Original	Decay	66	0.520718	[0.32, 0.68]	<0.001
Replication	Eta	139	0.756361	[0.67, 0.82]	<0.001
Replication	Beta	139	0.725363	[0.64, 0.8]	<0.001
Replication	Decay	139	0.721618	[0.63, 0.79]	<0.001

Supplemental Table 6: Results from the parameter recovery analyses. In each sample, each of the three parameters was recoverable.

We also report the results of a model comparison in the supplement which we reproduce here for convenience:

We compared the model reported in the main text with a Feature RL model without a decay parameter, as well as two win-stay-lose-shift models on the primary sample. The Feature RL model without decay performs all the same updates as the primary model reported in **Methods: reinforcement learning** except for incrementing down of non-selected rules.

The win-stay-lose-shift models predict that if a reward was received for a rule, the participant will continue using that rule with some probability (1-epsilon). Otherwise, the participant will either shift to a random rule of the four (WSLS) or shift to one of the other three rules.



Supplemental Figure 8: Model comparison. The Feature RL model with decay outperforms all other models such that AIC is lower for all participants with that model compared to the other 3.

- Related to 4, since the theory that motivated the study is latent cause inference, it could be interesting to fit a latent cause inference to the data (e.g., Wilson and Niv, 2014; Mirea et al., 2024).

We thank the reviewer for highlighting this interesting possibility. While we also agree that fitting LCI models would be interesting, porting the models of Mirea and colleagues over to this task is not trivial. Moreover, in previous work from Niv and colleagues, models of this type are generally outperformed by the decay+feature RL model that we have fit here (Niv et al. 2015). We have tried to adapt the LCI models from Mirea et al. though there are two major points of discrepancy. First, in the Mirea et al. model the number of latent causes is unknown to the participant, in our case participants are told the exact latent causes that can occur. Second, in our experiment we do not have certainty over which latent cause participants are choosing to respond to at the current moment. For instance, when they see baseball and say “yes” they could either be testing “small” or “manmade” as their hypothesis. Due to these discrepancies, we believe that current LCI models are not appropriate for this data and task. However, to try and approximate the LCI process we created a bespoke measure of the certainty over the rules. This measure of certainty is intuitively related to the distribution of weights in the RL model. When the weights are all very similar to one another the RL model is said to be “uncertain” and when the weight is much higher for one rule than the others it has higher certainty. We describe this metric in the methods and reproduce it here for convenience:

“In addition to examining the relationship between behavior and reward prediction errors from this model, we examined its relationship with the “certainty” of the weights over the rules. First, we converted the weight matrix (W) to a set of probabilities using a logistic transformation:

$$p(\text{rule}) = \frac{e^{(W_{\text{rule}} - \max(W))}}{\sum_{i \in \text{rules}} e^{(W_i - \max(W))}}$$

This produces a probability distribution of rules for each trial in the task. If the probability distribution is uniform, uncertainty about the current is maximally large. Therefore, deviances from a uniform distribution, measured using Kullback-Leiber (KL) divergence (Kullback and Leibler 1951) indicate increased rule certainty.” (page 8)

This measure of certainty is a strong predictor of recall both in the primary and replication samples. It is our belief that this measure helps form a bridge to the LCI literature.

5. The RT analyses:

1. Are the two models testing for RT effects independent, and should they be interpreted as different results? What happens if the regressors of `previous_no_reward`, `boundary_label`, and `trial_within_event` are together in one model, such that the effect could all be modeled together?
2. RTs do not distribute normally, please either log-transform the data or follow recommendations for a general-linear model with an inverse Gaussian link function (Lo and Andrews, 2015).

We appreciate the reviewer highlighting these two points. First, the reviewer is correct in noticing that we should combine the two RT effects into one model. In another comment the reviewer suggested that the interaction term with a post-boundary indicator is potentially superfluous. We have chosen thus to model the RT in the task by first log transforming them. The new combined linear model reads:

$$RT \sim trial_within_event + previous_no_reward + boundary_label + (1| participant) + (1| word)$$

The new model is more parsimonious and models the effect of post-error slowing and the boundary-labels while also accounting for the more general speedup of responses over the course of the event. We have altered the methods section to better explain this new model and appreciate the reviewer's suggestion to combine them.

“Response time and post-error slowing

As participants adapted to a new hidden rule over the course of an event, we expected response times to decrease. Along with this decrease, we expected to find post-error response times to be elevated as a measure of inference. We reasoned that trials following no reward would be slower than trials following reward, indicating post-error slowing (Danielmeier and Ullsperger 2011). Finally, we wanted to account for boundary-related slowing over and above these more general post-error slowing and event-related speedup effects. To examine these potential effects on response times, we ran a hierarchical mixed-effects model of the following form:

$$\log(RT) \sim trial_within_event + previous_no_reward + boundary_label + (1| participant) + (1| word)$$

Here, the first trial within an objective event (defined by a rule shift) is modeled as 0 and each trial afterwards is incremented by 1. The variable indicating whether a previous trial provided reward was coded as 0 if the previous trial was correct and 1 if the previous trial was incorrect such that a positive effect indicates post-error slowing. Finally, the boundary label variable is a categorical regressor, where the item's position is coded such that the trial at $t + 1$ after the boundary is called the “post-boundary” trial, $t - 1$ is called the “pre-boundary”, and all other trials are marked as “non-boundary”. Non-boundary is modeled as the base factor. We log transform the RT data according to recommendations by Lo and Andrews (2015).” (page 9)

We have also altered the results section to discuss the outcome of the combined model as follows:

“Another common behavioral indicator of event segmentation is post-boundary slowing (Zwaan and Radvansky, 1998; DuBrow & Davachi, 2014; Heusser et al. 2018). However, we note that post-boundary slowing could be linked to task-switch costs (Rogers & Monsell, 1995). Consistent with this phenomenon, a linear mixed effects model showed that participants slowed down after the rule shift ($\beta_{post_boundary} = 0.038$, $CI95\% = [0.012, 0.064]$, $p = 0.004$; Figure 2B). Importantly, this model accounted for post-error slowing ($\beta_{prev_no_reward} = 76.70$, $CI95\% = [21.28, 132.13]$, $p = 0.007$) and general speeding over the course of an event ($\beta_{trial_within_event} = -0.011$, $CI95\% = [-0.016, -0.006]$, $p < 0.001$). Therefore, the slowing after a rule shift can be uniquely attributed to the participants finding out that the active rule has changed. We also found a general speedup for the boundary trial in this model ($\beta_{boundary} = -0.058$, $CI95\% = [-0.085, -0.032]$, $p < 0.001$). There was no difference in speed for pre-boundary and non-boundary items when accounting for general post-error slowing and within-event speedup ($\beta_{pre_boundary} = 0.003$, $CI95\% = [-0.020, 0.026]$, $p = 0.814$).” (page 12)

6. p. 3: I think that the contrast between temporal/associative memory and item memory should be made clearer for the general reader, along with the notion that they are potentially subserved by different mechanisms, so we cannot extrapolate from one to the other.

We thank the reviewer for this suggestion. We have rewritten this section of the Introduction to highlight the differences between associative and item memory as follows:

“EST in particular makes predictions about changes to both associative memory and item memory through separate processes. According to EST, differences in associative memory are thought to occur because of binding processes within-events and separation processes due to boundaries between events. The coming together of items within events and separation across them has been borne out by numerous studies (Davachi and DuBrow 2016; Heusser et al. 2018; Pu et al 2022; Wang and Egnér 2022). The theory also predicts that memory for information around boundary points will be enhanced. This enhancement is thought to occur due to a sharpening of representations at the end of an event and the opening of perceptual gates and heightening of attention at the beginning of a new event (Swallow, Zacks, and Abrams 2009; also see Richmond and Zacks 2017; Clewett, DuBrow, and Davachi 2019 for review). In short, EST argues that prediction errors during event boundaries establish the need for a new event model, and that this process leads to enhanced within-event associative memory and enhanced memory for items occurring near boundaries.”

7. p. 3: the authors discuss item vs. associative memory but then move on to discuss action. This reads to me like a separate issue – please clarify better whether it is separate or meant to answer the question raised about item memory.

We thank the reviewer for pointing out the lack of a smooth transition and in response to this concern and others we have rewritten our introduction to better clarify our line of argument and how all these mnemonic features relate. We now place interaction and inference first and then begin to talk about associative and item memory. We hope that the restructuring of the introduction proves to be helpful in clarifying our line of argument.

8. p. 3: “For instance, merely observing people play a hand of poker might be represented and remembered differently than if you were playing the hand yourself. “ I agree, but please elaborate on why, in the authors’ view, that would be the case. Is it that people’s attention is guided differently? Do their goals differ? And these are known to influence how memories are formed?

We thank the reviewer for highlighting the need for further evidence here. We have specified the shift and goals and the further effects on memory:

“For instance, merely observing people play a hand of poker might be represented and remembered differently than if you were playing the hand yourself. Such differences could arise due to the rapidly changing nature of goals as one plays the game leading to different attentional settings (Cowan et al. 2024, Wang and Egnér 2022). This reveals an important gap in our understanding of event segmentation and its effects on memory: to what extent do theories such

as EST and its body of supporting evidence extend to dynamic, active environments where individuals are making choices to resolve uncertainty about the events in question?” (page 3)

9. p.3: The link btw event segmentation and LCI has been made before in recent years, I'd recommend citing these papers/preprints (Franklin et al., 2020; Shin and DuBrow, 2021; Varga, Morton and Preston, 2022; Bein and Niv, 2023; Lu et al., 2024; Buekers et al., 2024).

We thank the reviewer for highlighting this oversight. We have rewritten the discussion of latent cause inference in the introduction as follows:

“A parallel framework used to interpret event cognition is that of latent cause inference (Shin and DuBrow 2021). In latent cause inference, people are thought to generate and test hypotheses about the state of the world. Here, as in EST, prediction errors can also play a key role signaling the need to reevaluate hypotheses. This suggests an intriguing analogy between the hypotheses formed during latent cause inference and event models in EST. For example, hypotheses about the environment lead to sets of predictions that can be actively tested, which is akin to testing candidate event models (Shin and DuBrow 2021). Though such mechanisms have been proposed (Franklin et al., 2020; Shin and DuBrow, 2021; Varga, Morton and Preston, 2022; Bein and Niv, 2023; Lu et al., 2024; Beukers et al., 2024), specific experimental evidence for this claim is currently limited. The primary benefit of the latent cause inference approach is that it has explicitly tested how individuals interact with their environment and resolve uncertainty (Niv et al. 2015; Wilson and Niv 2012; Mirea et al. 2024; see Radulescu, Shin, and Niv 2021 for review). Inspired by this work, we seek to bridge these parallel lines of research by asking how actively inferring latent states influences event segmentation, and how this, in turn, affects long-term memory.” (page 4)

As well as in the discussion:

“...Future experiments could aim to better disentangle latent cause inference and reinforcement learning to enable different modeling approaches, as previous work has highlighted the strong connections between latent cause inference and event cognition (Franklin et al., 2020; Shin and DuBrow, 2021; Varga, Morton and Preston, 2022; Bein and Niv, 2023; Lu et al., 2024; Beukers et al., 2024).” (page 19)

10. p.4: Unclarity in the line of argument: the authors talk about action not being studied in previous event segmentation studies, then say that LCI is a way of actively inferring different states while interaction/hypotheses testing, then discuss alternating goals, followed by previous studies didn't have a component of uncertainty – what's the theoretical line here and what are the authors trying to test? How are all of these: LCI, goals, uncertainty, action, related? I believe this can be laid out much more coherently in the introduction, especially for the non-expert reader, and that would strongly benefit the paper.

We thank the reviewer for this comment and agree that the work would benefit from increased clarity here. In response to this point as well as a few others we have substantially reconfigured

the introduction to highlight the core contribution of the paper as well as to better lay out the interconnection of all of the disparate topics we invoke.

11. p. 5: Figure 1, the section of “Pseudo-random rule transition” is unclear because boundaries are between each letter, so it’s unclear what the lines between every four letters represent. Figure 1, please indicate in the caption what “Y/N” means, and what +0,+1 mean. I assume it’s the response and the reward, but that would be better clearly laid out.

We thank the reviewer for catching this. We have adjusted figure 1 to be clearer and reproduce it here for convenience along with the updated figure caption to better clarify what is meant by “Y/N” and +0 and +1.

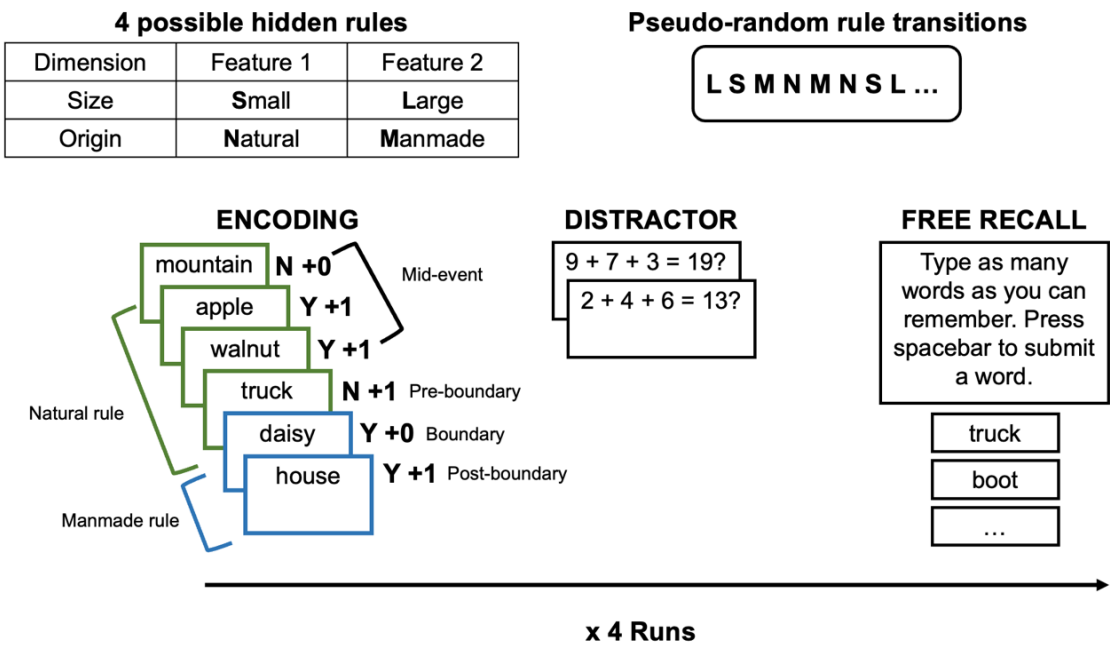


Figure 1: Experimental design. In the WRIT, participants indicate whether a word agrees with one of four hidden rules. The active rule repeats across events that last 6-8 trials. Then, it transitions pseudo-randomly such that each rule is experienced an equivalent number of times in each run. Here the word “daisy” marks the change from the natural rule (green) to the manmade rule (blue) (color added as aid, not shown to participants). The capital “N” or “Y” next to each box in the ENCODING section denotes the participant’s response while the “+0” and “+1” denote received reward. After performing the word rule inference task, participants perform a distractor task for 10 seconds and are then asked to perform typed free recall. They perform this sequence 4 times with run-unique words each time. (page 5)

12. Performance model: Were runs and position relative to boundary treated as factors? If so, please mention, if not, what do the authors mean by having the first run/initial trial treated as the reference? Is it just that these were modeled as 0?

We thank the reviewer for these questions. Position relative to boundary is a value that takes its value based on where the nearest subjective boundaries occurred. If the nearest subjective boundary is before the trial, it will have a positive distance. If the nearest subjective boundary trial is after the trial of interest, it will have a negative distance. In the case where the trials are equidistant from a previous and next boundary we default to a positive value. We have altered the text in the methods to enhance clarity about the treatment of factors as well as what is meant by the reference:

“For the run variable, the first run was modeled as 0 and each further run incremented by 1. This enabled us to examine the degree to which participants improved at the primary task across runs. For the position relative to boundary, the position was the number of trials between a given word and the nearest boundary. It was positive if it was preceded by the nearest boundary and negative if it occurred before the nearest boundary.” (pages 8-9)

13. RT model:

1. Is trial_within_event the same as position_relative_to_boundary in the performance model? Please clarify.

We thank the reviewer for flagging the lack of clarity here. Trial_within_event treats the first trial in a block as the 0 index, and then increments by 1 for each subsequent trial, whereas the position_relative_to_boundary contains negative and positive values as it treats the boundary timepoint as the 0 index. We swapped to using trial_within_event as the regressor for the RT analyses, because the position_relative_to_boundary term is collinear with the boundary_label factors of interest (position relative to boundary 0 is Boundary, 1 is Post-boundary, -1 is Pre-boundary, otherwise Non-boundary).

2. While I understand why one would include a linear trial within-event predictor and a 1(post-boundary) predictor in the same model, I’m not sure how to interpret that interaction term. If all not-post-boundary items are coded as zero, and all post-boundary items are by definition “1”, wouldn’t the interaction predictor be identical to the 1(post- boundary) predictor? Please provide more details.

We thank the reviewer for this question. As we have now combined the RT models as they previously suggested the 1(post-boundary) predictor is no longer used. However, to clarify there are cases where the trial_within_event was equal to 1 and it was not marked as a post-boundary. This has to do with how we quantified subjective boundary points whereas the trial_within_event uses the objective shifts in latent states.

14. Post-error model: what do the authors mean by “position relative to the event boundary”? The model does not include such a predictor. Please clarify.

We thank the reviewer for flagging this confusing switch in use of our terms. We have altered the text to be more explicit:

“Finally, the boundary label variable is a categorical regressor, where the item’s position is coded such that the trial at $t + 1$ after the boundary is called the “post-boundary” trial, $t - 1$ is

called the “pre-boundary”, and all other trials are marked as “non-boundary”. Non-boundary is modeled as the base factor. We log transform the RT data according to recommendations by Lo and Andrews (2015).” (page 9)

15. p. 10: WRIT accuracy: if the position is relative to the boundary, and accuracy increases as the position relative to the boundary increases, why is the beta estimate negative?

The beta estimate is negative due to the coding of pre-boundary items as negative and post-boundary items as positive. We have included clarification of this effect in the results:

“A linear mixed effects model showed that response accuracy decreased following a rule shift compared to the stable performance they achieved before, and then increased across the event ($\beta_{\text{position_relative_to_bound}} = -0.115$, $CI_{95\%} = [-0.15, -0.08]$, $p < 0.001$) (Figure 2A). This effect is negative as the items prior to the boundary have a negative coding and the items after a boundary have a positive position.” (page 10)

16. p. 10: WRIT response time: please be consistent between the Methods and Results sections. Is trial_within_block the same as trial_within_event?

We appreciate the reviewer catching this error. We have corrected all occurrences to read “trial_within_event”.

17. p. 10: Previous work linked higher RT at the boundary to poor memory across events (Heusser et al., 2018). I wonder if there’s some analogous analysis here, e.g., the clustering, that could further corroborate the relationship between inferred boundaries/RPE and memory.

We thank the reviewer for highlighting this interesting analytic possibility. We have run a Spearman rank correlation to assess the degree that block-level clustering in recall correlated with post-boundary RT and boundary RT. We found no evidence that either of the response time measures correlated with event-level clustering in recall.

18. p. 13: “but that transitions across events tended to occur in the forward direction.” – what’s the evidence for that? A couple of sentences before, the authors said that they did not find evidence for asymmetry.

We apologize for the lack of clarity here. The forward asymmetry is not present at a lag of 1 but overall, the positive event lags are more common than negative ones as we state earlier in that paragraph. We have adjusted that paragraph to enhance clarity. It now reads as follows:

“We found no evidence for forward asymmetry for the recall of next event items at lag 1, though this could be due to a lack of power to test this specific comparison (test of lag 1 vs lag -1 $t(65) = 1.52$, $Cohen's d = 0.27$, $p = 0.134$). However, when examining all forward vs backward across-event transitions we found that transitions tended to occur in the forward order (test of negative lags vs positive ones $t(65) = 7.1$, $Cohen's d = 1.28$, $p < 0.001$. In sum, these

analyses revealed a tendency to stay within an event during recall, and that transitions across events tended to occur in the forward direction.” (page 15)

19. While many studies reported an effect of boundaries on temporal memory, there has been a previous study finding no effect on recall (Heusser, 2018). I understand that EST might have predicted improved recall. But, I think it should be mentioned that this lack of better recall of boundary items was reported before.

We thank the reviewer for catching this, we have included a statement to this effect here:

“Boundary items were not recalled more often than non-boundary items ($t(65) = -1.65$, *Cohen's d* = -0.19 , $p = 0.103$). Pre-boundary items were however, recalled more often than non-boundary ($t(65) = 2.22$, *Cohen's d* = 0.25 , $p = 0.030$). While lack of a boundary-specific enhancement has been reported before (Heusser et al. 2018), we found evidence that event boundaries impaired memory.” (page 14)

20. RL Model:

1. It is unclear what is the “value of the stimulus“. Isn't the value in the model calculated per choice? The participant does not reject or accept a stimulus. They answer a question about the stimulus, and the value is per the answer.
2. The value is updated for all the feature weights that were present in the stimulus if the choice was ‘yes’. Why not just the feature on which the participants responded?

We thank the reviewer for this opportunity to clarify the model.

The answer a participant provides of “yes” or “no” is equivalent to accepting or rejecting the stimulus on the screen as a fit for whichever rule may be active at the current time. We have added this clarification to the methods section here:

“This prediction is multiplied by the learning rate η , a free parameter, to update all the feature weights that were present in the stimulus if the choice was ‘yes’, or the weights that were not in the stimulus if the choice was ‘no’. We update the value of all the feature weights associated with the stimulus or its inverse as we cannot be sure of the exact feature the participant is responding to. The weights of the non-chosen features are decayed to 0 according to a decay parameter d , another free parameter (Niv et al. 2015; Rouhani and Niv, 2021).” (page 7)

In response to the second point, we cannot update just the feature which the participants responded as we only have a yes or no response, so we have no way to know for certain whether they are saying “yes” for a size or an origin judgment. Therefore, we update both of the feature weights at once.