

¹ The psychological arrow of time drives temporal asymmetries in
² inferring unobserved past and future events

³ Xinming Xu¹, Ziyan Zhu², Xueyao Zheng³, and Jeremy R. Manning^{1,*}

⁴ ¹Dartmouth College, Hanover, NH, USA

⁵ ²Peking University, Beijing, China

⁶ ³Beijing Normal University, Beijing, China

⁷ *Address correspondence to jeremy.r.manning@dartmouth.edu

⁸ September 26, 2023

⁹ **Abstract**

¹⁰ How much can we infer about the past and future, given our knowledge of the present? Unlike temporally
¹¹ symmetric inferences about simple sequences, inferences about our own lives are asymmetric: we are better
¹² able to infer the past than the future, since we remember our past but not our future (i.e., the psychological
¹³ arrow of time). What happens when both the past and future are unobserved, as when we make inferences
¹⁴ about *other* people's lives? We had participants in two experiments view segments of two character-driven
¹⁵ television dramas. They wrote out what would happen just before or after each just-watched segment.
¹⁶ Participants were better at inferring past (versus future) events. This asymmetry was driven by participants'
¹⁷ reliance on characters' conversational references in the narrative, which tended to favor the past. We also
¹⁸ carried out a meta analysis to estimate the prevalence of these asymmetries in hundreds of millions of
¹⁹ dialogues from television shows, popular movies, novels, and written and spoken natural conversations. We
²⁰ found that, on average, references to the past are roughly 1.5–2 times more prevalent in human conversations
²¹ than references to the future. Our work reveals a temporal asymmetry in how observations of other people's
²² behaviors can inform us about the past and future.

²³ **Keywords:** arrow of time, prediction, retrodiction, narrative, conversation

²⁴ Introduction

²⁵ What we experience in the current moment tells us about *now*—but what does it tell us about the
²⁶ past or future? And does the current moment tell us, as human observers, *more* about the past or
²⁷ about the future? One way of examining these questions is to consider highly simplified scenarios
²⁸ that are artificially constructed in the laboratory (e.g., Maheu et al., 2022). At one extreme, for
²⁹ deterministic sequences with *known* rules, knowing the current state provides the observer with
³⁰ sufficient information to exactly reconstruct the entire past and future history of the stimulus. At
³¹ another extreme, for purely random sequences, observing the current state provides no information
³² about the past *or* future.

³³ Sequences generated by stochastic processes fall somewhere between these two extremes. For
³⁴ Markov processes, where each state is solely dependent on the immediately preceding state,
³⁵ Shannon entropy may be used to quantify the uncertainty of the past and future states, given the
³⁶ present state. Cover (1994) showed that, for any stationary process (i.e., processes in equilibrium),
³⁷ Markov or otherwise, the present state provides equal information (i.e., mutual information) about
³⁸ past and future states (also see Bialek et al., 2001; Ellison et al., 2009). Further, there is some
³⁹ evidence that humans are similarly adept at inferring the most likely previous and next items in
⁴⁰ sequences governed by stochastic Markov processes (Jones and Pashler, 2007).

⁴¹ Deterministic, random, and probabilistic sequences (in equilibrium) are all symmetric: the
⁴² present state of these sequences is equally informative about past versus future states. In contrast,
⁴³ our subjective experience in everyday life is that we know more about our own past than our
⁴⁴ future (e.g., Horwich, 1987). We have memories of our past that we carry with us into the
⁴⁵ present moment, but we do not have memories of our yet-to-be-experienced future. This temporal
⁴⁶ asymmetry imposes an “arrow of time” on our subjective experience, known as the *psychological*
⁴⁷ *arrow of time* (e.g., Hawking, 1985).

⁴⁸ Although the psychological arrow of time implies that we should be better able to infer our
⁴⁹ past than our future, how generally does this temporal asymmetry hold? And does the asymmetry
⁵⁰ hold only for our own experiences (due to our memories), or is the asymmetry a general property

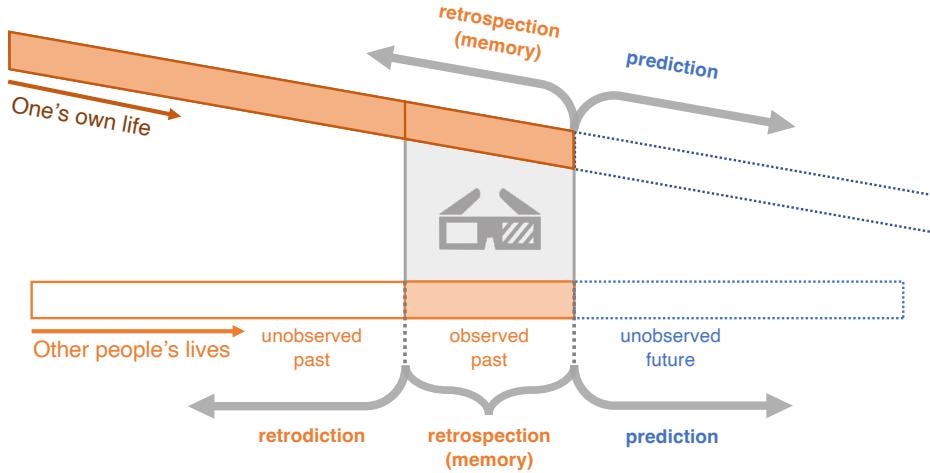


Figure 1: Retrodiction, retrospection, and prediction. In one’s own life, one may draw on memory to retrospect (i.e., review or re-evaluate) the past or predict the future. This process is time-asymmetric, since our own past is (typically) observed whereas our future is not. When we make inferences about *other* people’s lives, however, we often have uncertainty about both their past and future, since we may have observed neither. We may *retrodict* the unobserved past and predict the unobserved future of other people’s lives.

of any real-life event sequence? In real-world situations (and narratives) where we are *equally* ignorant of the past and future, as for *other* people’s lives where we lack memories of the relevant past, are our inferences about the past and future symmetric or asymmetric? For example, imagine that you are meeting a stranger for the first time. At the moment of your meeting, you lack both memories of their past and knowledge about what they might do in the future. After your first encounter with the stranger, would you be able to more accurately or easily form inferences about what had happened in their past (*retrodiction*) or what will happen in their future (*prediction*; Fig. 1)? Or suppose you started watching a movie partway through. Again, you would enter the moment of watching without memories of prior parts of the movie. Given your observations in the present, would your guesses about what had happened before you started watching be more (or less) accurate than your guesses about what will happen next? In general, when the past and future are *both* unobserved, are we better at inferring the past or the future in real-world settings? Narrative stimuli, such as stories and movies, can provide a useful testbed for exploring several of

64 these questions.

65 Although narratives are unlikely to be confused with one's own experiences, narratives mirror
66 some of the structure of real-world experiences. Character behaviors and interactions are often
67 designed in a way that helps the audience connect with or relate to the characters. Events in
68 narratives also unfold in ways that are intended to build rapport or engagement with the audience.
69 This might be accomplished by having events follow a believable structure that is reminiscent of
70 real-world experiences, or by designing the audience's experiences in ways that communicate clear
71 "rules" or "features" that help to immerse the audience in the narrative's universe. The characters
72 in a realistic narrative can also be written to behave in ways reminiscent of real-world people.
73 These same aspects of narratives that authors use to drive engagement with events and characters
74 can lead narratives to replicate some core aspects of real-world experiences that are typically lost or
75 overlooked in traditional sequence learning paradigms. Narratives can drive the audience to build
76 situation models (Radvansky and Copeland, 2006; Zwaan and Radvansky, 1998) of the narrative's
77 universe, or to form a theory of mind of and make predictions about the characters (Tamir and
78 Thornton, 2018; Koster-Hale and Saxe, 2013). Events in narratives may unfold in a consistent or
79 logical way, but they also exhibit complex and meaningful interactions across events reminiscent of
80 real-world experiences (but not necessarily the simple sequences traditionally used in the statistical
81 learning literature).

82 One key difference between simple artificial sequences and more naturalistic (real or narrative)
83 sequences is that naturalistic sequences often incorporate other people. Despite the past and fu-
84 ture being equally unknown to *the observer* prior to the current moment, other people, and realistic
85 characters in narratives, have their own psychological arrows of time. Specifically, they have mem-
86 ories of their own pasts. Other people's asymmetric knowledge about their *own* pasts and futures
87 might affect their behaviors (e.g., conversations). In turn, this might provide time-asymmetric
88 clues that favor the past (e.g., other people might talk more about their own pasts than their
89 futures; Demiray et al., 2018). If observers leverage these clues from other people's asymmetric
90 knowledge, then observers should also be better at inferring the past (versus the future) of other
91 people's lives. Alternatively, inferences about other people's lives may be more like inferences

92 about artificial statistical sequences (e.g., perhaps solely relying on statistical regularities like event
93 schemas, scripts, or situation models; Radvansky and Copeland, 2006; Zwaan and Radvansky,
94 1998; Bower et al., 1979; Ranganath and Ritchey, 2012; Baldassano et al., 2018). If so, then the
95 accuracy of inferences about the past and the future of others' lives should be approximately equal.
96 We note that the aforementioned authors make no specific claims about temporal symmetries or
97 asymmetries. Rather, we claim that statistical regularities might *imply* symmetry (e.g., if you are
98 on step n of an unfolding schema, this suggests you have just completed step $n - 1$ and that you
99 are likely to next encounter step $n + 1$).

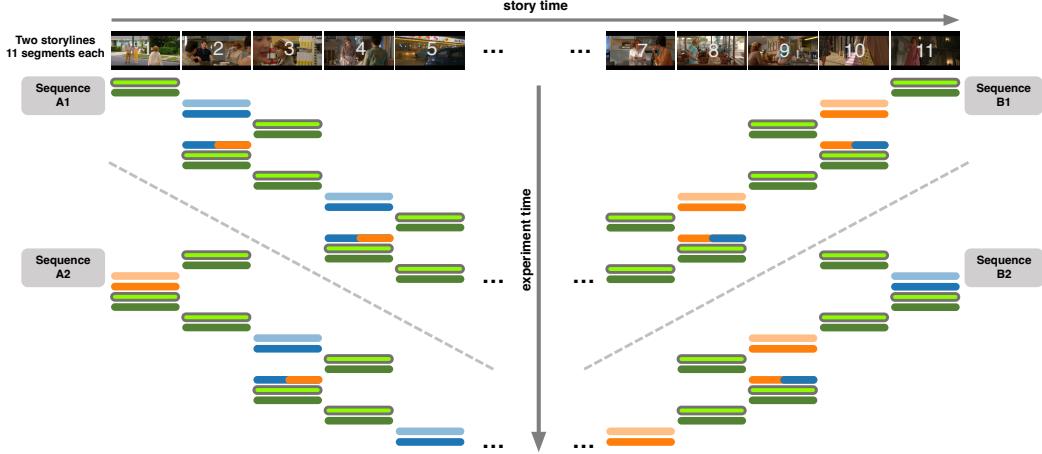
100 We designed a naturalistic paradigm for exposing participants to scenarios where the past
101 and future were equally unobserved. We asked our participants to watch a series of movie
102 segments drawn from a character-driven dramatic television show. Across the conditions and
103 trials in the experiment, participants made free-form text responses to either retrodict what had
104 happened in the previous segment, predict what would happen in the next segment, or recall
105 what happened in the just-watched segment. We used manual annotations and sentence-level
106 natural language processing models to characterize participants' responses. To foreshadow our
107 results, we found that participants were overall better at retrodicting the past than predicting the
108 future. This appeared to be driven by two main factors. First, characters more often referred to
109 past events than future (e.g., planned) events, and this influenced participants' responses. Second,
110 associations and dependencies between temporally adjacent events enabled participants to form
111 estimates about nearby events (e.g., to a just-watched scene or a past or future event referenced
112 in an observed conversation). We also ran a pre-registered replication study to confirm that these
113 findings generalized to another television show and group of participants. Finally, we ran a meta
114 analysis using natural language processing to estimate the prevalence of references to past and
115 future events in hundreds of millions of dialogues drawn from television shows, popular movies,
116 novels, and written and spoken natural conversations. Taken together, our work reveals a temporal
117 asymmetry in how observations of other humans' behaviors inform us about the past versus the
118 future.

¹¹⁹ **Results**

¹²⁰ Participants in our main experiment ($n = 36$) watched segments from two storylines, drawn
¹²¹ from the CBS television show *Why Women Kill*. Each storyline comprised 11 segments (mean
¹²² duration: 2.05 min; range: 0.97–3.87 min, Table S1). We asked participants to use free-form
¹²³ (typed) text responses to retrodict what had happened prior to a just-watched segment, predict
¹²⁴ what would happen next, or recall what they had just watched (Fig. 2, *Task design*). We referred
¹²⁵ to the to-be-retrodicted, to-be-predicted, or to-be-recalled segment as the *target segment* for each
¹²⁶ response. We systematically varied whether participants watched the segments in forward or
¹²⁷ reverse chronological order, and how many segments they had seen prior to making a response
¹²⁸ (see *Methods*).

¹²⁹ We asked participants in our main experiment to generate four types of responses after watching
¹³⁰ each video segment: uncued responses, character-cued responses, updated responses, and recalls
¹³¹ (Fig. 2, *Data overview*). To generate *uncued* responses, we asked participants to either retrodict
¹³² (uncued retrodiction; *u-R*) what happened shortly before or predict (uncued prediction; *u-P*) what
¹³³ happened shortly after the just-watched segment. To generate *character-cued* responses, we asked
¹³⁴ participants to retrodict (character-cued retrodiction; *c-R*) or predict (character-cued prediction;
¹³⁵ *c-P*) what came before or after the just-watched segment, but we provided additional information
¹³⁶ to the participant about which character(s) would be present in the target (to-be-retrodicted or to-
¹³⁷ be-predicted) segment. We hypothesized that character-cued responses should be more accurate
¹³⁸ than uncued responses, to the extent that participants incorporate the character information we
¹³⁹ provided to them into their retrodictions and predictions. To generate updated responses, we
¹⁴⁰ asked participants to watch an additional segment that came just prior to or just after the target
¹⁴¹ segment, and then to update their retrodiction (*c-RP*) or prediction (*c-PR*) about the target segment.
¹⁴² Results on updated responses are not reported in this paper. Finally, we also asked participants to
¹⁴³ *recall* what happened in the just-watched segment. We labeled these responses according to which
¹⁴⁴ other segments participants had watched prior to the just-watched target. Retrodiction-matched
¹⁴⁵ recall (*re(R)*) responses were made during the retrodiction sequences (B1 and B2; Fig. 2), whereas

Task design



Conditions

- Watch
- u-R: uncued retrodiction
- u-P: uncued prediction
- c-R: character-cued retrodiction
- c-P: character-cued prediction
- c-RP: updated retrodiction (after watching one segment earlier)
- c-PR: updated prediction (after watching one segment later)
- Recall
- re(R): retrodiction-matched recall
- re(P): prediction-matched recall
- ...

Data overview

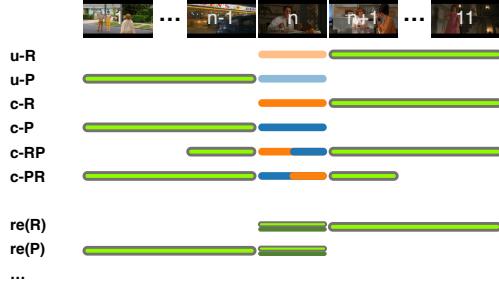


Figure 2: Task overview. Participants in our main experiment watched segments of two storylines from the television series *Why Women Kill*. They made free-form text responses to either retrodict what had happened in the previous segment, predict what would happen in the next segment, or recall what happened in the just-watched segment. Across four counterbalanced sequences, we systematically varied whether participants watched the segments in forward or reverse chronological order, whether (or not) responses were cued using the main characters in the target segment, and which other segments participants had watched prior to making a response. For each segment, we collected several retrodiction, prediction, and/or recall responses across different experimental conditions. Experiment time is denoted along the vertical axis, storyline segments are indicated along the horizontal axis, and the colors denote experimental tasks (conditions). For an analogous depiction of our replication experiment's design, see Fig. S4.

146 prediction-matched recall ($re(P)$) responses were made during the prediction sequences (A1 and A2;
147 Fig. 2). Whereas retrodiction and prediction responses reflect what participants *estimate* they would
148 remember after watching the (inferred) target segment, recall responses provide a benchmark for
149 comparison by measuring what they *actually* remember about the target segment. Our replication
150 experiment (Fig. S4) used a similar design, but did not have participants generate recall, $re(R)$, or
151 $re(P)$ responses.

152 For each retrodiction and prediction, participants were asked to generate at least one, and not
153 more than three, responses that constituted “the sorts of things [the participant would] expect
154 to have remembered if [they] had watched the [target] segment.” They were asked to generate
155 multiple responses only if those additional responses were (in their judgement) of equal likelihood
156 to occur. On average, participants in our main experiment generated 1.08 responses per prompt;
157 therefore we chose to consider only participants’ first (“most probable” or “most important”)
158 responses to each prompt. We also discarded a small number ($n = 20$) of character-cued responses
159 that did not contain references to all cued characters, along with one additional response due to
160 the participant’s misunderstanding of the task instructions during that trial. We carried out our
161 analyses on the remaining 2084 retrodiction, prediction, and recall responses. (Our replication
162 experiment analyses were carried out on XXX responses.)

163 We used two general approaches to assess the quality of participants’ responses (see *Meth-*
164 *ods*, Figs. 3A). One approach entailed manually annotating events in the video and counting the
165 number of matched events in participants’ responses. We identified a total of 117 unique events
166 reflected across the 22 video segments in our main experiment (range: 3–9 per segment; see *Meth-*
167 *ods*, Table S1). We assigned one “point” to each of these video events. We also identified 23
168 additional events in participants’ responses that were either summaries of several events or that
169 were partial matches to the manually identified video events. We assigned 0.5 point to each of
170 these additional events. This point system enabled us to compute the numbers and proportions
171 (*hit rates*) of correctly retrodicted, predicted, and recalled events contained in each response. Our
172 second approach entailed using a natural language processing model (Cer et al., 2018) to embed
173 annotations and responses in a 512-dimensional feature space. This approach was designed to

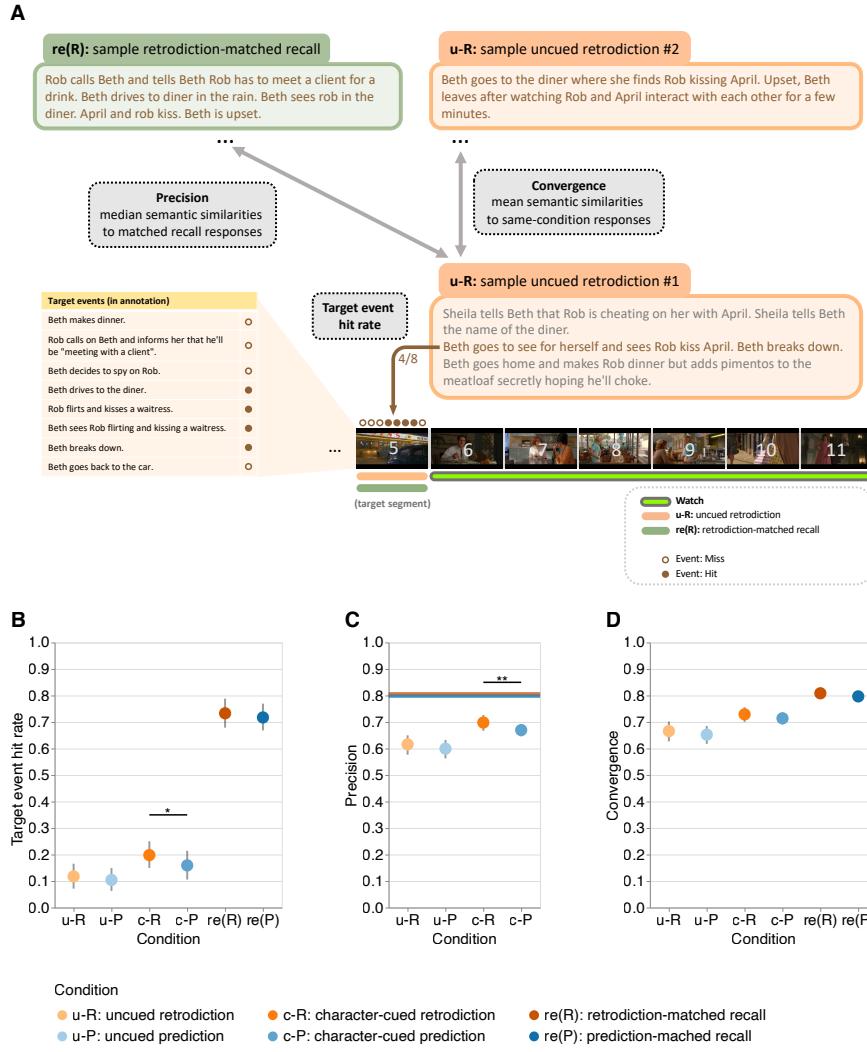


Figure 3: Retrodiction, prediction, and recall performance by experimental condition in our main experiment. **A. Methods schematic.** For each retrodiction, prediction, and recall response, we calculated the hit rate for events in the target segment, the response precision (see *Methods*), and the response convergence across participants (see *Methods*). **B. Target event hit rate.** Mean proportions of target events that were contained in participants' responses, for each response type, averaged across target segments. **C. Response precision.** Mean precisions of participants' responses, for each response type, averaged across target segments. The horizontal lines denote the mean pairwise semantic similarities (see *Methods*) across recall responses (re(R): orange; re(P): blue). **D. Response convergence.** Mean (across-participant) convergence of participants' responses, for each response type, averaged across target segments. All panels: error bars denote bootstrapped 95% confidence intervals. Asterisks indicate significance in the (generalized) linear mixed models: * denotes $p < 0.05$ and ** denotes $p < 0.01$. See Figure S5 for analogous results from our replication experiment.

capture conceptual overlap between responses that were not necessarily tied to specific events. To quantify this conceptual overlap, we computed the similarities between the embeddings of different sets of responses. Following Heusser et al. (2021), we defined the *precision* of each participants' retrodictions or predictions about a target segment as the median cosine similarities between the embeddings of (a) the participant's retrodiction or prediction response for the target segment and (b) each *other* participant's recalls of the same segment. In other words, precision is designed to measure the extent to which retrodictions and predictions captured the conceptual content that (other) participants remembered. We also developed a related measure, which we call *convergence*, to characterize response similarities across participants. In particular, we defined convergence as the mean cosine similarity between the embeddings of a participant's responses to a target segment and all other participants' responses (of the same type) to the same segment. We analyzed the data using generalized linear mixed models, with participant and stimulus (e.g., target segment) identities as crossed random effects (see *Methods*).

First we sought to validate a main effect of response type (i.e., uncued responses, character-cued responses, and recalls), irrespective of the temporal direction (retrodition versus prediction). Across these three types of responses, participants have access to increasing amounts of information about the target segment. Therefore, across these response types, we hypothesized that participants' responses should become both more accurate and more convergent across individuals. Consistent with this hypothesis, participants' character-cued retrodictions and predictions were associated with higher target event hit rates than uncued retrodictions and predictions in our main experiment (odds ratio (OR): 2.65, $Z = 4.24$, $p < 0.001$, 95% confidence interval (CI): 1.69 to 4.16; Fig. 3B). These character-cued responses were also more precise ($b = 0.13$, $t(18.1) = 9.43$, $p < 0.001$, CI: 0.10 to 0.16; Fig. 3C) and convergent across individuals ($b = 0.11$, $t(18.6) = 6.21$, $p < 0.001$, CI: 0.07 to 0.15; Fig. 3D). Relative to character-cued responses, participants' recalls showed higher target event hit rates (OR = 21.83, $Z = 10.61$, $p < 0.001$, CI: 12.35 to 38.59) and were more convergent across individuals ($b = 0.20$, $t(19.4) = 9.10$, $p < 0.001$, CI: 0.16 to 0.25). These results are consistent with the common-sense notion that access to more information about a target segment yields better performance (i.e., higher hit rates, precision, and convergence across individuals).

202 uals). These findings also held for our replication experiment (Fig. S5; hit rates of character-cued
203 vs. uncued responses: OR: XXX, Z = XXX, p = XXX, 95% confidence interval (CI): XXX to XXX;
204 precisions of character-cued vs. uncued responses: b = XXX, t (XXX) = XXX, p = XXX, CI: XXX
205 to XXX; convergence of character-cued vs. uncued responses: b = XXX, t (XXX) = XXX, p = XXX,
206 CI: XXX to XXX).

207 Next we carried out a series of analyses specifically aimed at characterizing temporal direc-
208 tion effects— i.e, the relative quality of retrodictions versus predictions across different types of
209 responses. We hoped that these analyses might provide insights into our central question about
210 whether inferences about the past and future are equally accurate. Across both uncued and
211 character-cued responses in our main experiment (Fig. 2), retrodictions had numerically higher
212 hit rates than predictions (Fig. 3B). However, these differences were only statistically reliable for
213 character-cued responses (uncued responses: OR = 1.17, Z = 0.35, p = 0.73, CI: 0.47 to 2.92;
214 character-cued responses: OR = 1.93, Z = 2.15, p = 0.03, CI: 1.06 to 3.52). We observed a similar
215 pattern of results for the precisions of participants' responses (Fig. 3C). Specifically, their responses
216 tended to be numerically more precise for retrodictions versus predictions, but the differences were
217 only statistically reliable for character-cued responses (uncued responses: b = 0.03, t (20.9) = 1.09,
218 p = 0.29, CI: -0.03 to 0.10; character-cued responses: b = 0.06, t (20.8) = 3.01, p = 0.007, CI: 0.02
219 to 0.11). We also consistently observed numerically higher convergence across participants for
220 retrodictions versus predictions (Fig. 3D), but neither of these differences were statistically reliable
221 (uncued responses: b = 0.03, t (17.9) = 0.75, p = 0.46, CI: -0.05 to 0.11; character-cued responses:
222 b = 0.04, t (17.4) = 1.46, p = 0.16, CI: -0.02 to 0.09). In our replication experiment (Fig. S5), partici-
223 pants were numerically better at making *predictions* than retrodictions, but none of these differences
224 were statistically reliable (hit rate for uncued responses: OR = XXX, Z = XXX, p = XXX, CI: XXX
225 to XXX; hit rate for character-cued responses: OR = XXX, Z = XXX, p = XXX, CI: XXX to XXX; precision
226 for uncued response: b = XXX, t (XXX) = XXX, p = XXX, CI: XXX to XXX; precision
227 for character-cued responses: b = XXX, t (XXX) = XXX, p = XXX, CI: XXX to XXX; convergence
228 for uncued responses: b = XXX, t (XXX) = XXX, p = XXX, CI: XXX to XXX; convergence for
229 character-cued responses: b = XXX, t (XXX) = XXX, p = XXX, CI: XXX to XXX). Taken together,

230 our results across our main and replication experiment suggest that whether participants are better
231 at retrodicting versus predicting the immediate past or future may be somewhat stimulus specific.
232 We also verified that this was not solely a consequence of how participants' memory performance
233 might have been affected by watching different segments (or making different responses to other
234 segments) across conditions by comparing recall responses in the retrodiction-matched recall ($re(R)$)
235 and prediction-matched recall ($re(P)$) conditions. Recall performance in our main experiment was
236 similar in both conditions (target event hit rate: OR = 1.12, Z = 1.07, p = 0.29, CI: 0.91 to 1.39;
237 convergence: b = 0.03, $t(19.3)$ = 1.89, p = 0.07, CI: 0.00 to 0.07). (We did not collect recall responses
238 in our replication experiment.)

239 The above analyses were focused solely on the target segment (i.e., retrodiction of segment n
240 after watching segments $(n + 1)\dots11$, or prediction of segment n after watching segments $1\dots(n - 1)$).
241 We wondered whether participants' responses might also contain longer-range information about
242 preceding or proceeding events. In order to carry out this analysis properly, we reasoned that
243 participants might reference past or future events that were *implied* to have occurred offscreen,
244 but not explicitly shown onscreen. For example, a character in location A during one scene might
245 appear in location B during the immediately following scene. Although it wasn't shown onscreen,
246 we can infer that the character traveled between locations A and B sometime between the time
247 intervals separating the scenes (Bordwell, 2008). In all, we manually identified a set of 74 *implicit*
248 offscreen events in our main experiment's stimuli that were implied to have occurred given what
249 was (explicitly) depicted onscreen (Fig. 4A), plus one additional partial event and one additional
250 summary event. We applied the same procedure to our replication experiment's stimuli and
251 identified XXX implicit offscreen events. We defined the just-watched segment as having a *lag* of 0.
252 We assigned the target segment of a participant's retrodiction or prediction (i.e., the immediately
253 preceding or proceeding segment) a lag of -1 or +1, respectively. The segment following the next
254 was assigned a lag of 2, and so on. We tagged offscreen events using half steps. For example, an
255 offscreen event that occurred after the prior segment but before the just-watched segment would
256 be assigned a lag of -0.5.

257 Because there is no "ground truth" number of offscreen events, we could not compute the hit

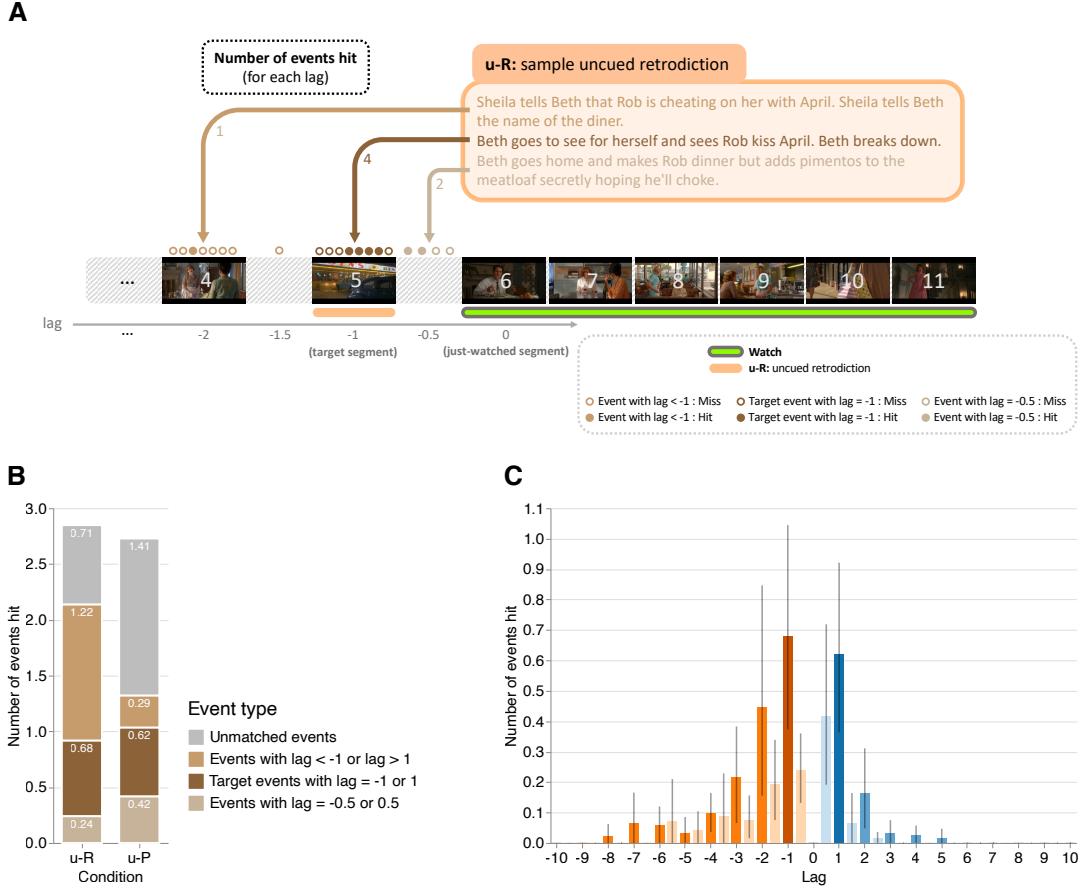


Figure 4: Retrodictions and predictions of temporally near and distant events. A. Illustration of annotation approach. For each uncued retrodiction and prediction response in our main experiment, we calculated the number of (retrodicted or predicted) events as a function of temporal distance from the target segment, or *lag*. Onscreen (explicit) events are tagged using integer-valued lags, whereas offscreen (implicit) events are tagged using half-step lags ($\pm 0.5, \pm 1.5$, etc.). **B. Number of events hit in participants' uncued retrodictions and predictions for each event type.** Here we separated events we identified in participants' responses according to whether they occurred in the target segment (lags of ± 1), during the interval between the target segment and the just-watched segment (lags of ± 0.5), at longer temporal distances ($|lag| > 1$), or were incorrect (unmatched with any past or future events in the narrative). The counts displayed in the panel are averaged across just-watched segments. **C. Number of events hit as a function of temporal distance.** Here the (across-segment) mean numbers of events hit in participants' uncued retrodictions (orange) and predictions (blue) are displayed as a function of temporal distance to the just-watched segment (lag). Error bars denote bootstrapped 95% confidence intervals. Colors denote temporal direction (orange: past; blue: future) and distance (darker shading: onscreen events from segments adjacent to the target segment; lighter shading: offscreen events). See Figure S6 for an analogous presentation of results from our replication study.

258 rates for offscreen events. Instead, we counted up the absolute *number* of retrodicted or predicted
259 events as a function of lag. In other words, given that the participant had just watched segment i ,
260 we asked how many events from segment $i + \text{lag}$ they retrodicted or predicted, on average, given
261 that they were aiming to retrodict or predict events at lags of ± 1 . We also counted the numbers of
262 *unmatched* events in participants' responses that did not correspond to any events in the relevant
263 segments of the narrative. We focused specifically on *uncued* retrodictions and predictions, which
264 we hypothesized would provide the cleanest characterizations of participants' initial estimates of
265 the unobserved past and future (i.e., without potential biases introduced by additional character
266 information, as in the character-cued responses). For participants in our main experiment, the
267 numbers of uncued retrodicted and predicted target (lag = ± 1) events were not reliably different
268 (OR = 0.92, Z = -0.15, $p = 0.88$, CI: 0.30 to 2.84). In other words, uncued retrodictions and
269 predictions over short timescales did not exhibit reliable asymmetries. This "null result" also
270 held in our replication study (OR = XXX, Z = XXX, $p = XXX$, CI: XXX to XXX). However, when
271 retrodicting, participants in both experiments mentioned events from the distant past (lag < -1)
272 more often than participants predicted events from the distant future (lag > 1; main experiment:
273 OR = 9.10, Z = 3.80, $p < 0.001$, CI: 2.92 to 28.39; Fig. 4B, C; replication experiment: OR = XXX,
274 Z = XXX, $p = XXX$, CI: XXX to XXX; Fig. S6; for results from the character-cued conditions,
275 see Fig. S2). Despite this asymmetry in the accuracies of participants' long-range retrodictions
276 versus predictions, there were no reliable differences in the *numbers* of uncued retrodicted versus
277 predicted events (across all lags; main experiment: OR = 1.05, Z = 0.75, $p = 0.45$, CI: 0.93 to 1.18;
278 replication experiment: OR = XXX, Z = XXX, $p = XXX$). Nor did we find any reliable differences in
279 the numbers of offscreen events immediately before or after the just-watched segment (lag = ± 0.5 ;
280 main experiment: OR = 0.75, Z = -0.36, $p = 0.72$, CI: 0.15 to 3.59; replication experiment: OR
281 = XXX, Z = XXX, $p = XXX$, CI: XXX to XXX). The apparent discrepancy between participants'
282 asymmetric accuracy but symmetric event counts was due to participants' tendencies to reference
283 "unmatched" events (i.e., events that did not correspond to any explicit or implicit event in the
284 story) more in their predictions than retrodictions (main experiment: OR = 0.36, Z = -4.53,
285 $p < 0.001$, CI: 0.23 to 0.56; replication experiment: OR = XXX, Z = XXX, $p = XXX$, CI: XXX to

286 XXX). We confirmed that the retrodiction advantage held when controlling for absolute lag (main
287 experiment: OR = 34.31, Z = 3.28, $p = 0.001$, CI: 4.16 to 283.20; replication experiment: OR = XXX,
288 Z = XXX, $p = XXX$, CI: XXX to XXX), for onscreen events alone (main experiment: OR = 47.54,
289 Z = 3.74, $p < 0.001$, CI: 6.27 to 360.60; replication experiment: OR = XXX, Z = XXX, $p = XXX$, CI:
290 XXX to XXX), and marginally for offscreen events alone (main experiment: OR = 24.76, Z = 1.71,
291 $p = 0.09$, CI: 0.63 to 975.27; replication experiment: OR = XXX, Z = XXX, $p = XXX$, CI: XXX
292 to XXX). Taken together, these analyses show that (in generating uncued responses) participants
293 tend to reach “further” into the unobserved past, and with greater accuracy, than the unobserved
294 future.

295 What might be driving participants to retrodict further and more accurately into the unobserved
296 past, compared with their predictions of the unobserved future? By inspecting the video content,
297 we noticed that characters frequently referenced both past events and (planned or predicted)
298 future events in their spoken conversations. We wondered whether the characters’ references
299 might show temporal asymmetries that might explain participants’ behaviors. Across all of the
300 characters’ conversations, and across all of the video segments from our main experiment, we
301 manually identified a total of 82 references to past or future events (i.e., that occurred onscreen or
302 offscreen before or after the events depicted in the current segment; Figs. 5A, S3A, S7). Characters in
303 our main experiment’s stimulus tended to reference the past (52 references) more than the future (30
304 references), consistent with previous work (Demiray et al., 2018). References to the past were also
305 skewed to more temporally distant events compared with references to the future (Figs. 5B, S3B, S7).
306 These asymmetries also held for characters in the replication experiment’s stimulus (Fig. 8). These
307 observations indicate that the characters in the stimulus display a “preference” for the past (versus
308 future) in their conversations. Might this asymmetry be driving the asymmetries in participants’
309 retrodictions versus predictions?

310 Controlling for temporal distance (lag), past and future events that story characters referenced
311 in their conversations were associated with higher hit rates than unreferenced events in our main
312 experiment (uncued retrodiction: OR = 12.70, Z = 10.94, $p < 0.001$, CI: 8.06 to 20.03; uncued
313 prediction: OR = 8.29, Z = 6.83, $p < 0.001$, CI: 4.52 to 15.20; Fig. 5E). This indicates that partici-

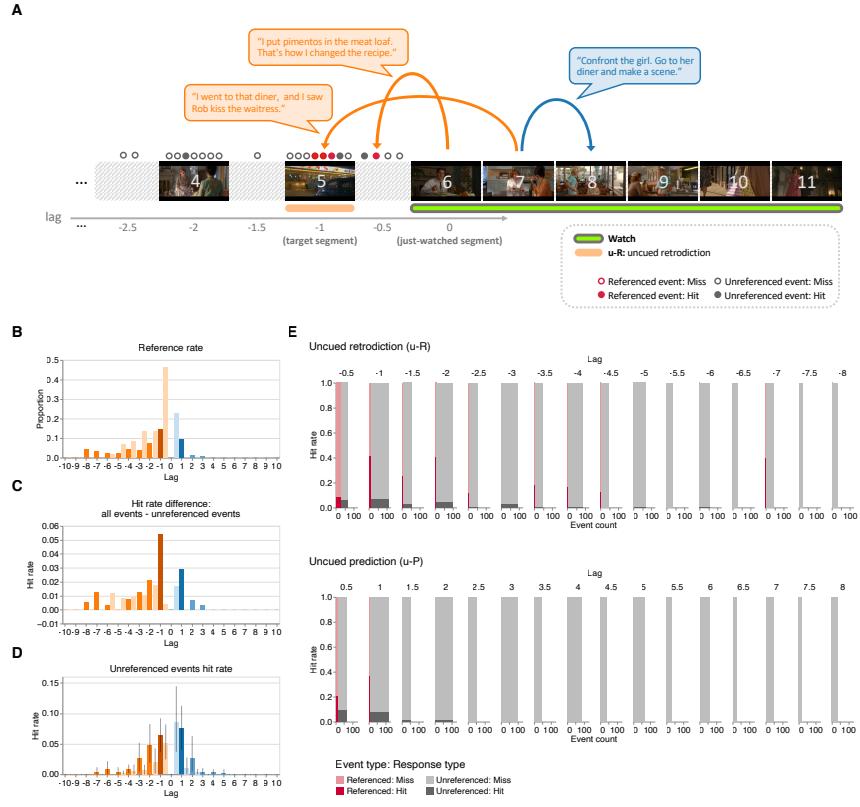


Figure 5: Characters' references drive participants' retrodiction and prediction performance. **A. Illustration of annotation approach.** We manually annotated references to events in past or future segments in characters' spoken conversations. We matched each such reference with its corresponding storyline event (and its corresponding segment number for onscreen events, or half-step segment number for offscreen events). We then tracked the hit rate separately for referenced versus unreferenced events in participants' uncued retrodictions and predictions. **B. Reference rate as a function of lag.** Across all possible just-watched segments (lag 0), the bar heights denote the average proportions of events referenced in other past (orange, negative lags) or future (blue, positive lags) segments in our main experiment's stimulus. **C. Difference in hit rates between all events and unreferenced events.** To highlight the effect of characters' references to past and future events on participants' retrodictions and predictions, here we display the difference in across-segment mean hit rates between all events and unreferenced events, as a function of temporal distance (lag) to the just-watched segment. **D. Hit rates for unreferenced events.** The average response hit rates for unreferenced events are displayed as a function of temporal distance to the just-watched segment. Error bars denote bootstrapped 95% confidence intervals. Panels B-D: colors are described in the Figure 4 caption. **E. Hit rates and counts of referenced and unreferenced events.** As a function of temporal distance to the just-watched segment, the sub-panels display the across-segment mean numbers (x -axes) and hit rates (y -axes) of referenced (red) and unreferenced (gray) events that participants hit (darker shading) or missed (lighter shading) in their uncued retrodictions (top sub-panel) and uncued predictions (bottom sub-panel). Intuitively, the widths of the rectangles at each lag denote the total number of events at each possible lag. The darker shading denotes the proportions of events that participants retrodicted or predicted, and the lighter shading denotes the proportions of events that participants "missed" in their responses. For an analogous presentation of results from the replication experiment, see Figure S7.

314 pants' responses are at least partially influenced by the characters' conversations. To estimate the
315 contributions of characters' references on hit rates, we computed the difference in hit rates between
316 all events (which comprised both referenced and unreferenced events) and unreferenced events,
317 as a function of lag. These differences exhibited a temporal asymmetry in favor of retrodiction
318 (Figs. 5C). This indicates that the asymmetries in participants' retrodictions versus predictions
319 are also at least partially influenced by the characters' conversations. However, these temporal
320 asymmetries in participants' retrodictions and predictions persisted even for events that char-
321 acters never referenced in their conversations (hit rates of uncued retrodicted versus predicted
322 unreferenced events: OR = 2.00, Z = 2.40, $p = 0.02$, CI: 1.14 to 3.51; Fig. 5D). When we further
323 separated the unreferenced events into onscreen events and offscreen events, we found that these
324 asymmetries held only for the onscreen events (onscreen: OR = 2.65, Z = 2.59, $p = 0.01$, CI: 1.27
325 to 5.54; offscreen: OR = 1.50, Z = 0.91, $p = 0.36$, CI: 0.63 to 3.62). We found similar patterns in
326 our replication experiment (Fig. S7; hit rates of uncued retrodictions for referenced events: OR =
327 XXX, Z = XXX, $p = XXX$, CI: XXX to XXX; uncued predictions for referenced events: OR = XXX,
328 Z = XXX, $p = XXX$, CI: XXX to XXX; hit rates of uncued retrodictions for *unreferenced* events: OR =
329 XXX, Z = XXX, $p = XXX$, CI: XXX to XXX; for predicted events: OR = XXX, Z = XXX, $p = XXX$, CI:
330 XXX to XXX). Taken together, these analyses suggest that asymmetries in the number of references
331 characters make to past and future events partially (but not entirely) explain why participants tend
332 to retrodict the past further and more accurately than they predict the future.

333 If characters' direct references cannot fully account for the temporal asymmetry in retrodicting
334 the unobserved past versus predicting the unobserved future, what other factors might explain this
335 phenomenon? The results above indicate that characters' references to specific unobserved events
336 in the past or future boost participants' estimates of these events. But might characters' references
337 have other effects on participants' responses *beyond* the referenced events? For example, real-world
338 experiences and events in realistic narratives are often characterized by temporal autocorrelations
339 (i.e., what is "happening now" will likely relate to what happens "a moment from now," and
340 so on). Real-world experiences and realistic narratives are also often structured into "schemas"
341 whereby experiences unfold according to a predictable pattern or formula that characterizes a

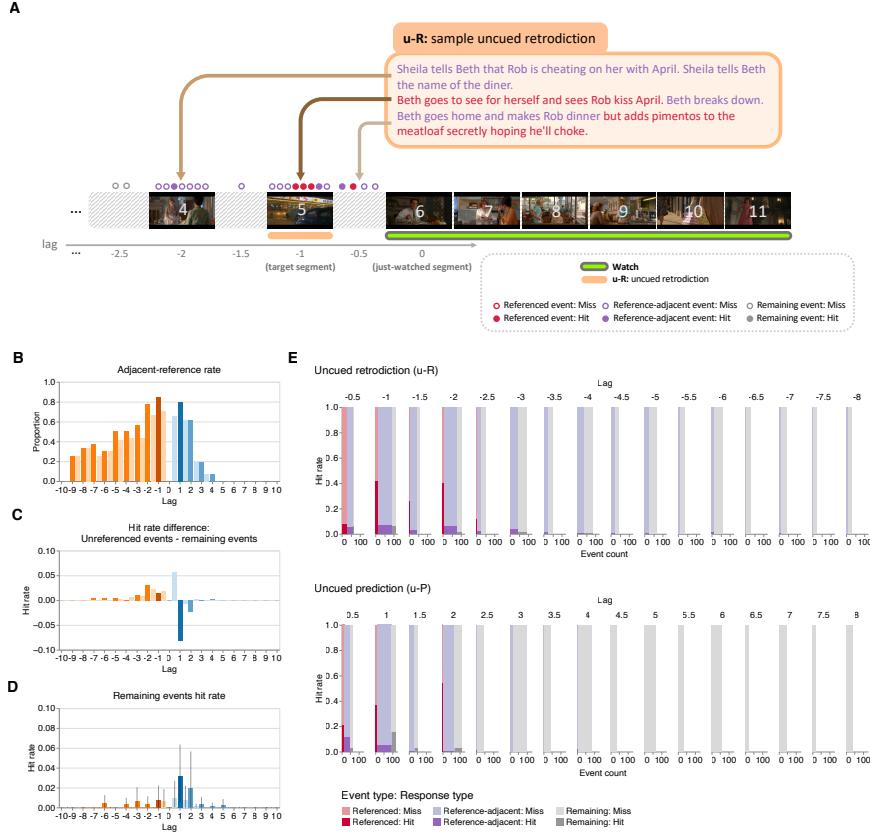


Figure 6: Reference-adjacent events are associated with higher hit rates (main experiment). **A. Illustration of annotation approach.** We extended the annotation procedure depicted in Figure 5A to also label unreferenced events that were either temporally adjacent to (i.e., immediately preceding or proceeding) a referenced event (reference-adjacent events) or not (remaining events). **B. Adjacent reference rate for unreferenced events as a function of lag.** Across all possible just-watched segments (lag 0), the bar heights denote the average proportion of unreferenced events in other past (orange, negative lags) or future (blue, positive lags) segments that were temporally adjacent to any referenced event. **C. Difference in hit rates between unreferenced events and remaining events.** To highlight the effect of reference adjacency on retrodiction and prediction of unreferenced events, here we display the difference in across-segment mean hit rates between unreferenced events and remaining events, as a function of temporal distance (lag) to the just-watched segment. **D. Hit rates for remaining events.** The across-segment mean response hit rates for unreferenced events that were *not* temporally adjacent to any referenced events are displayed as a function of temporal distance to the just-watched segment. Error bars denote bootstrapped 95% confidence intervals. Panels B–D: colors are described in the Figure 4 caption. **E. Hit rates and counts of referenced, reference-adjacent, and remaining events.** As a function of temporal distance to the just-watched segment, the sub-panels display the numbers (x-axes) and proportions (y-axes) of referenced (red), reference-adjacent (purple), and remaining (gray) events that participants hit (darker shading) or missed (lighter shading) in their uncued retrodictions (top sub-panel) and uncued predictions (bottom sub-panel). For an analogous depiction of results from our replication experiment see Fig. S8.

342 particular situation, such as going to a restaurant or catching a flight at the airport (Baldassano
343 et al., 2018). If there are associations or temporal dependencies between temporally nearby events
344 in the television show participants watched, participants might be able to pick up on these patterns
345 in forming their responses. This would be reflected in an inference “boost” for events that were
346 *nearby in time* to events that characters referred to in their conversations, in addition to the referenced
347 events themselves (Fig. 6A).

348 Because characters tended to refer to past events more often than future events, the proportions
349 of unreferenced events that were adjacent to referenced events should show a similar temporal
350 asymmetry in favor of the past. We tested this intuition by computing the proportions of unrefer-
351 enced events in the stimulus that were temporally adjacent to past or future events referenced by
352 the characters during a given segment. Here we defined *temporally adjacent* as any event within
353 an absolute lag of one relative to a referenced onscreen event, or within an absolute lag of 0.5 to a
354 referenced offscreen event. We also defined *remaining* events as unreferenced events that were not
355 temporally adjacent to any referenced events. As shown in Figure 6B, in our main experiment we
356 observed higher proportions of unreferenced past than future events that were temporally adjacent
357 to referenced events. Further, these reference-adjacent events had higher hit rates than remaining
358 events after controlling for absolute lag (uncued retrodiction: OR = 7.15, Z = 2.40, $p = 0.02$, CI: 1.44
359 to 35.58; uncued prediction: OR = 3.11, Z = 2.30, $p = 0.02$, CI: 1.18 to 8.21; Fig. 6E). These findings
360 also held in our replication experiment (uncued retrodiction: OR = XXX, Z = XXX, $p = XXX$, CI:
361 XXX to XXX; uncued prediction: OR = XXX, Z = XXX, $p = XXX$, CI: XXX to XXX; Fig. S8). To esti-
362 mate the contributions of reference adjacency on hit rates, we computed the difference in hit rates
363 between unreferenced events (which comprised both reference-adjacent and remaining events)
364 and remaining events, as a function of lag. These differences exhibited a temporal asymmetry in
365 favor of retrodiction. This suggests that reference-adjacent events also contribute to participants’
366 retrodiction advantage. Remaining events did *not* exhibit a reliable temporal asymmetry (main
367 experiment: OR = 0.75, Z = 0.33, $p = 0.74$, CI: 0.14 to 4.08, Fig. 6D; replication experiment: OR =
368 XXX, Z = XXX, $p = XXX$, CI: XXX to XXX, Fig. S8D), suggesting that, after accounting for temporal
369 adjacency, character’s references to past and future events can explain participants’ retrodiction

370 advantage.

371 The preceding analyses show that when characters reference past or future events, those refer-
372 enced events, and other events that are temporally adjacent to the referenced events, are more likely
373 to be retrodicted and predicted. In other words, referring to a past or future event in conversation
374 leads to a “boost” in that event’s hit rate. We wondered whether this boost was bi-directional. In
375 particular: when a character refers (during a *referring event*) to another event (i.e., the *referenced*
376 *event*), does this boost only the referenced event’s hit rate, or does the referring event also receive a
377 boost? We labeled each event as a “referring event,” a “referenced event,” or a “other event” (i.e.,
378 not referring or referenced; Fig. 7A, B). We limited our analysis to references to onscreen (explicit)
379 events. Consistent with our analysis of the proportions of referenced events (Fig. 5B), the propor-
380 tions of *referring* events exhibited a *forward* temporal asymmetry (Fig. 7C). Controlling for absolute
381 lag, we found that referring events were associated with lower hit rates than referenced events
382 in our main experiment (uncued retrodiction: OR = 0.03, Z = -4.81, $p < 0.001$, CI: 0.01 to 0.11;
383 uncued prediction: OR = 0.04, Z = -5.84, $p < 0.001$, CI: 0.01 to 0.12; Fig. 7D) and had no reliable
384 differences in hit rates compared with other events (uncued retrodiction: OR = 0.37, Z = -1.46,
385 $p = 0.15$, CI: 0.10 to 1.41; uncued prediction: OR = 2.16, Z = 1.68, $p = 0.09$, CI: 0.88 to 5.30). We also
386 observed this phenomenon in our replication experiment (referenced events, uncued retrodiction:
387 OR = XXX, Z = XXX, $p = XXX$, CI: XXX to XXX; referenced events, uncued prediction: OR = XXX,
388 Z = XXX, $p = XXX$, CI: XXX to XXX; other events, uncued retrodiction: OR = XXX, Z = XXX,
389 $p = XXX$, CI: XXX to XXX; other events, uncued prediction: OR = XXX, Z = XXX, $p = XXX$,
390 CI: XXX to XXX; Fig. S9). Taken together, this indicates that only referenced events received a
391 hit rate boost (relative to other events), suggesting that the retrodictive and predictive benefits of
392 references are directed (i.e., asymmetric).

393 The above analyses show that characters in the television shows we used as stimuli in our
394 main experiment and replication experiment refer more often to the past than to the future. This
395 appears to bias participants’ inferences about the past and future. But how universal is this pattern?
396 For example, were the television shows we happened to select for our experiment representative
397 of television shows more generally? Or perhaps narratives created for entertainment purposes



Figure 7: Referenced events are associated with higher hit rates, but referring events are not. **A. Illustration of annotation approach.** We extended the annotation procedure depicted in Figure 5A to also label which events in our main experiment’s stimuli contained references to events in other segments. **B. Referenced versus referring events.** During event i , when a character makes a reference to another event (j), we define i as the *referring* event and j as the *referenced* event. **C. Referring rate as a function of lag.** Across all possible just-watched segments (lag 0), the bar heights denote the across-segment mean proportions of events containing references to events in other past (orange, negative lags) or future (blue, positive lags) segments in our main experiment’s stimuli. The bar colors are described in the Figure 4 caption. **D. Hit rates and counts of referenced, referring, and other events.** As a function of temporal distance to the just-watched segment, the sub-panels display the numbers (x-axes) and hit rates (y-axes) of referenced (red), referring (green), and other (gray) events that participants hit (darker shading) or missed (lighter shading) in their uncued retrodictions (top sub-panel) and uncued predictions (bottom sub-panel). For a display of analogous results from our replication experiment see Figure S9.

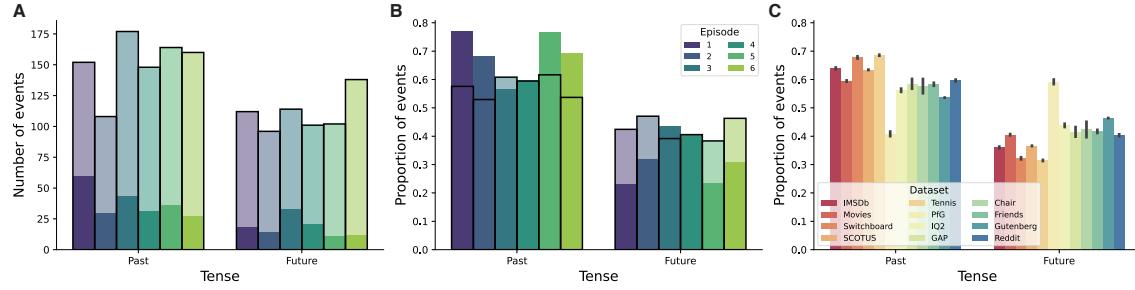


Figure 8: Meta analysis. We used natural language processing to automatically identify references to past or future events across a variety of sources. **A. Numbers of past and future events in *The Chair*, Season 1, Episodes 1–6.** The bar heights indicate the raw numbers of manually identified (lighter shading) and automatically identified (darker shading) past and future events from each episode (color). We used Episode 1 from this series as the stimulus in our replication experiment. **B. Proportions of past and future events in *The Chair*, Season 1, Episodes 1–6.** The Panel is in the same format as Panel A, but here the bar heights have been divided by the total numbers of past and future events (per episode). **C. Proportions of past and future events in movies, television shows, and natural conversations.** As in Panel B, the bar heights denote the proportions of past and future events detected in each dataset (color). The datasets are described in Table S6. Error bars denote bootstrap-estimated 95% confidence intervals.

tends to have a bias towards the past in order to keep the story engaging and unpredictable. To better understand temporal biases in conversations, we carried out a meta analysis using extracted conversation data from several large datasets, comprising over 17 million documents. The data comprised transcripts from television shows and popular films, novels, and spoken and written utterances from natural conversations. A summary of the data we analyzed may be found in Table S6. As summarized in Figure 8, we used natural language processing to identify references to past or future events in each conversation (also see *Meta analysis of conversation data*).

To validate our basic approach, we compared the numbers (Fig. 8A) and proportions (Fig. 8B) of automatically and manually identified references to past and future events, across six episodes of the television show *The Chair*. (The first episode was used as the stimulus in our replication study.) In general, our automated tagging procedure tended to overcount the numbers of references. From manually “spot checking” hundreds of example tags, we noticed that our automated tagging procedure often counts the “same” references multiple times. Specifically, the manually generated tags sought to identify references to specific events that occurred or were implied to occur in other parts of the narrative. In contrast, as a heuristic, we designed the automatic tagging procedure

413 to identify uses of the past or future *tense* as a proxy for references to past or future *events*.
414 Individual conversations often contains multiple references to a given (past or future) event.
415 Whereas the manually generated tags counted these as “single” references, our automated tagging
416 procedure has no means of differentiating between several references to the same event versus
417 the same number of references to different events. This leads the automated tagging procedure to
418 overestimate the numbers of distinct events being referenced. Nevertheless, this discrepancy did
419 not appear to bias the balance of the overall *proportions* of past or future references.

420 In all, across all of the datasets we examined in our meta analysis, we identified a total of
421 36,008,500 references to past or future events. A total of 19,464,741 (54.06%) of these were refer-
422 ences to past events, and the remaining 16,543,759 (45.94%) were references to future events. We
423 also computed the average proportions of references to past and future events across documents
424 within each individual dataset. Across the 12 datasets we examined (Fig. 8, Tab. S6), there were sig-
425 nificantly more references to the past than the to the future (mean \pm standard deviation proportion
426 of references to past events: $58.99\% \pm 7.28\%$; $t(11) = 4.28, p = 0.0013$). This bias towards the past
427 also held for each dataset individually ($ts \geq 5.14, ps < 0.01$) except for one dataset, “Persuasion
428 for Good,” which comprised natural conversations between pairs of Amazon Mechanical Turk
429 workers wherein one participant tried to convince the other participant to donate to a charity in
430 the future. In that dataset, references to the future were significantly more common than refer-
431 ences to the past ($t(11438) = -22.65, p < 0.001$). This latter example provided a nice sanity check
432 for verifying that our general approach was not itself biased in favor of the past, e.g., even in
433 conversations that were actually biased towards the future. Taken together, the results from our
434 meta analysis indicate that people tend to refer to the past more than they refer to the future, across
435 a wide variety of situations (including in both fictional and real conversations). Although (as in
436 the Persuasion for Good dataset) there may be specific exceptions to this bias, it seems that a bias
437 in favor of the past is a common element of many (and perhaps even *most*) human conversations.

438 **Discussion**

439 We asked participants in our main experiment to watch sequences of movie segments from a
440 character-driven television drama and then either retrodict what had happened prior to a just-
441 watched segment, predict what would happen next, or recall what they had just watched. We
442 found that participants tended to more accurately and more readily retrodict the unobserved
443 past than predict the unobserved future. We traced this temporal asymmetry to (a) characters'
444 tendencies to refer to past events more than future events in their ongoing conversations, and
445 (b) associations between temporally proximal events (Fig. 9). Essentially, associations between
446 temporally proximal events serve to enhance asymmetries in inferences driven by conversational
447 references (light orange and blue bars in Fig. 9). Our findings show that other peoples' psycholog-
448 ical arrows of time can affect external observers' inferences about the unobserved past and future.
449 We confirmed our main behavioral findings in a pre-registered replication study. We also carried
450 out a meta analysis of tens of millions of utterances from television shows, movies, novels, and
451 natural spoken and written conversations. We found that the tendency to refer more often to the
452 past than the future appears to be a widespread characteristic of human conversation.

453 When people communicate through language or other observable behaviors, they can transmit
454 their knowledge and memories to others (Hirst and Echterhoff, 2012; Mahr and Csibra, 2018;
455 Dessalles, 2007; Zadbood et al., 2017). A consequence of this sharing across people is that biases or
456 limitations in one person's knowledge and memories may also be transmitted to external observers.
457 Although people *can* communicate their intentions and future plans (i.e., information about their
458 future), because people know *more* about their pasts than their futures, the knowledge transmitted
459 to observers is inherently biased in favor of the past (Fig. 9; Demiray et al., 2018). Since observers
460 leverage communicated knowledge to reconstruct the unobserved past and future, this explains
461 why observers' inferences about observed people's lives also favor the past.

462 People's knowledge asymmetries are not always directly observable. For example, in a con-
463 versation where someone talks exclusively about their future plans, a passive observer might gain
464 more insight into the speaker's unobserved future than their unobserved past. However, because

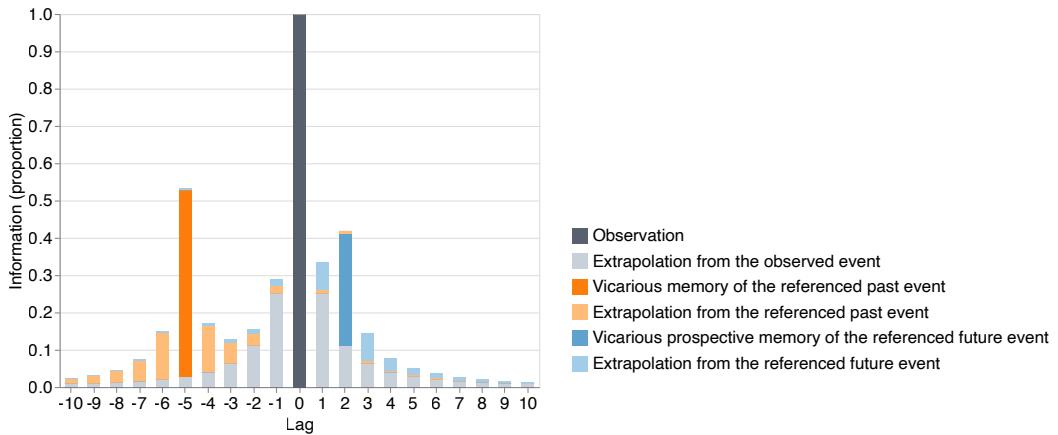


Figure 9: How much information about the past and future can be inferred by observing the present? By definition, let us say that the present moment (lag 0) contains all information about itself (dark gray). Given learned statistical regularities, one might extrapolate from the present moment into the past or future (light gray). As illustrated in this schematic, the information contained in the present about other moments in time falls off with absolute lag. This falloff is approximately time-symmetric. References in the present to past events (dark orange) or future events (dark blue) provide additional information about those referenced moments in time, beyond what could be inferred solely from statistical regularities. This additional information about those referenced moments can also be extrapolated to other moments that are temporally nearby to them (light orange and blue). The data in this schematic are hypothetical.

465 the speaker is also guided by their own psychological arrow of time, the “upper limit” of knowledge
466 about their past is still higher than that of their future. Therefore, after accounting for knowledge
467 that *could* be revealed through active participation in the conversation, the seemingly future-biased
468 conversation masks an underlying knowledge asymmetry in favor of the past. This hypothesized
469 “unmasking” effect of interaction implies that the influence of other people’s psychological arrows
470 of time should be more robust when the receiver is an active participant in the conversation. Other
471 social dimensions, such as trust, motivation or level of engagement, personal goals, and beliefs,
472 might serve to modulate the effective “gain” of the communication channel– i.e., how much the
473 speaker’s knowledge influences the observer’s knowledge. Some recent work (e.g., Tamir and
474 Mitchell, 2013; Meyer et al., 2019) also suggests that people might gain insights into other people
475 using “mental simulations” of how they might respond in particular situations (e.g., in the future),
476 or of which sorts of prior experiences might have led someone to behave a particular way in the
477 present.

478 In typical statistical sequences used in laboratory studies, there is no temporal asymmetry,
479 either theoretically (Cover, 1994; Bialek et al., 2001; Ellison et al., 2009), or empirically (Jones and
480 Pashler, 2007). What makes narratives and real-world event sequences time-asymmetric? Of
481 course there are many superficial differences between simple laboratory-manufactured sequences
482 and real-world experiences. As one example, real-world experiences often involve other people
483 who have their own memories and goals. At a deeper level, however, are our subjective experi-
484 ences essentially more complicated versions of laboratory-manufactured sequences? Or are there
485 fundamental differences? One possibility is that real-life event sequences are not stationary (i.e.,
486 not in equilibrium; Cover, 1994). For example, real-life events might start from a special initial
487 condition (Albert, 2000; Feynman, 1965; Cover, 1994) and proceed through a series of transitions
488 from more-ordered to less-ordered states, thus exhibiting an arrow time. When we retrodict, it is
489 possible that we only consider possible past events that are compatible with the highly-ordered
490 special initial state (Carroll, 2010, 2016). For example, when we see a broken egg we might infer
491 that the egg had been intact at some point in the past. But it would be difficult to guess at what
492 states or forms the broken egg might take in the future (Carroll, 2010, 2016). In other words, the

493 procession from order to disorder might result in better retrodiction performance compared with
494 that of (implicitly less-restricted) prediction tasks. The special initial state might also explain why
495 we remember the past, but not the future. Some recent work suggests that the psychological arrow
496 of time might be explained by a related concept in the statistical physics literature, termed the
497 “thermodynamic” arrow of time (Mlodinow and Brun, 2014; Rovelli, 2022). However, the relation
498 between the thermodynamic and psychological arrows of time is still under debate (Gołosz, 2021;
499 Hemmo and Shenker, 2019).

500 Beyond forming inferences about unobserved past and future events, our work also relates
501 to prior studies of how people perceive time (Block and Gruber, 2014; Howard, 2018; Eagleman,
502 2008; Ivry and Schlerf, 2008; Wearden, 2016), and how we “move” through time in our memories
503 of our past experiences (Manning, 2021; Manning et al., 2011; Howard et al., 2012; Manns et al.,
504 2007; Shankar and Howard, 2012; Kahana, 1996; Polyn and Kahana, 2008; Schacter and Tulving,
505 1994) or in our imagined (past or future) experiences (Schacter, 2012; Josselyn and Tonegawa, 2020;
506 Schacter et al., 1998; Momennejad and Howard, 2018). For example, a well-studied phenomenon
507 in the episodic memory literature concerns how remembering a given event cues our memories of
508 other events that we experienced nearby in time (i.e., the *contiguity effect*; Kahana, 1996). Across
509 a large number of studies there appears to be a nearly universal tendency for people to move
510 *forwards* in time in their memories, whereby recalling an “event” (e.g., a word on a previously
511 studied list) is about twice as likely to be followed by recalling the event that immediately followed
512 as compared with the event immediately preceding the just-recalled event (Healey and Kahana,
513 2014). Superficially our current study appears to report the *opposite* pattern, whereby participants
514 display a *backwards* temporal bias. However, the two sets of findings may be reconciled when
515 one considers the frame of reference (and current mental context; e.g., Howard and Kahana, 2002)
516 of the participant at the moment they make their response. In our study, participants observe
517 an event in the present, and they make guesses about what happened in the unobserved past or
518 future, relative to the just-observed event. (Our findings imply that participants are more facile
519 at moving backwards in time than forwards in time, relative to “now.”) In contrast, the classic
520 contiguity effect in episodic memory studies refers to how people move through time relative to

521 a just *remembered* event. The forward asymmetry in the contiguity effect follows from the notion
522 that the moment of remembering has greater contextual overlap with events *after* the remembered
523 event from the past than events that happened before it (for review also see Manning et al., 2015;
524 Manning, 2020).

525 In our study, we explicitly designed participants' experiences such that both the past and future
526 were unobserved. How representative is this scenario of everyday life? For example, we might
527 try to speculate about the unobserved future when making plans or goals, but when might we
528 encounter situations where the past is unobserved but still useful for us to speculate about? Real-life
529 events have long-range dependencies. In general, because the future depends on what happened
530 in the past, discovering or estimating information about the unobserved past can help us form
531 predictions about the future. We illustrate this point in Figure 9 by showing that the additional
532 information contributed by a referenced past event can also extend into the future (light orange bars
533 at lags > 0). This might explain why humans devote substantial effort and resources to attempting
534 to figure out what happened in the unobserved past: history, anthropology, geology, detective and
535 forensic science, and other related fields are each primarily focused on understanding, retrodicting,
536 or reconstructing unobserved past events.

537 Methods

538 Participants

539 **Main experiment.** A total of 36 participants (25 female, mean age 21.47 years, range 19–50 years)
540 were recruited from the Dartmouth College community for our main experiment. All participants
541 had self-reported normal or corrected-to-normal vision, hearing, and memory, and had not watched
542 any episodes of *Why Women Kill* before the experiment. Participants gave written consent to enroll
543 in the study under a protocol approved by the Committee for the Protection of Human Subjects at
544 Dartmouth College. Participants received course credit or monetary compensation for their time.
545 Two participants completed only the first half of the study and one participant's data from the

546 second half of their testing session was lost due to a technical error. All available data were used
547 in the analyses.

548 **Replication experiment.** A total of XXX participants (XXX female, mean age XXX years, range
549 XXX–XXX years) were recruited from the Dartmouth College community for our pre-registered
550 replication experiment. All participants had self-reported normal or corrected-to-normal vision,
551 hearing, and memory, and had not watched any episodes of *The Chair* before the experiment.
552 Participants gave written consent to enroll in the study under a protocol approved by the Commit-
553 tee for the Protection of Human Subjects at Dartmouth College. Participants received monetary
554 compensation for their time. All available data were used in the analyses.

555 **Stimuli**

556 **Main experiment.** The stimuli used in our main experiment were segments of the CBS television
557 series *Why Women Kill* Season 1. The TV series contained three distinct storylines depicting three
558 women's marital relationships. The three storylines, which took place in the 1960s, 1980s, and
559 2019, were shown in an interleaved fashion in the original episodes. The first 11 segments from the
560 1960s and 1980s storylines, across the first and second episodes, were used in our study. Segments
561 were divided based on major scene cuts, which primarily corresponded to storyline shifts in the
562 original episodes. The mean length of the segments was 2.05 min (range 0.97–3.87 min). We chose
563 this TV series based on its strictly linear storytelling (within each storyline) and its realistic settings
564 where most events depicted everyday life. The plots were focused on the main characters (Beth in
565 storyline 1 and Simone in storyline 2), who were present in all the segments in the corresponding
566 storylines.

567 **Replication experiment.** The stimuli used in our replication experiment were segments of the
568 first episode of the Netflix television show *The Chair*, Season 1. **JRM NOTE: Describe the show,**
569 **like you did for Why Women Kill.** The mean length of the segments was XXX min (range
570 XXX–XXX min). As for the stimulus we used in our main experiment, we chose this stimulus for

571 our replication experiment for its linear storytelling (again, within each storyline) and its realistic
572 depictions of everyday events. *JRM NOTE: The plots were focused on... (fill in something analogous to*
573 *what you wrote for the main experiment stimulus...)*

574 **Task design and procedure**

575 **Main experiment.** Our experimental paradigm was divided across two testing sessions. In each
576 session, participants performed a sequence of tasks on segments from one storyline (Fig. 2). For
577 each storyline, there were four different task sequences: two forward chronological order sequences
578 and two backward chronological order sequences. Participants completed one task sequence in
579 forward chronological order for one storyline, and one in backward chronological order for the
580 other storyline. The order of the two sessions (forward chronological order sequence first or
581 backward chronological order sequence first), and the pairing of task sequences with storylines,
582 were counterbalanced across participants.

583 Tasks in each sequence alternated between watching, recall, and retrodiction or prediction,
584 with the specific order of tasks differing across the four sequences. For example, in sequence A1,
585 participants first watched segment 1, followed by an immediate recall of segment 1. Then they
586 predicted what would happen in segment 2 (first uncued and then character-cued). Participants
587 then watched segment 3 and recalled segment 3. After that, participants guessed what happened in
588 segment 2 again, which we termed “updated prediction”. Then they watched segment 2, recalled
589 segment 2, and so on as depicted in Figure 2. This procedure was repeated to cover all possible
590 segments. We also note several edge cases at the start and end of the narrative sequences. Since
591 no segments precede the first segment, participants could never make “prediction” responses with
592 the first segment as their target. For analogous reasons, participants never made “retrodiction”
593 responses with the last segment as their target. Another edge case occurred in task sequences
594 B2 and A2 (Fig. 2). In the A1 and A2 sequences, participants experience the narrative in the
595 original (forward) order, predicting one segment ahead along the way. In the B1 and B2 sequences,
596 participants experience the narrative in the reverse order, retrodicting one segment ahead along
597 the way. However, because A2 and B2 are offset from A1 and B1 by one segment, the initial A2

598 responses are *retrodictions*, and the initial B2 responses are *predictions* (i.e., they conflict with the
599 temporal directions of the remaining responses in those conditions). We therefore excluded from
600 our analysis those initial retrodiction responses from the A2 condition, and the initial prediction
601 responses from the B2 condition.

602 Before watching each segment, participants were given the following task instructions. After
603 watching the video, participants were instructed to type their responses (retrodiction, prediction,
604 or recall) in 1–4 sentences. Participants were also asked to specify the characters' names in their
605 responses, i.e., avoiding use of characters' pronouns. For the recall task, the names of the characters
606 in the recall segment were displayed, and participants were asked to summarize the major plot
607 points in the present tense. For the retrodiction and prediction tasks, participants were instructed
608 to retrodict or predict the major plot points of the segment (also in the present tense), as though
609 they had watched the segment and were writing a plot synopsis. They were also instructed to
610 avoid speculation words (e.g., “I *think* Beth will...”). For the uncued retrodiction and prediction
611 tasks, participants made retrodictions or predictions without any cues provided, so they had to
612 guess which of the characters would be present in the segment. For character-cued retrodictions
613 and predictions, the characters in the target segment were revealed on the screen, alongside
614 participants' previous responses. Participants were instructed to include or incorporate those
615 characters into their character-cued responses, if their previous responses did not contain all the
616 characters provided. They were also told that the characters were not necessarily listed in their
617 order of appearance in the segment, and that only the main characters would be given. Also, the
618 characters given did not necessarily interact with each other in that segment, and they could appear
619 in successive events in that segment. If participants' previous responses included all the characters
620 given, then they could directly proceed to the next task without updating their responses. For
621 all of the prediction and retrodiction tasks, participants were instructed to provide at least one
622 response, but they were given the opportunity enter up to three responses if they felt that multiple
623 possibilities were more or less equally likely. Each response (including recall) was followed by a
624 confidence rating on a 1–5 point scale. However, these confidence data were not analyzed in the
625 present study.

626 Before their first testing session, participants were given a practice session, where they watched
627 the first segment of storyline 3 followed by a recall trial, an uncued prediction trial, and a character-
628 cued prediction trial. Participants' responses were checked by the experimenter to ensure compli-
629 ance with the instructions. To provide participants with sufficient background information about
630 the storyline (especially for the backward chronological sequences), at the beginning of each ses-
631 sion, participants were shown the time, location, and the main characters (with pictures) of the
632 storyline. The first session was approximately 1.5 h long and the second session was approximately
633 1 h long. We allowed participants, at their own discretion and convenience, to sign up for two
634 consecutive testing time-slots (i.e., with their testing sessions occurring in immediate succession),
635 or for testing sessions on two different days. The mean inter-session interval was 0.73 days (range:
636 0–4 days). The experiment was conducted in a sound- and light-attenuated testing room. Videos
637 were displayed using a 27-inch iMac desktop computer (resolution: 5120 × 2880) and sound was
638 presented using the iMac's built-in speakers. The experiment was implemented using jsPsych (de
639 Leeuw, 2015) and JATOS (Lange et al., 2015).

640 **Replication experiment.** JRM NOTE: briefly describe replication experiment methods, refer-
641 ring to Fig. S4. Since the methods will have been similar for the replication study, just highlight
642 the differences rather than re-describing everything.

643 Video annotation

644 **Main experiment.** Events in the first 11 segments of the two storylines were identified by the
645 first author (X.X.), corresponding to major plot points (total: 117; mean: 5.32 per segment; range
646 3–9). Additionally, 74 offscreen events were identified. Of these 74 offscreen events, 43 events
647 were identified from references in conversations during onscreen events. Another 16 events were
648 identified based on characters' implied movements and travels. For example, if in segment 1
649 character A was in place A and in segment 2 she was in place B, then the transit from place A to B
650 for character A would be identified as an offscreen event. The remaining 15 offscreen events were
651 identified based on logical inferences. For example, if a photograph was shown in an onscreen

652 event (but not the act of the photograph being taken), then the action that someone took the
653 photograph would be identified as an offscreen event. Offscreen events always occurred between
654 two contiguous segments, or before the first segment. The purpose of identifying offscreen events
655 was to match participants' responses to video events; thus our identification of these offscreen
656 events was not intended to be exhaustive.

657 **Replication experiment.** Events in the first XXX segments of the two storylines were identified
658 by the first author (X.X.), corresponding to major plot points (total: XXX; mean: XXX per segment;
659 range XXX–XXX). Additionally, XXX offscreen events were identified. Of these XXX offscreen
660 events, XXX events were identified from references in conversations during onscreen events.
661 Another XXX events were identified based on characters' implied movements and travels. The
662 remaining XXX offscreen events were identified based on logical inferences.

663 Response analyses

664 Participants' retrodiction, prediction, and recall responses were minimally processed to correct
665 obvious typos (e.g., in characters' names) and remove speculation descriptions (e.g., "I predict
666 that..."). All responses were manually coded and matched to events from the video annotations.
667 Retrodiction and prediction responses were coded by two coders (main experiment: X.X. and Z.Z.;
668 replication experiment: X.X. and X.Z.). Recall responses were coded by one coder (X.X.). While
669 most responses were clearly identifiable as either matching specific storyline events or as not
670 matching any storyline events, several ambiguous cases arose. First, some responses combined or
671 summarized over several (distinct) storyline events. Second, some responses lacked any specific
672 detail (e.g., "character A and B talk" without describing the specific topic(s) of conversation or
673 providing other relevant details). Based on participants' responses, in addition to the original
674 117 onscreen events and 74 offscreen events in the main experiment's stimulus, we added 25 new
675 events (23 onscreen, 2 offscreen) that either summarized across several events or partially matched
676 the annotated events. In our replication study we used the same procedure to add XXX new events
677 (XXX onscreen, XXX offscreen). Whereas the original events were each assigned a value of one

678 point, we assigned these additional events a half point. This point system enabled us to directly
679 match events in participants' responses to the annotated events. In our analyses of retrodictions,
680 predictions, and recalls, we added up the number of points earned for each response to estimate
681 participants' event hit rates.

682 We coded only the first retrodiction or prediction response in each trial. For these responses,
683 we also only considered storyline events that were in the same temporal direction as the target
684 segment. For example, if a participant was asked to retrodict what happened in segment n , only
685 events from segments 1... n were considered in our analysis. When coding recall responses, we
686 considered only events from the target segment.

687 An additional ambiguous case arose in one main experiment participant's responses pertaining
688 to segment 12, storyline 2, whereby the participant correctly identified an onscreen event that had
689 not been included in our original annotations. To account for this participant's response, we
690 retroactively added that event to our annotations of that segment. We also identified and counted
691 unmatched events in participants' responses (i.e., events that did not match any annotated events).
692 Cases where the two coders' independent scoring disagreed were resolved through discussions
693 between the two coders.

694 To estimate the semantic similarities between pairs of responses, we first transformed each
695 response into a 512-dimensional vector (embedding) using the Universal Sentence Encoder (Trans-
696 former USE, Cer et al., 2018). We defined *similarity* as the cosine of the angle formed by the
697 responses' vectors. Following Heusser et al. (2021), we defined the *precision* of participants' re-
698 sponses as the median similarity between that response's vector and the embedding vectors for
699 all other participants' recalls of the target segment. We defined the *convergence* of a given response
700 as the mean similarity between that response's vector and all other participants' responses to the
701 corresponding segment, in the same condition. To compute these median or mean similarities we
702 first applied the Fisher z-transformation to the similarity values, then took the median or mean
703 of the z-transformed similarities, and finally applied the inverse z-transformation to obtain the
704 precision or convergence score.

705 To test the validity and reliability of the USE embeddings, we performed a classification analysis

706 of recall responses using a leave-one-out approach. For each recall response, we calculated its
707 semantic similarity with all other recall responses for the same storyline. We took the segment
708 with the highest median semantic similarity (to the recall response) as the “predicted” segment.
709 Across all responses, the predicted segments matched the true recalled segments’ labels 98.6% of
710 the time (1088 out of 1103 predictions; chance level: 9%). We note that this validation analysis
711 could only be carried out with data from our main experiment, since we did not collect recall
712 responses in our replication experiment.

713 Reference coding

714 Two coders (main experiment: X.X. and Z.Z.; replication experiment: X.X. and X.Z.) identified
715 character dialogues in the narrative that referred to past events or future (onscreen or offscreen)
716 events. Only references to events that occurred in a different segment were included in this tagging
717 procedure. For each reference, the source (referring) segment and the referred event number were
718 recorded. A total of 82 references were identified in the main experiment stimulus, and XXX were
719 identified in the replication experiment stimulus. Of these references in the main experiment, 30
720 referred to onscreen events and 52 referred to offscreen events. In the replication experiment, XXX
721 referred to onscreen events and XXX referred to offscreen events. For these referenced events, their
722 corresponding summary events or partial events were also labelled as referenced. In instances
723 where the coders disagreed about a given tag, disagreements were resolved through discussions
724 between the two coders. In our analyses, each storyline event was coded according to whether
725 or not it had been referenced in the segment(s) that the participant had viewed thus far in the
726 experiment.

727 In principle, a given event could receive multiple labels. For example, during event *A*, a
728 character might speak about another event, *B*, during which a reference to a third event (*C*) was
729 made. In this scenario, event *B* could be both a “referring event” ($B \rightarrow C$) and a referenced event
730 ($A \rightarrow B$). In practice, however, this scenario was quite rare, accounting for only one out of a total
731 of 30 onscreen events in our main experiment and XXX events in our replication experiment.

732 **Statistical analysis**

733 We used (generalized) linear mixed models to analyze the hit rates and numbers of events retrodicted,
734 predicted, and recalled, as well as the precisions and convergences of participants' responses.
735 Our models were implemented in R using the `afer` package. We carried out comparisons or contrasts,
736 and extracted *p*-values, using the `emmeans` package. Participants and stimuli (e.g., segment
737 identity) were modeled as crossed random effects (as specified below). Random effects were selected
738 as the maximal structure that allowed model convergence. All of our statistical tests were
739 two-sided.

740 For our tests of the target event hit rates across four levels (uncued, character-cued, updated,
741 and recall; Fig. 3B), we fit a generalized linear mixed model with a binomial link function:

```
742   cbind(thp, ttp - thp) ~ direction * level * seg_cnt * storyline +  
743     (direction * level | target) +  
744     (direction * level * seg_cnt | subject)
```

745 where for analyses of our main experiment `thp` was the number of points hit for the target segment,
746 `ttp` was the total number of points for the target segment (from its annotations), `direction` was
747 either retrodiction or prediction, `level` had four levels (uncued, character-cued, updated, and
748 recall), `seg_cnt` represented the number of segments in the storyline that had been watched (1–10,
749 centered), `storyline` had two levels (1 or 2), and `target` had 22 levels according to the identity of
750 the target segment. For our analyses of our replication experiment, `level` had two levels (uncued
751 and character-cued), the `storyline` parameter was omitted since there was only a single storyline,
752 and `target` had XXX levels according to the identity of the target segment.

753 For our tests of precision and convergence (Fig. 3C, D), we fit linear mixed models using the
754 same formula. To test the effect of `direction` (retrodiction or prediction) on target event hit rates,
755 precision, and convergence, we fit a (generalized) linear mixed model separately for each of the
756 three levels (uncued, character-cued, and recall).

757 For our tests comparing the numbers of hits for different types of events (Fig. 4B), we fit
758 generalized linear mixed models using the same formula, but with a Poisson link function. For

759 these models, we manually doubled the point counts to ensure that half points were mapped onto
760 integers, ensuring compatibility with the Poisson link function.

761 For our analyses of the numbers of events hit, controlling for lag (Fig. 4C), we fit a generalized
762 linear mixed model with a Poisson link function:

```
763 hp_lag ~ direction * full_stp * lag * storyline +  
764 (direction | base_seg) + (1 | base_seg_pair) +  
765 (direction * full_stp * lag * storyline | subject)
```

766 where `hp_lag` is the number of “points” earned (for each lag) in each trial (we manually doubled
767 the point counts to ensure that half points were mapped onto integers, for compatibility with the
768 Poisson link function), `full_stp` denoted whether the given events (of the given lag) were onscreen
769 (i.e., full step) or offscreen (i.e., half step), `lag` denotes the (centered) absolute lag, `base_seg` denotes
770 the identity of the just-watched segment (main experiment: 22 levels; replication experiment: XXX
771 levels), and `base_seg_pair` denotes the pairing of the just-watched segment and the segment at
772 each lag (main experiment: 440 levels; replication experiment: XXX levels).

773 For our analyses of the proportions of events hit for referenced versus unreferenced events
774 (Fig. 5D, E), we fit a generalized linear model with a binomial link function:

```
775 cbind(hp_lag, tp_lag - hp_lag) ~ direction * reference * full_stp +  
776 lag + (direction | base_seg) +  
777 (1 | base_seg_pair) +  
778 (direction * reference * full_stp + lag | subject)
```

779 where `hp_lag` denotes the number of earned hit points for each reference type (referenced or
780 unreferenced) at each lag, `tp_lag` denotes the total number of possible hit points for each reference
781 type at each lag, and the other variables adhered to the same notation used in the above formulas.

782 For our tests of the proportions of events hit for all three reference types (referenced, reference-
783 adjacent, and remaining; Fig. 6D, E; or referenced, referring, and other; Fig. 7D), we fit a generalized
784 linear mixed model using the same formula as above, but with three (rather than two) reference
785 levels.

786 Several of our analyses entailed comparing the relative hit rates or probabilities of two different
787 conditions or outcomes. We used the `emmeans` package to compute the odds ratios given the
788 generalized linear mixed models we fit for the given analysis. These odds ratios reflect the
789 chances (“odds”) of a particular outcome (e.g., making a response about a particular event) given
790 a scenario (e.g., the event occurred *prior* to the just-watched segment) compared with the chances
791 of the outcome occurring in the alternative scenario (e.g., the event occurred *after* the just-watched
792 segment).

793 Meta analysis

794 At a high level, the goal of our meta analysis was to predict in-text references to past and future
795 events. Manually identifying these references is labor and time intensive, so it is impractical to scale
796 up manual tagging to millions of documents. Instead, we defined a set of heuristics for *predicting*
797 when text is referring to real or hypothetical past or future events. Our approach comprises four
798 main steps.

799 First, we use the `nltk` package (Bird et al., 2009) to segment each document into individual
800 sentences. Each sentence is processed independently of the others. Second, we handle contractions
801 using the `contractions` package (e.g., “we’ll” is split into “we will,” and so on). Third, we define
802 two sets of “keywords” (words and phrases) that tend to be indicative of referring to the past
803 (Tab. S4) or future (Tab. S5). We used ChatGPT (OpenAI, 2023) to generate each list, with exactly
804 50 templates per list, using the following prompt:

805 I’m designing a heuristic algorithm for identifying references (in text) to
806 past and future events. Part of the algorithm will involve looking for specific
807 keywords or phrases that suggest that the text is referring to something that
808 happened (or will happen) in the past and/or future. Could you help me generate
809 a list of 50 keywords or phrases to include in each list (one list for identifying
810 references to the past and a second list for identifying references to the
811 future)? I’d like to be able to paste the lists you generate into two plain

812 text documents with one row per keyword or phrase, and no other content. Please
813 output the lists as a "code" block (enclosed by '```').

814 Fourth, we use part-of-speech tagging (again, using the `nltk` package) to look for verbs or verb
815 phrases that are in past or future tenses. After the words were tagged with their predicted parts
816 of speech, we use regular expressions (applied to the sequences of tags) to label each verb or verb
817 phrase with a human readable verb form (e.g., "future perfect continuous passive," "conditional
818 perfect continuous passive," and so on). The regular expressions we used to generate these labels
819 are shown in Table S2, and the part of speech tags are defined in Table S3.

820 We treated each keyword match (of past or future keywords) as a single "reference" (to a past or
821 future event, respectively), and if any past or future verb forms were detected we treat those as (up
822 to) one additional reference. We then tallied up the numbers of past and/or future references across
823 sentences within the given document. The meta analysis results reported in Figure 8C display the
824 average numbers of references aggregated across all documents within each dataset we analyzed
825 (described in Tab. S6).

826 **Code and data availability**

827 All of the code and data generated for the current manuscript are available online at:

828 <https://github.com/ContextLab/prediction-retrodiction-paper>

829 **References**

830 Albert, D. Z. (2000). *Time and chance*. Harvard University Press, Cambridge, Mass.

831 Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas
832 during narrative perception. *The Journal of Neuroscience*, 38(45):9689–9699.

833 Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural
834 Computation*, 13(11):2409–2463.

- 835 Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text with*
836 *the natural language toolkit*. Reilly Media, Inc.
- 837 Block, R. A. and Gruber, R. P. (2014). Time perception, attention, and memory: a selective review.
838 *Acta Psychologica*, 149:129–133.
- 839 Bordwell, D. (2008). *Poetics of cinema*, chapter Three dimensions of film narrative, pages 85–134.
840 Routledge.
- 841 Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*,
842 11(2):177–220.
- 843 Carroll, S. (2010). *From eternity to here: the quest for the ultimate theory of time*. Penguin.
- 844 Carroll, S. (2016). *The big picture: on the origins of life, meaning, and the universe itself*. Dutton.
- 845 Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,
846 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.
847 *arXiv*, 1803.11175.
- 848 Cover, T. M. (1994). Which processes satisfy the second law? In Halliwell, J. J., Pérez-Mercader,
849 J., and Zurek, W. H., editors, *Physical Origins of Time Asymmetry*, pages 98–107. Cambridge
850 University Press, Cambridge, UK.
- 851 de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web
852 browser. *Behavior Research Methods*, 47(1):1–12.
- 853 Demiray, B., Mehl, M. R., and Martin, M. (2018). Conversational time travel: evidence of a
854 retrospective bias in real life conversations. *Frontiers in Psychology*, 9:2160.
- 855 Dessalles, J.-L. (2007). Storing events to retell them. *Behavioral and Brain Sciences*, 30(3):321–322.
- 856 Eagleman, D. M. (2008). Human time perception and its illusions. *Current Opinion in Neurobiology*,
857 18(2):131–136.

- 858 Ellison, C. J., Mahoney, J. R., and Crutchfield, J. P. (2009). Prediction, retrodiction, and the amount of
859 information stored in the present. *Journal of Statistical Physics*, 136(1005):doi.org/10.1007/s10955–
860 009–9808-z.
- 861 Feynman, R. (1965). *The character of physical law*. MIT Press.
- 862 Gołosz, J. (2021). Entropy and the direction of time. *Entropy*, 23(4):388.
- 863 Hawking, S. W. (1985). Arrow of time in cosmology. *Physical Review D*, 32(10):2489–2495.
- 864 Healey, M. K. and Kahana, M. J. (2014). Is memory search governed by universal principles or
865 idiosyncratic strategies? *Journal of Experimental Psychology: General*, 143(2):575–596.
- 866 Hemmo, M. and Shenker, O. (2019). The second law of thermodynamics and the psychological
867 arrow of time. *The British Journal for the Philosophy of Science*.
- 868 Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behavioral and
869 neural signatures of transforming experiences into memories. *Nature Human Behavior*, 5:905–919.
- 870 Hirst, W. and Echterhoff, G. (2012). Remembering in conversations: the social sharing and reshaping
871 of memories. *Annual Review of Psychology*, 63(1):55–79.
- 872 Horwich, P. (1987). *Asymmetries in time: problems in the philosophy of science*. MIT Press.
- 873 Howard, M. W. (2018). Memory as perception of the past: compressed time in mind and brain.
874 *Trends in Cognitive Sciences*, 22(2):124–136.
- 875 Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal
876 of Mathematical Psychology*, 46:269–299.
- 877 Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL
878 neurons “jump back in time” in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.
- 879 Ivry, R. B. and Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends in
880 Cognitive Sciences*, 12(7):273–280.

- 881 Jones, J. and Pashler, H. (2007). Is the mind inherently forward looking? comparing prediction and
882 retrodiction. *Psychonomic Bulletin and Review*, 14(2):295–300.
- 883 Josselyn, S. A. and Tonegawa, S. (2020). Memory engrams: recalling the past and imagining the
884 future. *Science*, 367(6473):eaaw4325.
- 885 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*, 24:103–109.
- 886 Koster-Hale, J. and Saxe, B. (2013). Theory of mind: A neural prediction problem. *Neuron*,
887 79(5):836–848.
- 888 Lange, K., Kühn, S., and Filevich, E. (2015). “Just Another Tool for Online Studies” (JATOS): an
889 easy solution for setup and management of web servers supporting online studies. *PLoS One*,
890 10(6):e0130834.
- 891 Maheu, M., Meyniel, F., and Dehaene, S. (2022). Rational arbitration between statistics and rules
892 in human sequence processing. *Nature Human Behaviour*, pages 1–17.
- 893 Mahr, J. B. and Csibra, G. (2018). Why do we remember? the communicative function of episodic
894 memory. *Behavioral and Brain Sciences*, 41:e1.
- 895 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors, *Handbook
896 of Human Memory*. Oxford University Press.
- 897 Manning, J. R. (2021). Episodic memory: mental time travel or a quantum “memory wave”
898 function? *Psychological Review*, 128(4):711–725.
- 899 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory.
900 In Gazzaniga, M., editor, *The Cognitive Neurosciences*, pages 557–566. MIT Press.
- 901 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns
902 in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National
903 Academy of Sciences, USA*, 108(31):12893–12897.

- 904 Manns, J. R., Howard, M. W., and Eichenbaum, H. (2007). Gradual changes in hippocampal activity
905 support remembering the order of events. *Neuron*, 56(3):530–540.
- 906 Meyer, M. L., Zhao, Z., and Tamir, D. I. (2019). Simulating other people changes the self. *Journal of*
907 *Experimental Psychology: General*, 148(11):1898–1914.
- 908 Mlodinow, L. and Brun, T. A. (2014). Relation between the psychological and thermodynamic
909 arrows of time. *Physical Review E*, 89(5):052102.
- 910 Momennejad, I. and Howard, M. W. (2018). Predicting the future with multi-scale successor
911 representations. *bioRxiv*, page doi.org/10.1101/449470.
- 912 OpenAI (2023). ChatGPT. Personal communication.
- 913 Polyn, S. M. and Kahana, M. J. (2008). Memory search and the neural representation of context.
914 *Trends in Cognitive Sciences*, 12:24–30.
- 915 Radvansky, G. A. and Copeland, D. E. (2006). Walking through doorways causes forgetting:
916 situation models and experienced space. *Memory and Cognition*, 34(5):1150–1156.
- 917 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature*
918 *Reviews Neuroscience*, 13:713–726.
- 919 Rovelli, C. (2022). Memory and entropy. *Entropy*, 24(8):1022.
- 920 Schacter, D. L. (2012). Constructive memory: past and future. *Dialogues in Clinical Neurosciences*,
921 1:7–18.
- 922 Schacter, D. L., Norman, K. A., and Koutstaal, W. (1998). The cognitive neuroscience of constructive
923 memory. *Annual Review of Psychology*, 49:289–318.
- 924 Schacter, D. L. and Tulving, E. (1994). *Memory systems 1994*. MIT Press, Cambridge, MA.
- 925 Shankar, K. H. and Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural*
926 *Computation*, 24:134–193.

- 927 Tamir, D. I. and Mitchell, J. P. (2013). Anchoring and adjustment during social inferences. *Journal*
928 *of Experimental Psychology: General*, 142(1):151–162.
- 929 Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive*
930 *Sciences*, 22(3):201–212.
- 931 Wearden, J. (2016). *The psychology of time perception*. Springer.
- 932 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit mem-
933 ories to other brains: constructing shared neural representations via communication. *Cerebral*
934 *Cortex*, 27(10):4988–5000.
- 935 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and
936 memory. *Psychological Bulletin*, 123(2):162–185.

937 Acknowledgements

938 We thank Luke Chang, Yi Fang, Paxton Fitzpatrick, Caroline Lee, Meghan Meyer, Lucy Owen, and
939 Kirsten Ziman for feedback and scientific discussions. Our work was supported in part by NSF
940 CAREER Award Number 2145172 to J.R.M. The content is solely the responsibility of the authors
941 and does not necessarily represent the official views of our supporting organizations. The funders
942 had no role in study design, data collection and analysis, decision to publish, or preparation of the
943 manuscript.

944 Author contributions

945 Conceptualization: X.X. and J.R.M.; Methodology: X.X. and J.R.M.; Software: X.X.; Analysis: X.X.,
946 Z.Z., X.Z., and J.R.M.; Writing, Reviewing, and Editing: X.X., Z.Z., X.Z., and J.R.M.; Supervision:
947 J.R.M.

⁹⁴⁸ **Competing interests**

⁹⁴⁹ The authors declare no competing interests.