Research Article

Psychological Science
1–19
© The Author(s) 2023
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09567976221128206
www.psychologicalscience.org/PS

# Evidence That Event Boundaries Are Access Points for Memory Retrieval

**Sebastian Michelmann[1], Uri Hasson[1,2], and Kenneth A. Norman[1,2]**
[1]Princeton Neuroscience Institute, Princeton University, and [2]Department of Psychology, Princeton University

## Abstract

When recalling memories, we often scan information-rich continuous episodes, for example, to find our keys. How does our brain access and search through those memories? We suggest that high-level structure, marked by event boundaries, guides us through this process: In our computational model, memory scanning is sped up by skipping ahead to the next event boundary upon reaching a decision threshold. In adult Mechanical Turk workers from the United States, we used a movie (normed for event boundaries; Study 1, $N = 203$) to prompt memory scanning of movie segments for answers (Study 2, $N = 298$) and mental simulation (Study 3, $N = 100$) of these segments. Confirming model predictions, we found that memory-scanning times varied as a function of the number of event boundaries within a segment and the distance of the search target to the previous boundary (the key diagnostic parameter). Mental simulation times were also described by a skipping process with a higher skipping threshold than memory scanning. These findings identify event boundaries as access points to memory.

## Keywords

episodic memory, event cognition, memory search, memory replay

Received 4/19/22; Revision accepted 8/30/22

Everyday human experience is information rich and extends over time, which poses unique challenges on our memory system: When accessing specific details (e.g., where we placed our keys), our brain needs to use an efficient strategy to sift through long periods where not all information is relevant to the memory search. What are the strategies that the brain uses when accessing and searching through memories of naturalistic stimuli (i.e., temporally continuous, information-rich material that contains a coherent narrative and event structure; Sonkusare et al., 2019)?

In naturalistic experience, we can identify high-level cognitive constructs—"events"—that chunk memories into meaningful units (e.g., a "restaurant event" or a "phone-call event"; Zacks et al., 2007). This construct "event" is different from the colloquial use of the word "event" (meaning "occurrence"); it can be described as the representation of a situation (Bartlett, 1932). Indeed, a neural substrate has recently been proposed in rodents, where the firing of "event cells" that represent laps in a maze is invariant to lap duration (Sun et al., 2020). Event structure of experience has consequences for how we remember: Participants' ability to detect event boundaries is associated with episodic memory performance (Sargent et al., 2013), and information learned in a previous event (e.g., before walking into a new room) suffers from more forgetting than information that is learned within the same event (i.e., the "doorway effect": Radvansky & Copeland, 2006). Further, memory-based duration judgments are increased by event boundaries under constant clock duration (Ezzyat & Davachi, 2014; Lositsky et al., 2016); for recent reviews of how event boundaries shape behavior and brain activity, see Clewett et al. (2019) and Brunec et al. (2018). Moreover, duration judgments can be

**Corresponding Author:**
Sebastian Michelmann, Princeton Neuroscience Institute, Princeton University
Email: s.michelmann@princeton.edu

modeled by presenting videos to a feedforward image classification network and tracking changes in the model's internal state (akin to event boundaries; Roseboom et al. 2019).

Events may therefore structure how we replay past experience in our mind's eye: When Jeunehomme and D'Argembeau (2020) had students walk around campus with wearable cameras, participants later took less time to recall the experience compared with its actual duration (temporal compression), but the amount of detail they recalled increased with the number of identified events in camera images. This suggests that events, rather than absolute time, are the unit of experience in memory, potentially enabling compressed replay. Note that despite evidence for the importance of events in memory, no study has directly addressed how event structure is used in the retrieval process, that is, what is the functional role of event structure? Studies in which participants have been asked to mentally simulate continuous memories (e.g., navigated routes) corroborate the importance of events in memory: When the event structure of experience was manipulated, controlling for other factors, the (temporally compressed) mental simulation duration took longer when experiences contained more events (Bonasia et al., 2016; Faber & Gennari, 2015). Some behavioral evidence also points to the idea that event boundaries could be used to access memories. In serial recall tasks of material that spanned event boundaries, items at event boundaries were more likely to be recalled out of order relative to control items (DuBrow & Davachi, 2016; see also Heusser et al., 2018, for a statistical trend suggesting preferred transitioning to boundary items during free recall). Note that these latter investigations are evidence that event memory can be studied without using naturalistic stimuli; that is, naturalistic stimuli do not recruit processes that are qualitatively inaccessible to other studies. However, we believe that naturalistic video stimuli are particularly well suited for assessing the role of event boundaries in the recall process, given the presence of rich event structure in these stimuli.

A magnetoencephalography (MEG) study in humans (Michelmann et al., 2019) explored memory scanning across artificial scenes that had some properties of naturalistic experience (information rich and temporally continuous). The study provides indirect indications of a mechanism of temporal compression and scaling by the number of events: Participants studied unique word cues superimposed on scenes in short video clips (each consisting of three scenes). During retrieval, participants decided whether they had seen each word in the first, second, or third scene. Behaviorally, participants were faster to recall words from earlier (vs. later) scenes, indicating forward memory scanning. Neural patterns, however, indicated that—although replay was compressed—forward replay within scenes proceeded at a slower rate than the overall level of compression would suggest. Michelmann et al. (2019) argued that participants speed up the memory-scanning process by skipping ahead to the next scene boundary. This explains compression (because participants skip) and slower replay within scenes than across multiple scenes (because replay across multiple scenes is sped up through multiple skips). A comparative study with macaque monkeys uniquely linked this dynamic replay strategy to humans (Zuo et al., 2020).

## Statement of Relevance

In our day-to-day lives, we routinely search through long periods in memory—for instance, when figuring out when we last had our keys. Does the structure that we perceive in our experience help in this process? Specifically, humans perceive events and boundaries between them (e.g., picking up the phone marks the beginning of a "phone-call event"). We hypothesized that people can speed up memory search by skipping directly to the next event boundary if the current event is very different from the memory being sought. To test this, we prompted human participants in large-scale online experiments to search their memory of a movie, and we measured the amount of time it took them to locate the sought-after memory. In line with our hypothesis, we found that search time can be explained using a model in which participants skip through all events except the last one, which needs to be played through in its entirety to find the sought-after memory that it contains. These results suggest that event boundaries can act as stepping stones to facilitate memory search.

The Michelmann et al. (2019) study explored memory scanning for artificial, loosely related scenes. Neural results were in line with a skipping mechanism; however, direct evidence is missing. Here, we used behavioral experiments and modeling to assess whether naturalistically occurring event boundaries function as access points in memories of naturalistic stimuli. Specifically, we hypothesized that when people retrieve continuous memories, they access them from naturalistically occurring event boundaries. Therefore, when people scan memories of naturalistic stimuli (e.g., trying to remember where they placed their keys), they can skip ahead in memory to the beginning of the next event (i.e., the next event boundary). To decide when it is time to skip, people leverage the similarity of the target of memory search (here, the keys) to each scanned moment in memory. When the total observed dissimilarity in the current event exceeds a threshold value, a decision is made to stop scanning and skip the rest of the current event.

We tested this idea in large online experiments: Participants watched movies and then rapidly answered questions that required memory scanning of segments from the movie. Across questions, segments varied in duration and contained different numbers of events. Our computational "stepping-stones" model predicted that scanning duration would depend on (a) the number of events leading up to the target stimulus (each partially scanned before skipping) and (b) the temporal distance of the target from the preceding event boundary—this final segment did not itself contain event boundaries and needed to be scanned without skipping (see Fig. 1). We identified the temporal distance of the target from the preceding event boundary as the key diagnostic parameter necessary to arbitrate between our model and alternative models, where memory scanning could, for instance, skip directly to the target of memory search on the basis of semantic relatedness or to other moments that were unrelated to event boundaries (e.g., random moments within the event). The rest of the article is composed of the following parts. In a first study (study 1), we normed a movie (adapted from *Gravity*; Cuarón, 2013) to determine where event boundaries are perceived. Second, from a computational model of continuous memory search that can access memories from event boundaries and skip ahead, we derived the parameters that (according to the model) should be diagnostic of memory-scanning duration. Third, in a large-scale online experiment, participants scanned segments from the movie in memory (study 2); we tested whether memory-scanning times were explained by the parameters identified by the model. Fourth, another large-scale online experiment (study 3) tested the model by asking participants to perform thorough memory scanning by mentally simulating segments from memory. All of the studies were reviewed and approved by the Princeton University Institutional Review Board. Approval included adherence to the legal requirements of the study country.

## Open Practices Statement

The data from the behavioral studies and analysis scripts are publicly accessible via Zenodo (https://doi.org/10.5281/zenodo.6972620). The studies were not preregistered.

## Event Boundary Norming

To investigate the influence of event boundaries on continuous memory scanning, we asked where participants typically perceive event boundaries within a movie. This section describes a norming study (study 1), in which participants mapped out the event structure of the movie. This information about event structure was used in all subsequent experiments in which participants performed different tasks.

### Method

***Stimulus material.*** The material consisted of two movies of 7-min 30-s duration that were gathered from the movie *Gravity* (Cuarón, 2013). The edited movies told a coherent story of 15-min duration. Although we cannot share these movies because of copyright issues, interested readers can view the original movie. The edited version used in this experiment resembled an extended trailer that spanned the whole movie. The two movie clips were further compiled so they spanned several naturalistic events: Examples are moments in which the protagonists are suddenly confronted with danger, enter or leave space stations, or set out to travel toward a new destination. The editing also aimed to preserve as much suspense as possible from the original movie, which was thought to help task engagement and memorability of the story.

***Data collection.*** A first sample of 104 participants was collected online using *Inquisit* (Millisecond Software, www.millisecond.com). The movies were reduced to a low-resolution version (320 × 180 pixels) to allow for seamless presentation via *Inquisit*. Another sample of 99 participants was collected on a custom-configured machine running *psiTurk* (Eargle et al., 2020; Gureckis et al., 2016). The total sample size was determined on the basis of a previous study that performed the norming for event boundaries in a naturalistic story (Michelmann et al., 2021); using a similar sample size ($N \approx 200$) was therefore considered appropriate to get reliable estimates on where event boundaries in the movie are located (see the Data Exclusion and Data Preparation section for additional considerations regarding sufficient sample size). In the second sample, movies were presented at a resolution of 1,280 × 720 pixels.

All participants were recruited via Amazon Mechanical Turk (Buhrmester et al., 2011). Cloudresearch (formerly TurkPrime, www.cloudresearch.com) was used to prevent workers from participating if they had a low overall acceptance rate across studies (< 80% of experimenters accepted participation as valid), to restrict data collection to the United States, and to facilitate payment and assignment of bonuses (Litman et al., 2017). Participants received monetary compensation of $3.50 ($7 per hour) for their participation in the experiment and an additional $1 performance-based bonus. There was no instruction on performance contingencies (i.e., how the bonus related to performance). The bonus was

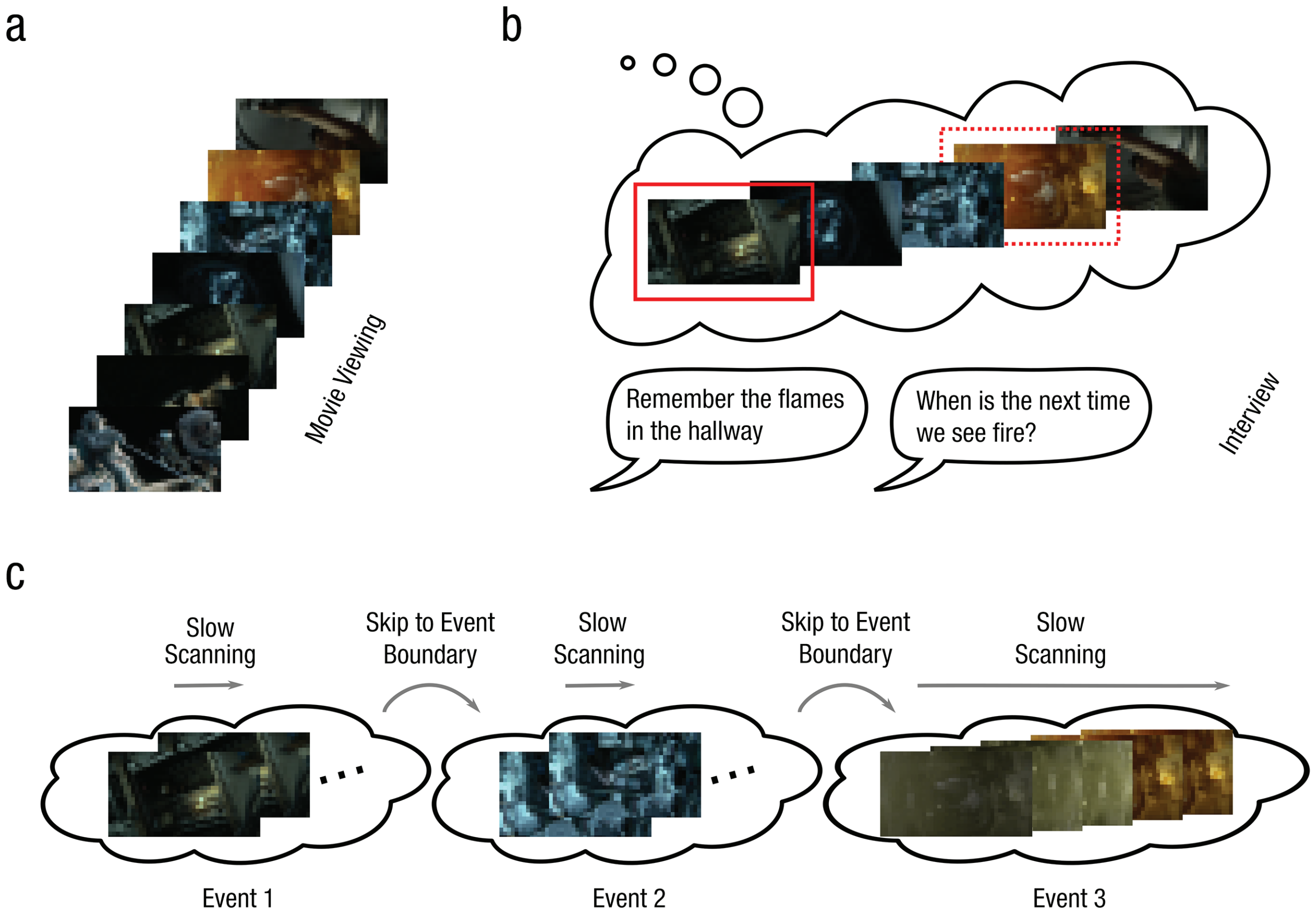

**Fig. 1.** Continuous memory scanning (pixelated still shots). Participants watched a short movie (a) and were subsequently asked questions about the movie (b). Note that still shots from the movie are pixelated here for copyright reasons. Each question started by describing a specific incident from the movie, thereby orienting participants to that moment. Subsequently, participants were asked about a later moment from the movie (typically, "When is the next time that…"). When participants knew the correct answer, they clicked a button labeled "respond" and then typed a description of that moment. Our hypothesis (c) was that participants would scan their memory of each event in the movie at a slow speed; when they decided that the target was not in the current event, they would skip ahead to the next event boundary. In the final event, scanning would proceed until the target was found. Under these assumptions, the number of event boundaries and the distance of the target to the previous event boundary would predict the memory-scanning duration.

mentioned in the advertisement and in the consent form, using the following wording: "You may receive a monetary bonus of up to $1 for the successful completion of the task." Only participants that did not respond at all throughout the whole experiment were denied the bonus.

***Procedure.*** In this boundary-norming experiment, we asked participants only to mark event boundaries within the movies. After providing informed consent, participants were told that they would watch two short movies. They were instructed to press the space bar whenever, in their opinion, one natural and meaningful unit ended and another began. A black dot was displayed above the movie for feedback whenever the space bar was pressed. After completing the first movie, they could take a short break and were subsequently asked to perform the same task on the second movie. Movies were presented in the same order to all participants.

***Data analysis.***

*Data exclusion and data preparation.* Data were analyzed in MATLAB (Version 2019a; The MathWorks, Natick, MA) using custom scripts. For the second behavioral sample, data were excluded if the effective duration of the movie presentation was 1.5 s more than the actual duration of the movie. These uncertainties were due to occasional lag in the movie presentation because of the higher resolution of the movies. A total of 33 data sets were excluded for the first movie, and 28 data sets were excluded for the second movie. Data were then aggregated in response vectors at a millisecond resolution.

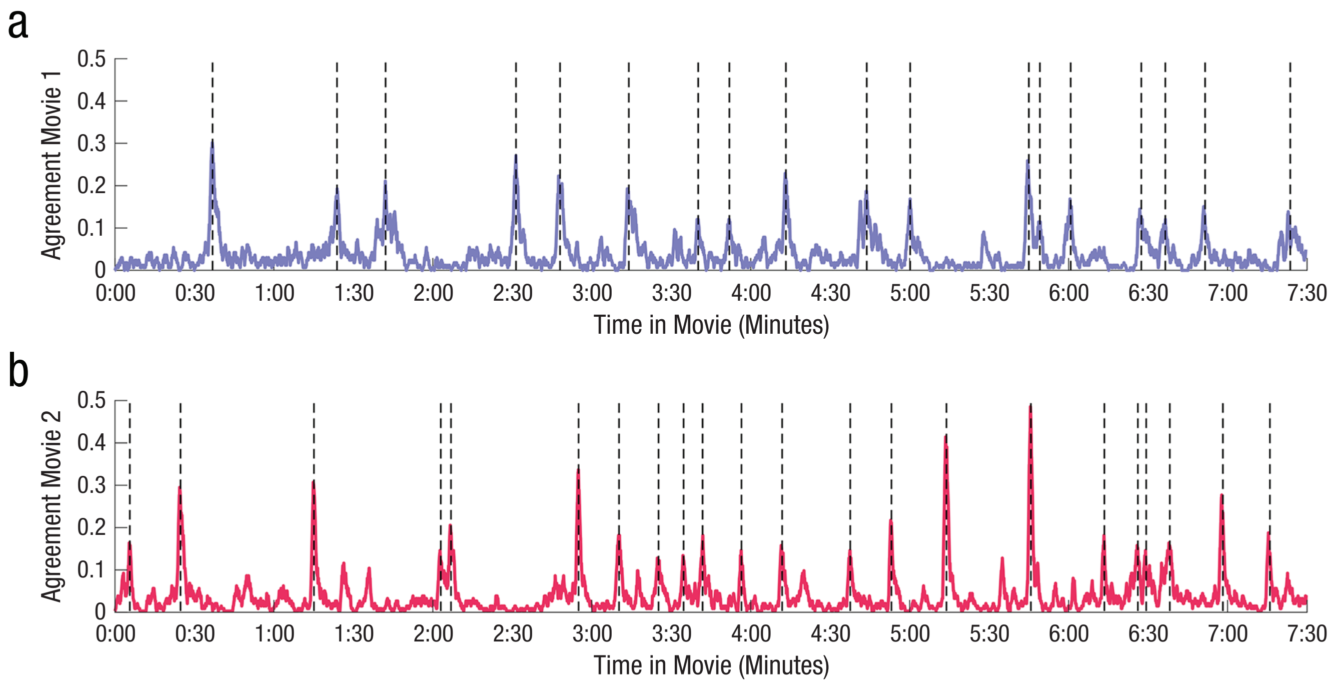

**Fig. 2.** Event boundaries in the (a) first and (b) second movie. Participants watched two movies that formed part of a coherent story and were asked to press a button whenever, in their judgment, one natural and meaningful event ended and another began. Local peaks in the time course of agreement between participants mark event boundaries for each movie.

Vectors were set to 1 at a given time point if a participant had pressed the space bar within the 1 s surrounding that moment and were set to 0 otherwise. This procedure is in line with the typical aggregation of responses at a 1-s resolution (Zacks et al., 2009) while maintaining millisecond precision of averages.

We next assessed whether it would be justified to combine the two samples. The average concatenated response vector in the first behavioral sample was highly similar to the average vector in the second sample (cosine similarity = 0.85), suggesting that our measure of boundary perception was reliable; we therefore decided to combine the two samples, resulting in 170 data sets for Movie 1 and 175 data sets for Movie 2. Given the high similarity that was obtained between the two samples, we also concluded that a good approximation of the typical perception of event boundaries in the movie had been reached and decided to stop data collection for the norming sample (i.e., not to replace excluded participants). To further improve the data quality, we subsequently excluded participants who responded atypically. This was done by comparing each participant's individual response vector with the average response across all other participant. If a participant's response vector for a given movie exceeded a cosine distance of 0.9, the vector was excluded from the average; this reduced the sample size to 166 for the first movie and 167 for the second movie.

*Boundary definition.* Boundaries were defined as local peaks in the average response vector across all participants. To identify these peaks, we first smoothed the vectors with a gaussian kernel of 3-s window width. Subsequently, the data were thresholded at the 90th percentile and grouped in clusters of neighboring points that exceeded this threshold. Each cluster's maximum was then taken as a local peak that marked an identified event boundary (cf. Michelmann et al., 2021).

## *Results*

The average response vectors revealed substantial agreement on event boundaries between participants (Fig. 2), with up to 50% of participants pressing the space bar within the same second on one occasion (note that aggregating across a broader time window increased agreement at the expense of temporal precision). A total of 18 event boundaries were identified as local peaks in the first movie, and 22 event boundaries were identified in the second movie.

## Stepping-Stones Model of Continuous Memory Scanning

We specify our hypotheses in a model that is concerned with sequential search through a memory. The model searches in a forward direction but uses the event structure of the memory to speed up its search: If the current event is sufficiently dissimilar from the target pattern, the model skips to the beginning of the next event.

## Method

The decision to skip ahead is modeled analogously to a drift-diffusion process with a single decision boundary: Evidence for the absence of the target is accumulated within each event until a certain threshold (the "skip threshold") is passed (Nosofsky & Palmeri, 2015; Ratcliff & McKoon, 2008). Our model formulation relied on three key assumptions. First, elements that form part of the same event are similar to one another. Second, event boundaries are access points for memory retrieval, and therefore the sequential search can skip ahead to the next event boundary. And third, sequence memory enables the sequential access of event boundaries

The first assumption is based on the prominent finding that—in high-level brain regions that represent events—neural patterns are more similar (on average) within events than across events (Baldassano et al., 2017; Baldassano et al., 2018; Geerligs et al., 2021; Geerligs et al., 2022). Note that the brain may concurrently represent other kinds of similarity, where (for instance) moments in different events with the same actor are represented as more similar than moments in the same event; however, our model is concerned with high-level representations where moments in an ongoing event are represented as being (on average) relatively similar to one another and relatively distinct from other events. The second assumption is supported by the findings that suggest that skipping to the beginning of new events can speed up the replay process (Michelmann et al., 2019)—which we discussed in detail in the introduction. The third and final assumption is based on evidence that humans represent sequential information in memory (for an overview, see Bellmund et al., 2020); we are agnostic about the mechanistic implementation of this ability to remember event boundaries in sequential order.

### *The memory sequence.*

*Elements.* A memory sequence is modeled as a sequence of elements; a single element is defined as a vector $\vec{x}_t \in \mathrm{R}^{100}$, where $t$ denotes the time point. Elements can be thought of as a moment of experience; they are the most fine-grained temporal unit of a continuous episode in the model and realize a discrete approximation of continuous experience. The number of elements that form an episode of fixed duration therefore determines the effective sampling rate of the memory sequence and was set arbitrarily with computational feasibility in mind.

*Events.* On the basis of the first assumption of the model—similarity within events—we define events via a correlation among their corresponding element vectors. Intuitively, the similarity structure of the memory sequence can be understood like a time-by-time representational similarity matrix (Kriegeskorte et al., 2008), where moments within the same event are similar to one another and moments from different events are dissimilar. Neurally, this representational structure has been observed in multivoxel patterns of functional MRI data (Baldassano et al., 2017; Geerligs et al., 2021). Correlated patterns are achieved by sampling the element vectors within each event $e$ from the same multivariate normal distribution $N(\vec{\mu}_e, \Sigma_e)$, where $\Sigma_e$ was generated by applying the eigenvectors of a random normally distributed symmetric matrix to a diagonal matrix $D$ with $\{d_{ij}, i = j \in R \mid 0 \leq d_{i,j} \leq 1\}$. We sampled the mean event vector $\mu_e$ $\{\mu_{ie} \in R \mid 0 \leq \mu_{ie} \leq 1.5\}$ from a wider range (0–1.5) to increase the strength of correlation within an event $e$ (a wider range of values in $\mu_e$ increases the variance within the mean pattern $\mu_e$; consequently, if $\Sigma_e$ is kept constant, two samples that are drawn from a multivariate distribution around $\mu_e$ will have a higher correlation).

*Target.* The target $\vec{y}$ is the element that defines the end of the memory search. It is first generated as part of the memory sequence and sampled from the multivariate normal distribution that defines the final event in the search. Once generated, the target is taken as is and is compared in the memory search with elements that form the memory sequence.

***The search process.*** The search process is defined as a sequence of comparison operations in which a target vector $\vec{y}$ is compared with elements within each event (Fig. 3a). Specifically, the cosine distance ($1 - \cos\theta$) between the target vector and the current element vector is computed (note that other distance metrics could have been used interchangeably, e.g., Euclidean distance). The search process ends when the target is reached, that is, when the target is compared with itself, which results in a cosine distance of 0. Importantly, the model is concerned with speeding up this sequential search process so that not all elements have to be compared with the target. Therefore, on top of the comparison operation, a decision process is modeled that realizes the second assumption of the model: The sequential search can skip ahead to the beginning of a new event. To make this decision, the model accumulates the cosine distance to encountered elements within the event and subtracts a bias $b$, that is, at every time point $t$ within an event $e$, the decision criterion $d$ is computed as follows:

$$d_{t,e} = d_{t-1,e} + \left(1 - \cos\theta\right) - b, \tag{1}$$

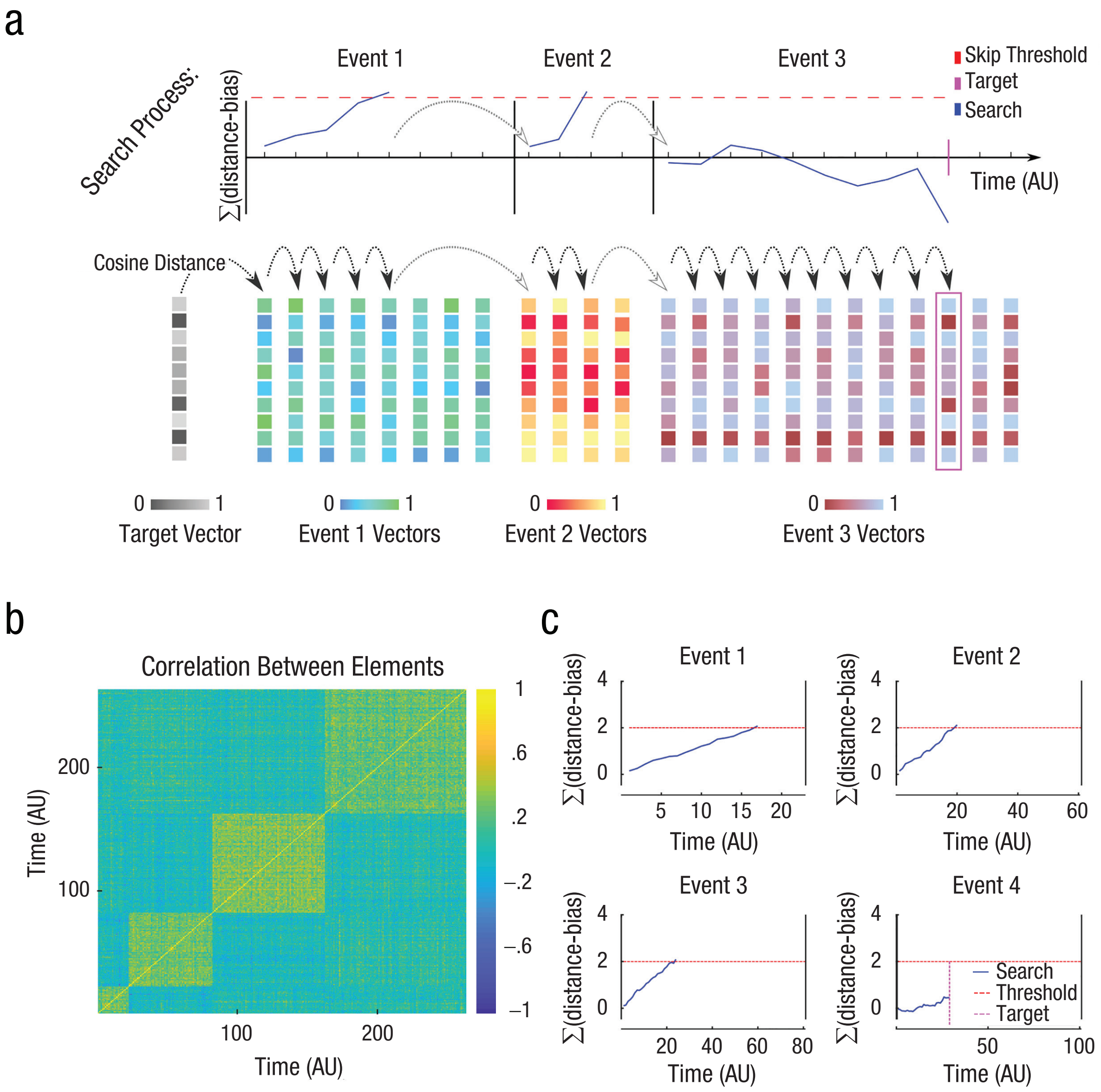

**Fig. 3.** Model of memory scanning. (a) A memory sequence consisted of element vectors that were correlated within events. In the search process, a target was compared with each element and dissimilarity was accumulated. If a threshold was passed, the model skipped to the beginning of the next event. In the final event, elements were similar to the target and dissimilarity therefore did not typically accumulate past the threshold. (b) Element-by-element correlation matrix for a simulated trial of four events. Correlation within events is larger than correlation across events. (c) Search process through the trial from (b). In Events 1 through 3, dissimilarity (blue) exceeds the threshold (red) after several time steps. In the final event, the search continues until the target (pink line) is found. AU = arbitrary units.

with $d_{0,e} = 0$. If $d_t$ exceeds a predefined threshold (i.e., $d_t > d_{\text{crit}}$), then the search process continues in the next event. In the final event, the similarity between the target $\vec{y}$ and other elements from the event causes the cosine distance to be low at each step of the sequential search, such that typically

$$1 - \cos\theta \leq b. \tag{2}$$

Consequently, in the final event, the search process typically continues until the target is reached. In the unlikely but possible event that the model skips in the final event—and therefore misses the target—this trial is noted as a miss (i.e., the correct target was not found and the trial is excluded from analysis).

***Response time computation.*** The response time (RT) of the model in a given trial is defined as the absolute number of steps (comparison operations) that are made in the transition from the start to the target. The steps within a given event $t_{\max,e}$ are the number of time points up to the moment where $d_t > d_{\text{crit}}$, or the moment where the target is found, i.e.,

$$t_{max,e} = \begin{cases} \arg\min_t \{d_{t,e} \mid d_{t,e} > d_{crit}\} & \text{if } e < n \\ \arg_t \{x_{t,e} \mid x_{t,e} = y_{t,e}\} & \text{if } e = n \end{cases} \tag{3}$$

The RT then amounts to $rt = \sum_{e=1}^{n} t_{max,e}$

***Model predictions.*** Through this computation of RTs, the model formalizes the following prediction: The time that it takes to scan a memory sequence for a target is explained by the distance of the target to the previous event boundary, together with the number of event boundaries in the memory sequence (i.e., the number of skips that the model makes). Predicting RTs on the basis of these two factors should outperform a model that takes into account only the total length of the scanned memory sequence, that is, the distance of the target to the beginning of the whole memory sequence that is scanned. This memory sequence in the model corresponds to movie segments that are defined by two moments within the movie (moments are actions or occurrences in the movie that can be described or asked for in a question); we used those segments later to prompt memory scanning or mental simulation (note that these movie segments can encompass several events). The total length of the model's memory sequence corresponds to the duration of such a segment in seconds.

## Results

***Simulation of a data set.*** To derive testable predictions about response data from our model, we simulated data for 50 participants, each performing memory scanning of 20 random trials, using MATLAB (Version 2020b; The MathWorks, Natick, MA). Note that simulating a small sample is sufficient under the high signal-to-noise ratio in simulated data. The bias term of the model was set to $b = 0.45$, and the skip threshold was set to $d_{\text{crit}} = 2$. A trial consisted of a random number of up to 10 events, and each event consisted of a random number of time samples (between 1 and 100 $\in$ N). For illustrative purposes, the final event length was set to 100 time samples, and a random element within that event was selected as the target (determining the end of the search). A participant-specific random intercept (1 − 100 $\in$ R) was added to all 20 trials of a given participant and a random noise term (±5 $\in$ R) was added to each trial (see Fig. 3b for the time-by-time correlation matrix of an example trial and Fig. 3c for the search process through that trial).

***Analysis of simulated data.*** Out of 1,000 trials, 32 were excluded from analysis because the model skipped in the final event before the target was found. A linear mixed-effects model (LMEM) was then fitted (restricted maximum likelihood estimation) onto the generated data using *RStudio* (Version 1.2.1335, www.rstudio.com) with the *lme4* package (Version 1.1-28; Bates et al., 2015). Corrected *p* values were derived via the *lmerTest* package (Version 3.1-3; Kuznetsova et al., 2017) approximating degrees of freedom via Satterthwaite's method. These analyses were meant to validate that data generated by the model showed the predicted effects (i.e., RTs were explained better by the number of event boundaries and the distance from the target to the previous event boundary, compared with duration alone). The LMEM was specified to explain RTs as a function of the number of event boundaries *(nEB)*, the distance of the target to the previous event boundary *(distEBpre)* and a participant specific intercept ($\text{sj}_{\text{id}}$):

$$\text{RT} \sim 1 + nEB + distEBpre + (1 \mid \text{sj}_{\text{id}}). \tag{4}$$

An effect of *nEB* confirmed that more events result in longer RTs of the model ($\beta = 22.802$, $SE = 0.558$, $t = 40.849$, $p < .001$), and an effect of distance to the previous event boundary (*distEBpre*) confirmed the model's key prediction that the distance of the target to the previous event boundary mattered ($\beta = 0.976$, $SE = 0.055$, $t = 17.614$, $p < .001$). We then refitted the LMEM using maximum likelihood estimation in order to compare it with another LMEM that took only the absolute segment duration (*dur*) as a predictor. The alternative LMEM formulation was as follows:

$$\text{RT} \sim 1 + dur + (1 \mid \text{sj}_{\text{id}}). \tag{5}$$

Because the alternative LMEM had fewer parameters, we used the Akaike information criterion (AIC) for comparison. Despite being penalized by AIC for having more parameters, the LMEM prediction based on the number of event boundaries and the distance of the target to the previous event boundary provided a substantially better fit to the data (ΔAIC = 78.8, ΔBayesian information criterion [BIC] = 73.9).

# Continuous-Memory-Scanning Experiments

We next wanted to prompt human participants to perform memory scanning through continuous memories of naturalistic stimuli (study 2). We presented participants with the movie material and subsequently asked interview questions (see Fig. 1) that first oriented participants to a specific moment in the movie and then elicited memory scanning for the answer.

## Method

In an initial sample, we collected data ($N = 80$) with a variety of questions (34 questions across two subsamples); we then collected a larger sample ($N = 100$) with 18 questions based on identified easy questions (participants had a high chance of answering these questions correctly because we wanted to maximize the number of correct responses in the new sample) and analyzed the time it took participants to find a response to the questions. To decide whether a participant's response was correct, we relied on independent samples of raters (total $N = 118$; for a thorough description of the rating process see the Rating Procedure section) who decided whether written responses described a given moment in the video that was concurrently presented to them on screen.

Considering the novelty of the hypothesized effects, we determined the sample size heuristically by taking into account previously reported sample sizes and by drawing on experience with data quality on Amazon Mechanical Turk (Michelmann et al., 2019, 2021; Rouhani et al., 2020). On the basis of these benchmarks from previous studies, we considered the total size of the sample to be appropriate for identifying psychologically meaningful effects.

**_Stimulus material._**

*Movies.* The video material consisted of the same two movies of 7-min 30-s duration from the movie *Gravity* (Cuarón, 2013), described in the event boundary norming study (study 1).

*Interview question properties.* For the first sample, we designed 34 unique questions that were split into sets of 17 across two sub-samples. Questions for the second sample were combined and adjusted from the previous 34 questions. Not all questions were compatible and could be in the same set together, that is, the setup description of one question might give away the answer to another question from the other set. Therefore, some new unique questions were combined from previous setup and target moments, resulting in 18 questions, seven new and 11 reused. Putting all of this together, we used 41 unique questions across the two samples. Every question had a beginning, defined by the moment described in its setup, and an end, defined by the moment described in its target (the beginning and end for each question were time stamped in the movies). Taking into account the identified event boundaries from the norming sample (Fig. 2), we therefore could identify the following properties of interest in a given question (which varied between questions): (a) the absolute duration of the segment in seconds, (b) the number of event boundaries between the beginning of the question and its end, and (c) the distance of the target to the previous event boundary (if no event boundary was crossed—seven of 34 questions in sample 1 and two of 18 questions in sample 2—the distance of the target to a previous event boundary was not defined).

Furthermore, we could identify (d) the distance of the target to the next event boundary (as a control property, expected to be of no relevance) and (e) the difficulty of a question (1 − the ratio of correct responses in the sample; see below) as a potentially relevant property of no interest. If a question started or ended in the vicinity of an event boundary, the number of event boundaries sometimes had to be adjusted manually (±1; 12 of 41 questions). Specifically, the delay in participants' responses in the boundary-norming task could cause a temporal misalignment, such that the empirically derived time point of the event boundary slightly extended into a new scene in the movie; in this case, the segment spanned by the question could erroneously include or exclude an event boundary. When this occurred, the property "number of event boundaries" for that question was corrected.

**_Experimental procedure._** After providing informed consent, participants received information about the movie and the characters to ensure that they would understand the plot. After that, instructions for the experiment were presented, in which participants were told that they would be asked a specific type of question: In these questions, they would first read a description of a moment in the movie (e.g., "[In the space station] we see little flames flying[...]into the hallway.") and then be asked about a different moment, typically introduced by: "When is the next time that . . ." (e.g., "we see fire?"; see Fig. 1b).

The instructions were followed by an example video of 9-s duration featuring Charlie Chaplin and by two practice questions that familiarized participants with the task. Participants then watched the first movie without interruption. After that, they briefly reviewed the task instructions and started the interview task. The questions that corresponded to the first movie were then presented in random order. A fixation cross was shown in the center of the screen for 1 s; then the setup moment was presented in written form, and participants could press a button labeled "next." They were instructed to make sure they knew which moment in the movie the setup referred to before they continued. After participants clicked this button, the target question appeared on the screen, and a button labeled "respond" became available on the bottom of the screen. Participants were instructed to press this button as soon as they knew the correct answer but not before that. When they clicked "respond," a text field appeared so they could type a description of the correct moment (i.e., the moment that the question asked about).

Participants were informed that their responses would be scored by another person, and they were encouraged to be precise and complete in their description. Participants were asked to type "I don't know the answer" if they did not find the correct moment. This was done to discourage clicking through the questions without much effort. When the questions about the first movie were complete, participants watched the second movie (because the movies formed part of a continuous narrative, the movies were always presented in order) and performed the same task; the corresponding interview questions once again were presented in random order.

Finally, participants rated their attentiveness, understanding of the movie, and understanding of instructions using a slider (left = 0, right = 100; participants only moved a slider, and values were assessed as 0 to 100 based on the final position of the slider). Participants' mean attentiveness rating was 93.806 (*SD* = 12.403; range = 65, 100 = fully attentive), mean rating of understanding of the movie was 78.961 (*SD* = 22.610, range = 90, 100 = fully understood everything), and mean rating of understanding of instructions was 15.867 (*SD* = 29.335, range = 100, 0 = fully understood everything; note that this scale was flipped, so the leftmost response corresponded to "fully understood everything"). We did not exclude entire participants because we observed that even participants who rated their understanding or attentiveness low were able to provide reasonable answers. Because our main analyses focused on correctly answered trials, and constant subject-level differences were already captured by the random intercept of LMEMs, we did not include these measures from the end of the experiment in our models.

***Rating procedure.*** To determine whether responses (total *N* = 3,160 typed text entries) were correct, we recruited a group of independent raters via Amazon Mechanical Turk. Raters were informed about the task and the instructions for the interview. They received the same introductory information about the movie and practiced the task with the two example questions following the Charlie Chaplin video. Raters were told that they would decide whether another person's description corresponded to the correct moment in the movie; this moment was presented as a video clip on screen while raters made their decision. Raters then watched the first movie before they rated participants' answers in batches of 10 participants per rater. After a fixation cross (200-ms duration), they first saw a setup description below a clip showing the corresponding moment in the movie. They had the option to replay the moment by clicking a "replay" button. After clicking "next," raters saw the target question below a video clip of the moment that represented the correct answer. Below the target question, a participant's answer was presented in red font with the introduction "This person answered:" A rater could then choose between the options "match," "mismatch," and "unclear." After the rater decided, a new response was presented under the same video clip and question. When all participants' responses to a given question were rated, a fixation cross and the next setup moment were presented. After rating all responses to the first movie, the rater watched the second movie. After that, they rated the corresponding answers. In addition to participants' answers, five fabricated wrong descriptions were included for each movie in order to assess the quality of a rater's assessments.

***Data collection.*** Data were collected on a custom configured machine running *psiTurk* (Eargle et al., 2020; Gureckis et al., 2016). All participants were recruited via Amazon Mechanical Turk. Cloudresearch (formerly TurkPrime, www.cloudresearch.com) was used to filter out participants with low acceptance rate across studies (< 80%), to restrict data collection to the United States, and to facilitate payment and assignment of bonuses (Litman et al., 2017). Participants, including raters, received monetary compensation of $7 per hour for their participation in the experiment and a performance-based bonus of up to $3 (bonus was increased from $1 after the first study). Videos were presented at a resolution of 1,280 × 720 pixels. The interview task was completed by an initial group of 80 participants who were tested on 34 questions. Some participants were tested or partially tested but did not follow the task instructions or gave no meaningful responses (*n* = 28). Of the 80 participants who completed the task successfully, 39 completed one set of 17 questions and 41 completed another set of 17 questions. Answers to these 34 questions were then rated by 49 raters for correctness (the rating was completed by a separate group of participants on Mechanical Turk; see the Rating Procedure section above).

Considering the difficulty estimates that were obtained from the ratings of correctness, another 100 participants completed the interview task with a final set of 18 easy questions, which was deemed a reasonable sample size (see above). An additional two participants were excluded from this sample because they reported lag or buffering. These participants' responses were rated by a sample of 69 raters. Bonuses were typically given in full, except for participants who did not give meaningful responses and raters who clearly did not follow instructions (e.g., by rating all included catch trials—including fabricated trials that were clearly wrong answers—as correct). In case of dispute, bonuses were paid.

***Analysis of RTs.*** RT in a given trial was computed in milliseconds from the time that the target question appeared on screen to the moment that the "respond" button was clicked. All analyses were implemented using LMEMs in the *lme4* package (Bates et al., 2015) in *RStudio* (Version 1.2.1335, www.rstudio.com). Models were fitted using restricted maximum likelihood estimation and corrected *p* values for predictors were derived via the *lmerTest* package (Kuznetsova et al., 2017) approximating degrees of freedom via Satterthwaite's method. To compare LMEMs, we refitted models using maximum likelihood estimation and compared the AICs.

***Data exclusion of memory-scanning responses.*** To reduce the influence of outliers, we excluded trials from analysis if the RT exceeded two interquartile ranges above the median across all trials (i.e., RT > 19,523 ms; 324 of 3,160 trials were excluded). We decided on median and interquartile ranges for outlier rejection because we assumed that very large values (potentially due to participants leaving the computer) may bias the estimate of mean and standard deviation; we chose two interquartile ranges as a conservative threshold that did not exclude too much data. However, we later confirmed that rejecting values that were 2.5 or 3 standard deviations above the mean yielded qualitatively identical and statistically significant reproductions of all the effects reported in this article.

***Aggregation of ratings and data exclusion of raters.*** Ratings were treated as a dichotomous variable (match vs. mismatch/unclear). Of the 49 raters who rated answers to the first 34 questions, six were excluded because they had rated more than two of the 10 fabricated wrong answers as correct. Every participant in the interview sample was rated by 5.38 raters on average (minimum = 3, maximum = 11). In order to aggregate ratings, we computed Cohen's κ between the raters. Subsequently, the rater who agreed the least with other raters was excluded until there was agreement of κ > 0.6 between all raters. With a total of 30 raters remaining, every participant was rated on average 3.75 times (minimum = 2, maximum = 11). A participant's response was then defined as correct if it was marked as correct by more than 50% of raters. Of the 69 raters who rated the final 18 questions, 27 were excluded because they rated at least one of the fabricated wrong answers as correct (fabricated questions were updated to be more obvious, hence the stricter exclusion criterion). Another 16 raters were excluded until there was agreement of κ > 0.6 between all raters. Each participant was rated by the remaining 26 raters 2.6 times on average (minimum = 2, maximum = 4), and answers were again defined as correct if labeled "correct" by more than 50% of raters. Note that inclusion criteria for raters were strict because we wanted to rely on the best possible raters to determine correctness of responses, that is, raters did not need to be representative of the population.

## Results

In a first analysis, all data from all memory-scanning samples were combined except excluded trials in which the RT exceeded two interquartile ranges above the median across all trials (i.e., RT > 19,523 ms; 324 of 3,160 trials were excluded, see above). We first tested whether questions that were answered correctly (factor: *correct*; $n = 1{,}972$) took less time to answer than trials that were not answered correctly ($n = 864$) by fitting the LMEM RT ~ 1 + correct + $(1 \mid \mathrm{sj_{id}})$ on all data. An effect of *correct* confirmed that RTs were on average 1,457.5 ms faster ($\beta = -1{,}457.5$, $SE = 153.4$, $t = -9.498$, $p < .001$) for correct responses. Despite considering it a nonspecific effect of no interest, we therefore included the difficulty of a question (*diff*) as a predictor in all further models, even if they modeled only correct responses (note that none of our results hinged on the inclusion of the predictor *diff*, but including a relevant predictor of no interest renders model estimates of other predictors more accurate). To test the stepping-stones model, we formulated an LMEM in which RT was modeled as a function of the number of event boundaries crossed between setup and target question (*nEB*) and the distance of the target to the previous event boundary (*distEBpre*). Based on the computational model, these predictors provide independent contributions to the forward scanning process, which is why we did not include interaction terms between the predictors in the LMEM (note that the inclusion of the interaction term does not result in a significant contribution of that predictor). We further included the difficulty of a question as a predictor of no interest.

Moreover, the stepping-stones model suggests that memory scanning proceeds in a forward direction from event boundaries, but memory scanning in a backward direction is a possible alternative (reported, for instance, by Wimmer et al., 2020). As a negative control predictor that we did not expect to contribute to the model prediction, we therefore included the distance of a target to the next event boundary (*distEBpost*), which would contribute to the model prediction only if participants scanned backward from the subsequent event boundary:

$$\mathrm{RT} \sim 1 + nEB + distEBpre + distEBpost + diff + \left(1 \middle| \mathrm{sj_{id}}\right). \quad (6)$$

This model was then fitted on all trials that were correctly answered ($n = 1{,}972$). We found an effect of "distance to the previous event boundary" (β = 47.342, *SE* = 12.655, *t* = 3.741, *p* < .001) and also an effect of the "number of event boundaries" (β = 98.752, *SE* = 40.871, *t* = 2.416, *p* = .016)—taken together, these findings support the stepping-stones model as a plausible process of continuous memory scanning. We further found a contribution of the "difficulty of the question" (β = 37.947, *SE* = 5.587, *t* = 6.792, *p* < .001) on RTs. The "distance of the target to the next event boundary," which was included as a control predictor, did not significantly improve model fit (β = 6.672, *SE* = 7.811, *t* = 0.854, *p* = .393). Excluding the control predictor, the final model consisted of

$$\text{RT} \sim 1 + nEB + distEBpre + diff + \left(1 \mid \text{sj}_{\text{id}}\right), \tag{7}$$

with contributions from the "number of event boundaries crossed" (β = 93.709, *SE* = 40.441, *t* = 2.317, *p* < .021), "distance to the previous event boundary" (β = 48.775, *SE* = 12.543, *t* = 3.889, *p* < .001), and "difficulty of the question" (β = 38.676, *SE* = 5.521, *t* = 7.005, *p* < .001).

We next wanted to test whether some participants were able to skip more event boundaries at a time than others (i.e., if there was variability in how much individual participants skip); specifically, we wanted to model individual data by allowing for a random slope of the predictor "number of event boundaries crossed." This LMEM is described as follows:

$$\text{RT} \sim 1 + nEB + distEBpre + diff + \left(1 + nEB \mid \text{sj}_{\text{id}}\right). \tag{8}$$

This new predictor, however, did not significantly improve the model (*p* = .158), that is, the null hypothesis that the number of event boundaries has a similar influence on scanning time across participants could not be rejected.

Furthermore, to ensure that the model does not simply fit better for data with a high number of boundaries rather than low number of boundaries, we performed a median split by the predictor *nEB*, finding a significant contribution of *distEBpre* in the upper and lower half of the split (*p*s < .05). We also tested a possible interaction between the predictors *nEB* and *distEBpre*, finding no significant interaction, and we tested for heteroscedasticity of the residuals when grouped by *nEB*. Based on Levene's test, however, the null hypothesis of homoscedasticity could not be rejected (*p* > .05).

In the introduction, we discussed studies that reported and estimated the compression of memory replay compared with experience (e.g., Lee & Wilson, 2002; Liu et al., 2019; Michelmann et al., 2019). Estimating the compression level of memory sequences is productive because it provides information about mechanisms of selection and compression of information in memory. In our computational model, we elucidate the mechanisms of memory scanning and are thereby able to account for two relevant factors: "number of event boundaries crossed" and "distance to the previous event boundary."

By relating the coefficient estimates to the original duration of the movie, we can interpret our findings as an estimate of the speed of memory scanning (with the cautionary note that task- and material-specific influences may be relevant). Specifically, if we take the coefficient for "distance to the previous event boundary" as our best estimate of the scanning rate when skipping does not take place, this implies that scanning of 1 s in the final event takes about 48.775 ms in memory. Combining this with the model's estimate of the scanning time per event (93.709 ms), this would mean that participants scan on average 93.709/48.775 = 1.921 s of each event before they skip ahead to the next event boundary (note that this interpretation assumes no temporal cost of skipping and may therefore overestimate the amount of time spent within each event).

In naturalistic experience—and in our data—the absolute duration of experience typically correlates with the number of events that it contains. In our stepping-stones model, the distance of the target of memory scanning to the previous event boundary further accounts for some of the total duration of the scanned memory. A potentially more sparse alternative account for the duration of memory scanning could therefore be given by a predictor that measures the duration of the experience, that is, the length of the scanned segment in memory. As our next step, we therefore wanted to test this hypothesis and compare the stepping-stones model with a simpler model that takes into account only the duration of the segment *dur*, that is, the total time in the movie between the setup and the target question. To this end, we refitted the LMEM (Equation 7) via maximum likelihood estimation and compared it with the LMEM:

$$\text{RT} \sim 1 + dur + diff + \left(1 \mid \text{sj}_{\text{id}}\right). \tag{9}$$

Despite being penalized by AIC for having more parameters, the LMEM that implements the predictions from the stepping-stones model (Equation 7; AIC = 31,795.1) achieved a substantially lower AIC than the segment duration model (Equation 9; AIC = 37,792.2, ΔAIC = 5,997.1, ΔBIC = 5,992.6). Taken together, these data are evidence of a dynamic memory-scanning process, in which participants can skip ahead to the beginning of a new event in their memory.

## Mental Simulation Experiments

The stepping-stones model predicts that the target's distance to the previous event boundary makes a high

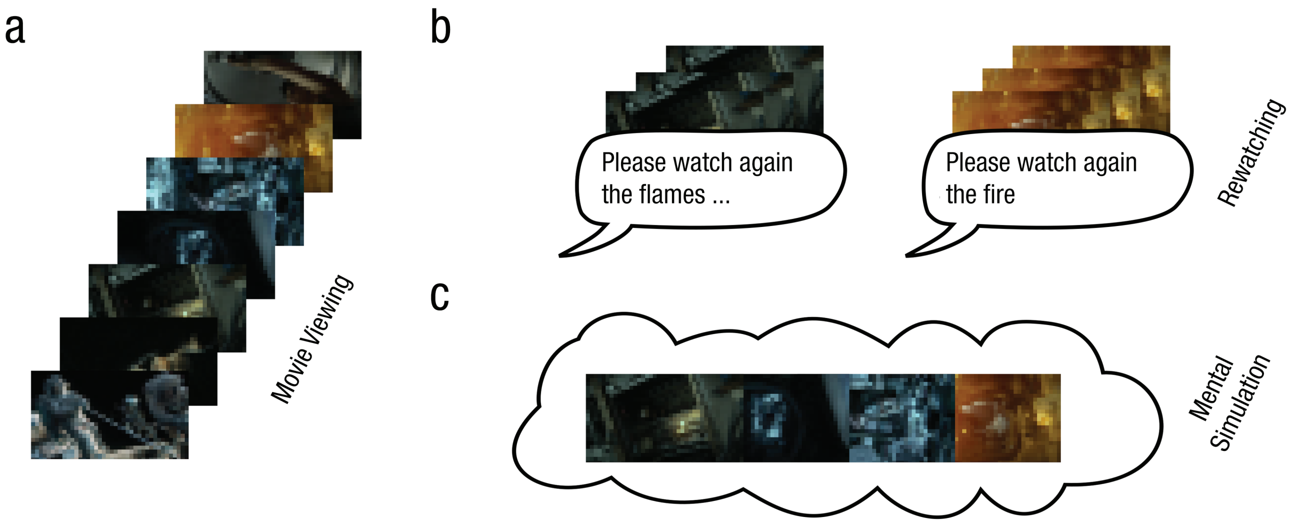
a
b
c
Movie Viewing
Please watch again the flames ...
Please watch again the fire
Rewatching
Mental Simulation

**Fig. 4.** Paradigm prompting mental simulation. Participants watched a short movie (a) and were subsequently asked to mentally simulate segments of the movie (b-c) For each segment the beginning and the end were shown as brief video clips (b). Subsequently, participants were asked to "rewatch" the clip in their mind's eye and indicate the start and stop of their mental simulation by pressing a button (c).

relative contribution to RTs because a low skipping threshold ensures that little time is spent within each event; the final event, however, is searched without skipping. Increasing the skip threshold should therefore make the length of the final event less predictive of RTs because participants spend proportionally more time thinking about the events *before* the final event. In this study (study 3), we aimed to elicit a thorough scanning of continuous memories via "mental simulation" instructions (see Experimental Procedure); that is, we employed a mental simulation task as a (less naturalistic) experimental manipulation of the memory-scanning process, with the goal of increasing the skip threshold and thereby testing predictions from the stepping-stones model.

## Method

### Stimulus material.

*Movies.* The video material consisted of the same two movies of 7-min 30-s duration from the movie *Gravity* (Cuarón, 2013) that were used in the event boundary norming study and the memory-scanning study.

*Segment properties.* For this experiment, we repurposed the 18 easy questions that we had identified for the memory-scanning experiments and asked participants to perform mental simulation of the corresponding segments in the movie, that is, we used the segments that were spanned by these final 18 easy questions. As in the memory-scanning experiment, we characterized segments in terms of their absolute duration in seconds, the number of event boundaries between the beginning of the segment and its end, and the distance of the end to the previous event boundary (if no event boundary was crossed—as was the case for 2 of the 18 segments—the distance was not defined).

***Experimental procedure.*** Experiments designed to elicit a very thorough scanning of continuous mnemonic representation often use a mnemonic task in which participants are asked to mentally simulate memory sequences (e.g., Bonasia et al., 2016; Faber & Gennari, 2015). To elicit more detailed memory scanning, we therefore gave participants mental simulation instructions, informing them of the beginning and the end of the to-be-scanned episode; participants tried to replay this episode in their mind's eye (Fig. 4). We hypothesized that, given these instructions, participants would adopt a higher skip threshold to scan memory more thoroughly; thereby, the length of the final segment would have less of an influence on memory-scanning time: After providing informed consent, participants received information about the movie and the characters to ensure that they would understand the plot. After that, instructions for the experiment were presented, and participants practiced the task using the short video clip featuring Charlie Chaplin.

Participants were instructed that they would first watch a movie. Then they would read the description of two different moments in the movie and be asked to mentally simulate the movie between the first moment and the second moment. Mental simulation was explained as “playing a video in your mind’s eye, like watching a video clip, however you just use your memory and your imagination to do that.” After watching the whole movie, mental simulation trials proceeded as follows. Participants saw the first moment as a short video with a description of the moment in written form under the video. Above the video, a button was available to replay the clip. Below the description, another button (“next”) was available. After participants proceeded, the second moment was shown in the same way: as a short video with a written description underneath. Thereafter, participants performed mental simulation between the first and the second moment in the movie. A timer button was presented on the screen; participants clicked the button when starting mental simulation and clicked it again when they finished their mental simulation (i.e., they had reached the second moment in their mind’s eye). Above the timer button, the descriptions of the two moments were displayed next to each other in columns of a table (column headings: “Moment 1,” “Moment 2”). The label on the timer was “start simulation”; when the button was clicked, the label changed to “stop (complete).” After two practice trials, participants watched the first movie without interruption. Afterward, participants were given a brief reminder of the instructions, and then they mentally simulated the first set of segments. When the task for the first movie was complete, they watched the second movie and then completed the remaining segments performing mental simulation (because the movies formed part of a continuous narrative, the movies were always presented in the same order).

Finally, participants rated their attentiveness, understanding of the movie, and understanding of instructions using a slider (left = 0, right = 100). Participants’ mean attentiveness rating was 95.38 (*SD* = 11.208, range = 79, 100 = fully attentive), mean rating of understanding of the movie was 84.44 (*SD* = 16.511, range = 68, 100 = fully understood everything), and mean rating of understanding of instructions was 95.05 (*SD* = 10.186, range = 68, 100 = fully understood everything). We did not exclude any participants on the basis of these ratings.

***Data collection.*** Data were collected on an online experiment server (www.cognition.run). All participants were recruited via Amazon Mechanical Turk. Cloudresearch (formerly TurkPrime, www.cloudresearch.com) was used to filter out participants with low acceptance rate across studies, to restrict data collection to the United States, and to facilitate payment and assignment of bonuses (Litman et al., 2017). Participants received monetary compensation of $7 for their participation in the experiment and a performance-based bonus of up to $3. Videos were presented at a resolution of 1,280 × 720 pixels. One hundred participants completed the mental simulation task successfully. Data collection was stopped when the predefined sample size was reached. Sample size was determined heuristically by taking into account previously reported sample sizes, by drawing on experience with data quality on Amazon Mechanical Turk (Michelmann et al., 2019, 2021; Rouhani et al., 2020), and by considering previous experience with the memory-scanning experiments. Given these benchmarks from previous studies, the total size of the sample was considered to be appropriate for identifying psychologically meaningful effects.

***Analysis of RTs.*** RTs (here, *mental simulation times*) from human participants were computed in milliseconds as the time between the two button clicks when participants started and stopped their mental simulation. Human participant data were analyzed using LMEMs in the same way that was described for the continuous-memory-scanning experiments (see above). RT of the model simulation in a given trial was defined as the absolute number of steps (comparison operations). To streamline the process, we analyzed RTs from the model simulations in MATLAB (Version 2020b; The MathWorks, Natick, MA) with generalized LMEMs using maximum pseudolikelihood (MPL) estimation (because the model was implemented in MATLAB, using the same analysis software facilitated the repeated fitting of LMEMs under different parameter settings). Models were compared using the AIC. Specifically, we evaluated the competing LMEMs that either explain the RT as a function of the number of event boundaries (*nEB*) crossed, the distance of the target to the previous event boundary (*distEBpre*), and a participant-specific ($\mathrm{sj}_{\mathrm{id}}$) intercept (Equation 5) or (alternatively) as a function of the segment duration (*dur*) and a participant-specific ($\mathrm{sj}_{\mathrm{id}}$) intercept (Equation 4).

***Data exclusion of mental simulation times.*** To reduce the influence of outliers, we excluded trials from analysis if the RT exceeded two interquartile ranges above the median across all trials (i.e., RT > 52,132.85 ms; 149 of 1,800 trials were excluded). We decided on median and interquartile ranges for outlier rejection because we assumed that very large values (potentially because of participants leaving the computer) may bias the estimate of mean and standard deviation; we chose two interquartile ranges as a conservative threshold that did not exclude too much data. However, we later confirmed that rejecting values that are 2.5 or 3 standard deviations above the mean yields qualitatively identical and statistically significant reproductions of all effects reported in this article.

***Simulations with the stepping-stones model.*** We expected that giving participants mental simulation instructions would lead them to increase their threshold for skipping to the next event (resulting in less skipping overall). Before running the mental simulation experiment on human participants, we ran simulations with the stepping-stones model to refine our predictions about how increasing the skip threshold would affect the relationship between RT, the number of events, and the distance of the end of the segment from the previous event boundary. As discussed above, we expected that increasing the skip threshold in the model would make the length of the final event less predictive of RT.

In our simulations, we generated data from 100 simulated participants, each of whom mentally simulated 18 trials. The properties of simulated trials (number of events, duration, distance of target from previous event boundary) were derived from the actual 18 segments. Time samples within each trial were created at a sampling rate of 10 Hz (i.e., 10 steps correspond to 1 s) and rounded to the nearest integer. The bias term of the model was set to $b = 0.45$. Crucially, the skip threshold was varied in steps of five, between $d_{\text{crit}} = 5$ and $d_{\text{crit}} = 80$. A participant-specific random intercept ($1 - 10{,}000 \in \mathrm{N}$) was added to all 18 trials of a given participant and a random noise term ($\pm 100 \in \mathrm{N}$) was added to each trial. We repeated and analyzed this simulation 50 times and report the median of the resulting parameters for robustness.

## Results

***Predictions from the stepping-stones model.*** Figure 5a shows, for simulated data, how the fits of these competing mixed-effects models varied as a function of the skip threshold. The stepping-stones model provided a better fit than the duration-only model for all but the highest skip threshold values (skip threshold ≥ 60). Next, we looked at how the skip threshold affected the significance of prediction and also parameter estimates in the stepping-stones model. As expected, we found that increasing the skip threshold reduced the significance of the predictor *distEBpre* (the distance of the end of the segment to the previous event boundary); *distEBpre* was no longer significant at a skip threshold of 25 (Fig. 5b), at which point the estimate for this predictor had decreased to −0.009 per sampling point of distance to the event boundary (Fig. 5c; note that, by construction, the correct estimate for this predictor should be 1). At very high levels of the skip threshold, this predictor became significant again, but this is an artifact caused by the fact that—in this fixed set of 18 questions—*distEBpre* happens to be negatively correlated with overall duration ($r = -.111$); as the skip threshold increases, *distEBpre* can overfit because of the "inherited" predictiveness from the duration variable. Finally, the predictor that captures the number of crossed event boundaries in the segment (*nEB*) remained highly significant with increasing skip threshold (*p* values even decreased numerically to zero). Because more time was spent within each event, the estimate increased with the threshold for skipping and then saturated when every event was fully scanned.

***Findings from the behavioral data.*** Following the predictions from the simulated data, we expected that—because the skip threshold would be substantially increased—the relative contribution of each event (*nEB*) to the overall RT would be dramatically increased. We further expected that the distance of the end of the segment to the previous event boundary (*distEBpre*) would no longer significantly contribute to explaining the variance in RTs. Finally, we expected that the stepping-stones model would still outperform the duration model unless participants performed a very thorough and exhaustive scan throughout the whole segment.

RTs were analyzed using LMEMs, except for excluded trials where the RT exceeded two interquartile ranges above the median across all trials (i.e., RT > 52,132.85 ms; 149 of 1,800 trials were excluded). When we fitted the LMEM

$$\mathrm{RT} \sim 1 + nEB + distEBpre + \left(1 \mid \mathrm{sj}_{\mathrm{id}}\right) \tag{10}$$

to the remaining 1,651 trials, only the predictor *nEB* explained significant variance in the data (β = 1,395.90, *SE* = 98.46, *t* = 14.177, *p* < .001), and the distance of the second moment to the previous event boundary *distEBpre* was no longer a significant predictor (*p* = .606). The estimated contribution of *nEB* per event of 1,395.90 ms was substantially higher than the contribution of *nEB* to the duration of memory scanning (93.709 ms; compare this value with findings from the continuous-memory-scanning experiments), that is, more time was spent within each event. Note that the high *t* value associated with the significant contribution of the predictor *nEB* speaks against an interpretation in which the mental simulation data are simply noisier than the memory-scanning data. Increased noise would affect the association between all predictors and the outcome variable (i.e., mental simulation time).

Finally, we refitted the LMEM using maximum likelihood estimation to compare it with the simpler model:

$$\mathrm{RT} \sim 1 + dur + \left(1 \mid \mathrm{sj}_{\mathrm{id}}\right). \tag{11}$$

Despite being penalized by AIC for having more parameters, the LMEM that implements predictions from the stepping-stones model resulted in a substantially

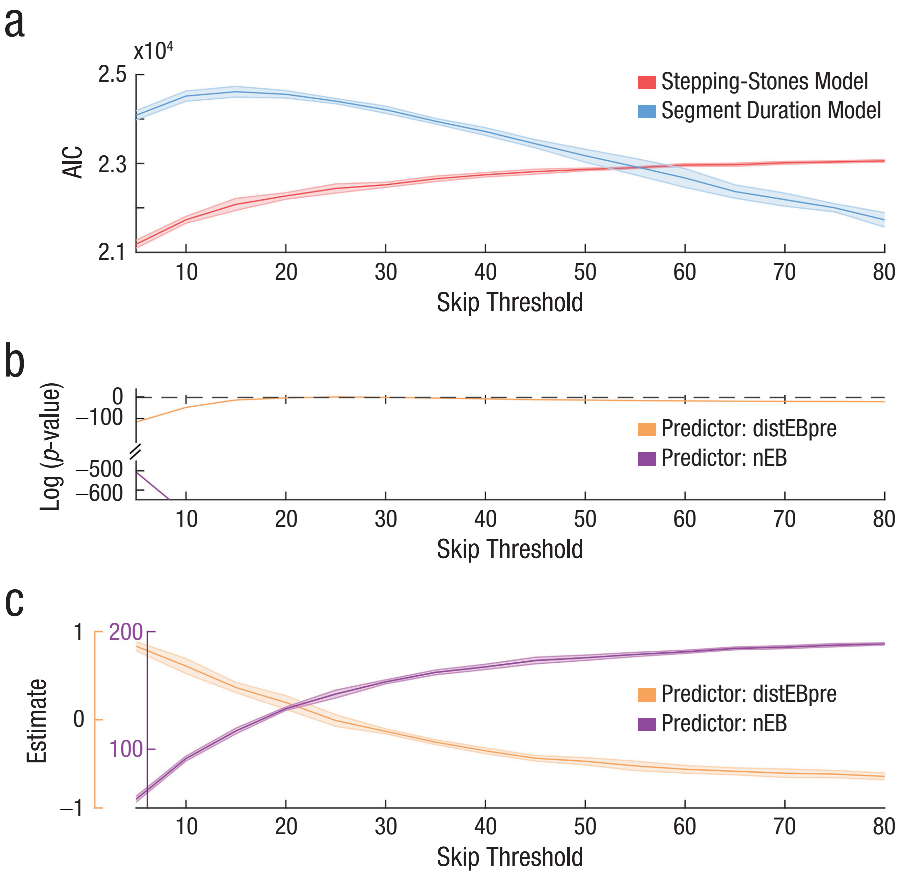

**Fig. 5.** Simulations of behavioral outcome with varying threshold using the stepping-stones model. (a) The stepping-stones model yields a lower (i.e., better) Akaike information criterion (AIC) than the duration model for all but the highest values of the skip threshold. (b) Increasing the skip threshold in the stepping-stones model reduces the significance of the "distance to previous event boundary" (*distEBpre*)" predictor (dashed horizontal line corresponds to $p$ = .05). The "number of event boundaries" (*nEB*) remains a significant predictor in the stepping-stones model for all values of the skip threshold. Note that the significant negative contribution of the *distEBpre* predictor at high thresholds is artifactual (see the text for explanation). (c) With increasing skip threshold, the parameter estimate for *nEB* becomes larger because there are more steps within each event (i.e., a single event accounts for more response time). The parameter estimate for *distEBpre* on the other hand, decreases as the skip threshold increases.

lower AIC (AIC = 30,488.4) compared with the LMEM that takes into account only the segment duration (AIC = 34,776.3, ΔAIC = 4,287.9, ΔBIC = 4,283.2). This suggests that, despite being instructed to scan the whole segment in memory, participants still engaged in some amount of skipping. Interestingly, the estimated contribution of *nEB* per event of 1,395.90 ms was very high in relation to the memory-scanning speed that we estimated in our continuous-memory-scanning study (48.775 ms per second based on the coefficient of the predictor *distEBpre* in that study). Taking the estimate from the memory-scanning experiments at face value, this would mean that participants could scan 28.619 s (1,395.90/48.775) worth of content for each event; however, the average segment duration per event boundary (when boundaries were present) was only 25.513 s. This contradiction suggests that participants may not, in fact, adopt the same scanning rate in memory-scanning and mental simulation tasks—during mental simulation, participants may undertake a more thorough (and slower) sequential scanning while still engaging in skipping.

## Discussion

Human experience is characterized by structure; it can be segmented into meaningful events such as a dinner or a phone call (Zacks et al., 2007). We addressed how event structure is used when we access memories of naturalistic stimuli by prompting participants to scan extended episodes in memory.

To test the hypothesis that event boundaries are access points for memory retrieval, we assessed whether—when scanning continuous memories—participants skip ahead to the beginning of new events to speed up memory scanning: In naturalistic experiments, participants watched movies characterized by event structure. By presenting interview questions that first oriented participants to a specific moment in the movie and then asked about a later moment, we were able to measure the time it takes to get from A to B in memory. We found that the number of events within a segment and the distance of the answer (moment B) to the previous event boundary significantly explained scanning time, outperforming models based on segment duration. This contests the idea of a rigid (exhaustive) memory-scanning process or scanning strategies that are based on semantic relatedness and not event structure. Specifically, a model that takes into account only the duration of a scanned segment is a simpler model of memory scanning and would consequently be preferred if it explained the data similarly well. Likewise, if processes unrelated to event boundaries determined the memory-scanning duration (e.g., if semantic relatedness orchestrated memory search), we would observe no significant contribution of an answer's distance to the previous event boundary when modeling memory-scanning times.

We formulated a computational model that implements memory scanning as a series of comparison operations between a search target and "moments" in memory. When scanning, the model skips ahead if sufficient evidence is accumulated that the target is not in the current event. In the final event, the model automatically lingers because different moments within events are similar (i.e., the target is also similar to moments within its containing event). This model formulation produces our key findings: Memory-scanning time (here, the number of comparison operations) is best explained by the number of event boundaries within a segment and temporal distance of the target to the previous event boundary.

Finally, we used this model to predict the effect of thorough memory scanning. In simulated data, increasing the skip threshold quickly eliminated the significant explanatory contribution of the predictor "distance of the target to the previous event boundary," whereas the relative contribution of the number of event boundaries in predicting RT increased. For all but the highest values of the skip threshold, RTs were better explained by the stepping-stones model than by a model that used only the duration of the segment.

Experimentally, we achieved thorough scanning by asking participants to perform mental simulation, which yielded evidence for these model predictions: Mental simulation times were still better explained by an event-skipping process than by segment durations, but the predictor "distance of the end of the segment to the previous event boundary" no longer contributed significantly. The relative contribution of the number of event boundaries to the RT, however, was dramatically increased, as predicted by our simulations.

These data offer a unique demonstration of how structure in naturalistic experience informs the functionality of our memory system, specifically how we access continuous memories. Prior studies have already established the importance of event boundaries for memory encoding of item and order information (e.g., DuBrow & Davachi, 2013; Pettijohn & Radvansky, 2016; Swallow et al., 2009). Neural evidence from functional MRI and electrocorticography furthermore suggest that the hippocampus encodes information preferably at event boundaries (Baldassano et al., 2017; Ben-Yakov et al., 2013; Michelmann et al., 2021), a finding that is supported by recent modeling work that highlights computational advantages of encoding near event boundaries (Lu et al., 2022). Our results suggest a functional role of event boundaries in the retrieval process. This makes predictions about neurophysiological processes supporting episodic memory under naturalistic conditions (e.g., information flow from hippocampus to cortex should coincide with event-pattern shifts in cortex).

Our findings provide novel behavioral evidence for a flexible mechanism of memory access that is based on the event structure of experience. Even when asking participants to engage in mental simulation, we found that RTs were still in line with a flexible skipping process that leverages event structure. Note that these findings may potentially be limited to adult populations who are comparable with Mechanical Turk workers. Interestingly, however, if we estimate the speed of replay from our memory-scanning experiments and apply this estimate to our mental simulation data, this yields a relatively high estimate of time that participants spend within events. It is therefore possible that the speed of replay can also slow down on a fine-grained level, given retrieval demands. Overall, these findings highlight the importance of event structure for memory search, showing how this structure allows us to efficiently scan through memories of continuously unfolding experiences.

## Transparency

*Action Editor:* Krishnankutty Sathian
*Editor:* Patricia J. Bauer
*Author Contributions*
**Sebastian Michelmann:** Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation;

Methodology; Software; Visualization; Writing – original draft; Writing – review & editing.

**Uri Hasson:** Conceptualization; Funding acquisition; Resources; Supervision.

**Kenneth A. Norman:** Conceptualization; Funding acquisition; Methodology; Resources; Supervision; Writing – original draft; Writing – review & editing.

*Declaration of Conflicting Interests*

The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

*Funding*

This work was supported by Grant R01 MH112357 from the National Institute of Mental Health awarded to U. Hasson and K.A. Norman. S. Michelmann is funded by Deutsche Forschungsgemeinschaft Project 437219953.

*Open Practices*

This article has received the badge for Open Data. More information about the Open Practices badges can be found at http://www.psychologicalscience.org/publications/badges.

## ORCID iD

Sebastian Michelmann https://orcid.org/0000-0002-5717-586X

## Acknowledgments

We thank Qiong Zhang and Elizabeth McDevitt for helpful discussions and comments on earlier versions of this manuscript and Ariel Goldstein and James Antony for helpful comments and discussions.

## References

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721.e5. https://doi.org/10.1016/j.neuron.2017.06.041

Baldassano, C., Hasson, U., & Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, *38*(45), 9689–9699. https://doi.org/10.1523/JNEUROSCI.0251-18.2018

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge University Press.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bellmund, J. L. S., Polti, I., & Doeller, C. F. (2020). Sequence memory in the hippocampal-entorhinal region. *Journal of Cognitive Neuroscience*, *32*(11), 2056–2070. https://doi.org/10.1162/jocn_a_01592

Ben-Yakov, A., Eshel, N., & Dudai, Y. (2013). Hippocampal immediate poststimulus activity in the encoding of consecutive naturalistic episodes. *Journal of Experimental Psychology: General*, *142*(4), 1255–1263. https://doi.org/10.1037/a0033558

Bonasia, K., Blommesteyn, J., & Moscovitch, M. (2016). Memory and navigation: Compression of space varies with route length and turns. *Hippocampus*, *26*(1), 9–12. https://doi.org/10.1002/hipo.22539

Brunec, I. K., Moscovitch, M., & Barense, M. D. (2018). Boundaries shape cognitive representations of spaces and events. *Trends in Cognitive Sciences*, *22*(7), 637–650. https://doi.org/10.1016/j.tics.2018.03.013

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science*, *6*(1), 3–5. https://doi.org/10.1177/1745691610393980

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

Cuarón, A. (Director). (2013). *Gravity* [Film]. Warner Bros.

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

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

Eargle, D., Gureckis, T., Rich, A. S., McDonnell, J., & Martin, J. B. (2020, October 13). *psiTurk: An open platform for science on Amazon Mechanical Turk* (Version 2.3.11). Zenodo. https://doi.org/10.5281/ZENODO.4082367

Ezzyat, Y., & Davachi, L. (2014). Similarity breeds proximity: Pattern similarity within and across contexts is related to later mnemonic judgments of temporal proximity. *Neuron*, *81*(5), 1179–1189. https://doi.org/10.1016/j.neuron.2014.01.042

Faber, M., & Gennari, S. P. (2015). In search of lost time: Reconstructing the unfolding of events from memory. *Cognition*, *143*, 193–202. https://doi.org/10.1016/j.cognition.2015.06.014

Geerligs, L., Gözükara, D., Oetringer, D., Campbell, K. L., van Gerven, M., & Güçlü, U. (2022). A partially nested cortical hierarchy of neural states underlies event segmentation in the human brain. *eLife*, *11*, Article e77430. https://doi.org/10.7554/eLife.77430

Geerligs, L., van Gerven, M., & Güçlü, U. (2021). Detecting neural state transitions underlying event segmentation. *Neuroimage*, *236*, Article 118085. https://doi.org/10.1016/j.neuroimage.2021.118085

Gureckis, T. M., Martin, J., McDonnell, J., Rich, A. S., Markant, D., Coenen, A., Halpern, D., Hamrick, J. B., & Chan, P. (2016). psiTurk: An open-source framework for conducting replicable behavioral experiments online. *Behavior Research Methods*, *48*(3), 829–842. https://doi.org/10.3758/s13428-015-0642-8

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

*Learning, Memory, and Cognition*, *44*(7), 1075–1090. https://doi.org/10.1037/xlm0000503

Jeunehomme, O., & D'Argembeau, A. (2020). Event segmentation and the temporal compression of experience in episodic memory. *Psychological Research*, *84*, 481–490. https://doi.org/10.1007/s00426-018-1047-y

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, Article 4. https://doi.org/10.3389/neuro.06.004.2008

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13). https://doi.org/10.18637/jss.v082.i13

Lee, A. K., & Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron*, *36*(6), 1183–1194. https://doi.org/10.1016/S0896-6273(02)01096-6

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Liu, Y., Dolan, R. J., Kurth-Nelson, Z., & Behrens, T. E. J. (2019). Human replay spontaneously reorganizes experience. *Cell*, *178*(3), 640–652.e14. https://doi.org/10.1016/j.cell.2019.06.012

Lositsky, O., Chen, J., Toker, D., Honey, C. J., Shvartsman, M., Poppenk, J. L., Hasson, U., & Norman, K. A. (2016). Neural pattern change during encoding of a narrative predicts retrospective duration estimates. *eLife*, *5*, Article 16070. https://doi.org/10.7554/eLife.16070

Lu, Q., Hasson, U., & Norman, K. A. (2022). A neural network model of when to retrieve and encode episodic memories. *eLife*, *11*, Article e74445. https://doi.org/10.7554/eLife.74445

Michelmann, S., Price, A. R., Aubrey, B., Strauss, C. K., Doyle, W. K., Friedman, D., Dugan, P. C., Devinsky, O., Devore, S., Flinker, A., Hasson, U., & Norman, K. A. (2021). Moment-by-moment tracking of naturalistic learning and its underlying hippocampo-cortical interactions. *Nature Communications*, *12*(1), Article 5394. https://doi.org/10.1038/s41467-021-25376-y

Michelmann, S., Staresina, B. P., Bowman, H., & Hanslmayr, S. (2019). Speed of time-compressed forward replay flexibly changes in human episodic memory. *Nature Human Behaviour*, *3*(2), 143–154. https://doi.org/10.1038/s41562-018-0491-4

Nosofsky, R. M., & Palmeri, T. J. (2015). An exemplar-based random-walk model of categorization and recognition. In J. R. Busemeyer, Z. Wang, J. T. Townsend, & A. Eidels (Eds.), *The Oxford handbook of computational and mathematical psychology* (pp. 142–164). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199957996.013.7

Pettijohn, K. A., & Radvansky, G. A. (2016). Walking through doorways causes forgetting: Environmental effects. *Journal of Cognitive Psychology*, *28*(3), 329–340. https://doi.org/10.1080/20445911.2015.1123712

Radvansky, G. A., & Copeland, D. E. (2006). Walking through doorways causes forgetting: Situation models and experienced space. *Memory & Cognition*, *34*(5), 1150–1156. https://doi.org/10.3758/BF03193261

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: Theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922. https://doi.org/10.1162/neco.2008.12-06-420

Roseboom, W., Fountas, Z., Nikiforou, K., Bhowmik, D., Shanahan, M., & Seth, A. K. (2019). Activity in perceptual classification networks as a basis for human subjective time perception. *Nature Communications*, *10*, 267. https://doi.org/10.1038/s41467-018-08194-7

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

Sargent, J. Q., Zacks, J. M., Hambrick, D. Z., Zacks, R. T., Kurby, C. A., Bailey, H. R., Eisenberg, M. L., & Beck, T. M. (2013). Event segmentation ability uniquely predicts event memory. *Cognition*, *129*(2), 241–255. https://doi.org/10.1016/j.cognition.2013.07.002

Sonkusare, S., Breakspear, M., & Guo, C. (2019). Naturalistic stimuli in neuroscience: Critically acclaimed. *Trends in Cognitive Sciences*, *23*(8), 699–714. https://doi.org/10.1016/j.tics.2019.05.004

Sun, C., Yang, W., Martin, J., & Tonegawa, S. (2020). Hippocampal neurons represent events as transferable units of experience. *Nature Neuroscience*, *23*(5), 651–663. https://doi.org/10.1038/s41593-020-0614-x

Swallow, K. M., Zacks, J. M., & Abrams, R. A. (2009). Event boundaries in perception affect memory encoding and updating. *Journal of Experimental Psychology: General*, *138*(2), 236–257. https://doi.org/10.1037/a0015631

Wimmer, G. E., Liu, Y., Vehar, N., Behrens, T. E. J., & Dolan, R. J. (2020). Episodic memory retrieval success is associated with rapid replay of episode content. *Nature Neuroscience*, *23*(8), 1025–1033. https://doi.org/10.1038/s41593-020-0649-z

Zacks, J. M., Kumar, S., Abrams, R. A., & Mehta, R. (2009). Using movement and intentions to understand human activity. *Cognition*, *112*(2), 201–216. https://doi.org/10.1016/j.cognition.2009.03.007

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: A mind-brain perspective. *Psychological Bulletin*, *133*(2), 273–293. https://doi.org/10.1037/0033-2909.133.2.273

Zuo, S., Wang, L., Shin, J. H., Cai, Y., Zhang, B., Lee, S. W., Appiah, K., Zhou, Y.-d., & Kwok, S. C. (2020). Behavioral evidence for memory replay of video episodes in the macaque. *eLife*, *9*, Article e54519. https://doi.org/10.7554/eLife.54519