

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

³ Xinming Xu¹, Ziyuan 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 18, 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 than references to the
²¹ future. Our work reveals a temporal asymmetry in how observations of other people's behaviors can inform
²² 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

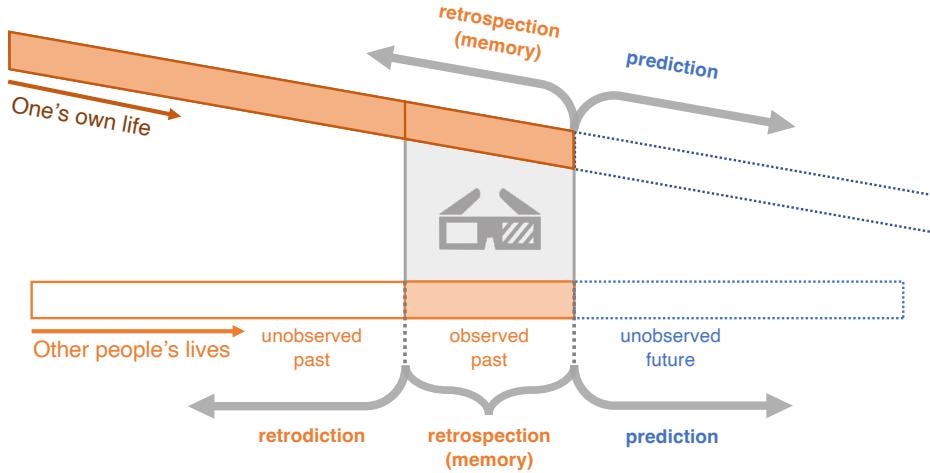


Figure 1: Retrodiction, retrospection, and prediction. In one’s own life, one may draw on memory to retrospect (i.e., review or re-evaluate) the past or predict the future. This process is time-asymmetric, since our own past is (typically) observed whereas our future is not. When we make inferences about *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 next likely to encounter step $n + 1$).

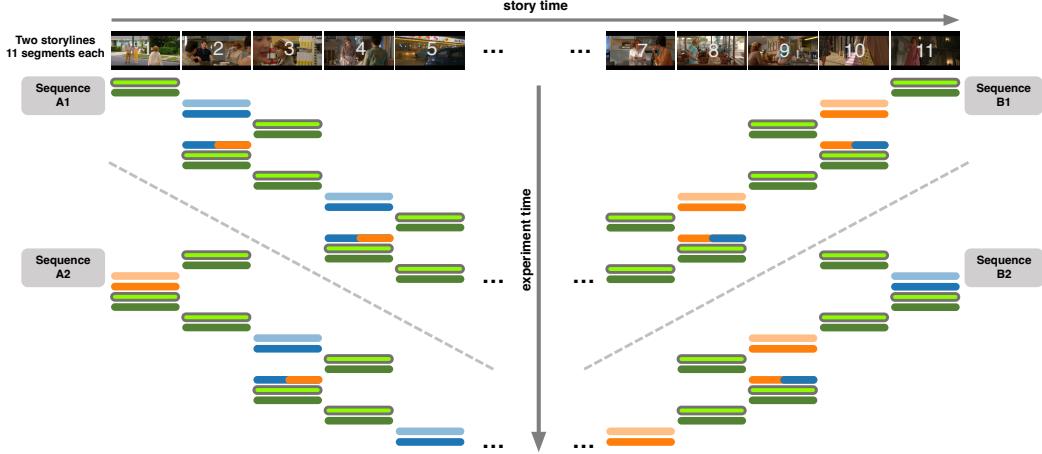
100 We designed a naturalistic paradigm for exposing participants to scenarios where the past
101 and future were equally unobserved. We asked our participants to watch a series of movie
102 segments drawn from a character-driven dramatic television show. Across the conditions and
103 trials in the experiment, 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

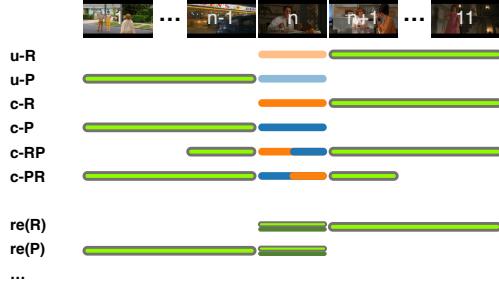


Figure 2: Task overview. Participants in our main experiment watched segments of two storylines from the television series *Why Women Kill*. They made free-form text responses to either retrodict what had happened in the previous segment, predict what would happen in the next segment, or recall what 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 generated 1.08 responses per prompt; therefore we chose to
157 consider only participants’ first (“most probable” or “most important”) responses to each prompt.
158 We also discarded a small number ($n = 20$) of character-cued responses that did not contain
159 references to all cued characters, along with one additional response due to the participant’s
160 misunderstanding of the task instructions during that trial. We carried out our analyses on the
161 remaining 2084 retrodiction, prediction, and recall responses.

162 We used two general approaches to assess the quality of participants’ responses (see *Methods*,
163 Figs. 3A). One approach entailed manually annotating events in the video and counting the number
164 of matched events in participants’ responses. We identified a total of 117 unique events reflected
165 across the 22 video segments (range: 3–9 per segment; see *Methods*, Table S1). We assigned
166 one “point” to each of these video events. We also identified 23 additional events in participants’
167 responses that were either summaries of several events or that were partial matches to the manually
168 identified video events. We assigned 0.5 point to each of these additional events. This point
169 system enabled us to compute the numbers and proportions (*hit rates*) of correctly retrodicted,
170 predicted, and recalled events contained in each response. Our second approach entailed using
171 a natural language processing model (Cer et al., 2018) to embed annotations and responses in
172 a 512-dimensional feature space. This approach was designed to capture conceptual overlap
173 between responses that were not necessarily tied to specific events. To quantify this conceptual

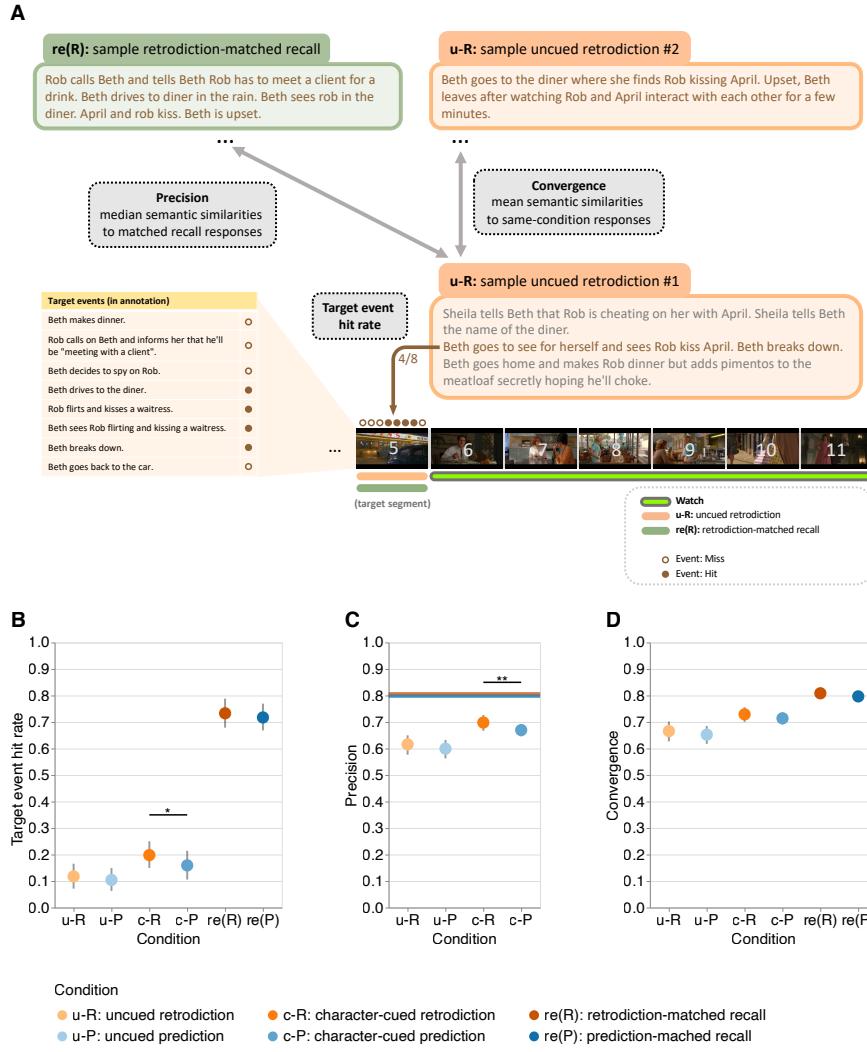


Figure 3: Retrodiction, prediction, and recall performance by experimental condition. **A. Methods schematic.** For each retrodiction, prediction, and recall response, we calculated the hit rate for events in the target segment, the response precision (see *Methods*), and the response convergence across 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.

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 (retrodiction 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 (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). These findings also held for our replication experiment (Fig. S5; hit rates of character-cued vs. uncued responses:

202 OR: XXX, Z = XXX, p = XXX, 95% confidence interval (CI): XXX to XXX; precisions of character-
203 cued vs. uncued responses: b = XXX, t (XXX) = XXX, p = XXX, CI: XXX to XXX; convergence of
204 character-cued vs. uncued responses: b = XXX, t (XXX) = XXX, p = XXX, CI: XXX to XXX).

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

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

237 The above analyses were focused solely on the target segment (i.e., retrodiction of segment n
238 after watching segments $(n+1)\dots 11$, or prediction of segment n after watching segments $1\dots(n-1)$).
239 We wondered whether participants' responses might also contain longer-range information about
240 preceding or proceeding events. In order to carry out this analysis properly, we reasoned that
241 participants might reference past or future events that were *implied* to have occurred offscreen,
242 but not explicitly shown onscreen. For example, a character in location A during one scene might
243 appear in location B during the immediately following scene. Although it wasn't shown onscreen,
244 we can infer that the character traveled between locations A and B sometime between the time
245 intervals separating the scenes (Bordwell, 2008). In all, we manually identified a set of 74 *implicit*
246 offscreen events that were implied to have occurred given what was (explicitly) depicted onscreen
247 (Fig. 4A), plus one additional partial event and one additional summary event. We defined the
248 just-watched segment as having a *lag* of 0. We assigned the target segment of a participant's
249 retrodiction or prediction (i.e., the immediately preceding or proceeding segment) a lag of -1 or
250 +1, respectively. The segment following the next was assigned a lag of 2, and so on. We tagged
251 offscreen events using half steps. For example, an offscreen event that occurred after the prior
252 segment but before the just-watched segment would be assigned a lag of -0.5.

253 Because there is no "ground truth" number of offscreen events, we could not compute the hit
254 rates for offscreen events. Instead, we counted up the absolute *number* of retrodicted or predicted
255 events as a function of lag. In other words, given that the participant had just watched segment i ,
256 we asked how many events from segment $i + lag$ they retrodicted or predicted, on average, given
257 that they were aiming to retrodict or predict events at lags of ± 1 . We also counted the numbers of

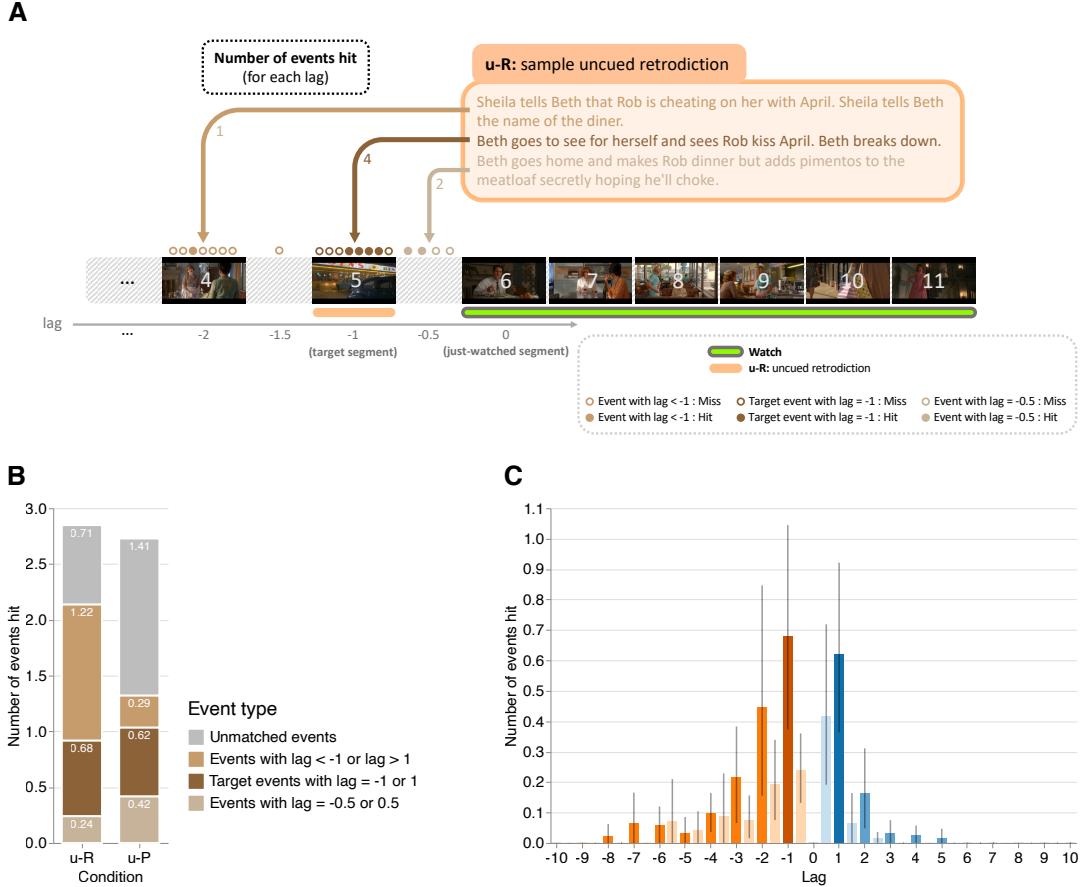


Figure 4: Retrodictions and predictions of temporally near and distant events. A. Illustration of annotation approach. For each uncued retrodiction and prediction response in our main experiment, we calculated the number of (retrodicted or predicted) events as a function of temporal distance from the 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 *unmatched* events in participants' responses that did not correspond to any events in the relevant
259 segments of the narrative. We focused specifically on *uncued* retrodictions and predictions, which
260 we hypothesized would provide the cleanest characterizations of participants' initial estimates of
261 the unobserved past and future (i.e., without potential biases introduced by additional character
262 information, as in the character-cued responses). For participants in our main experiment, the
263 numbers of uncued retrodicted and predicted target ($lag = \pm 1$) events were not reliably different
264 ($OR = 0.92, Z = -0.15, p = 0.88, CI: 0.30$ to 2.84). In other words, uncued retrodictions and
265 predictions over short timescales did not exhibit reliable asymmetries. This "null result" also
266 held in our replication study ($OR = XXX, Z = XXX, p = XXX, CI: XXX$ to XXX). However, when
267 retrodicting, participants in both experiments mentioned events from the distant past ($lag < -1$)
268 more often than participants predicted events from the distant future ($lag > 1$; main experiment:
269 $OR = 9.10, Z = 3.80, p < 0.001, CI: 2.92$ to 28.39 ; Fig. 4B, C; replication experiment: $OR = XXX,$
270 $Z = XXX, p = XXX, CI: XXX$ to XXX ; Fig. S6; for results from the character-cued conditions,
271 see Fig. S2). Despite this asymmetry in the accuracies of participants' long-range retrodictions
272 versus predictions, there were no reliable differences in the *numbers* of uncued retrodicted versus
273 predicted events (across all lags; main experiment: $OR = 1.05, Z = 0.75, p = 0.45, CI: 0.93$ to 1.18 ;
274 replication experiment: $OR = XXX, Z = XXX, p = XXX$). Nor did we find any reliable differences in
275 the numbers of offscreen events immediately before or after the just-watched segment ($lag = \pm 0.5$;
276 main experiment: $OR = 0.75, Z = -0.36, p = 0.72, CI: 0.15$ to 3.59 ; replication experiment: OR
277 $= XXX, Z = XXX, p = XXX, CI: XXX$ to XXX). The apparent discrepancy between participants'
278 asymmetric accuracy but symmetric event counts was due to participants' tendencies to reference
279 "unmatched" events (i.e., events that did not correspond to any explicit or implicit event in the
280 story) more in their predictions than retrodictions (main experiment: $OR = 0.36, Z = -4.53,$
281 $p < 0.001, CI: 0.23$ to 0.56 ; replication experiment: $OR = XXX, Z = XXX, p = XXX, CI: XXX$ to
282 XXX). We confirmed that the retrodiction advantage held when controlling for absolute lag (main
283 experiment: $OR = 34.31, Z = 3.28, p = 0.001, CI: 4.16$ to 283.20 ; replication experiment: $OR = XXX,$
284 $Z = XXX, p = XXX, CI: XXX$ to XXX), for onscreen events alone (main experiment: $OR = 47.54,$
285 $Z = 3.74, p < 0.001, CI: 6.27$ to 360.60 ; replication experiment: $OR = XXX, Z = XXX, p = XXX, CI:$

286 XXX to XXX), and marginally for offscreen events alone (main experiment: OR = 24.76, Z = 1.71,
287 $p = 0.09$, CI: 0.63 to 975.27; replication experiment: OR = XXX, Z = XXX, $p = XXX$, CI: XXX
288 to XXX). Taken together, these analyses show that (in generating uncued responses) participants
289 tend to reach “further” into the unobserved past, and with greater accuracy, than the unobserved
290 future.

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

306 Controlling for temporal distance (lag), past and future events that story characters referenced
307 in their conversations were associated with higher hit rates than unreferenced events in our main
308 experiment (uncued retrodiction: OR = 12.70, Z = 10.94, $p < 0.001$, CI: 8.06 to 20.03; uncued
309 prediction: OR = 8.29, Z = 6.83, $p < 0.001$, CI: 4.52 to 15.20; Fig. 5E). This indicates that partici-
310 pants’ responses are at least partially influenced by the characters’ conversations. To estimate the
311 contributions of characters’ references on hit rates, we computed the difference in hit rates between
312 all events (which comprised both referenced and unreferenced events) and unreferenced events,
313 as a function of lag. These differences exhibited a temporal asymmetry in favor of retrodiction

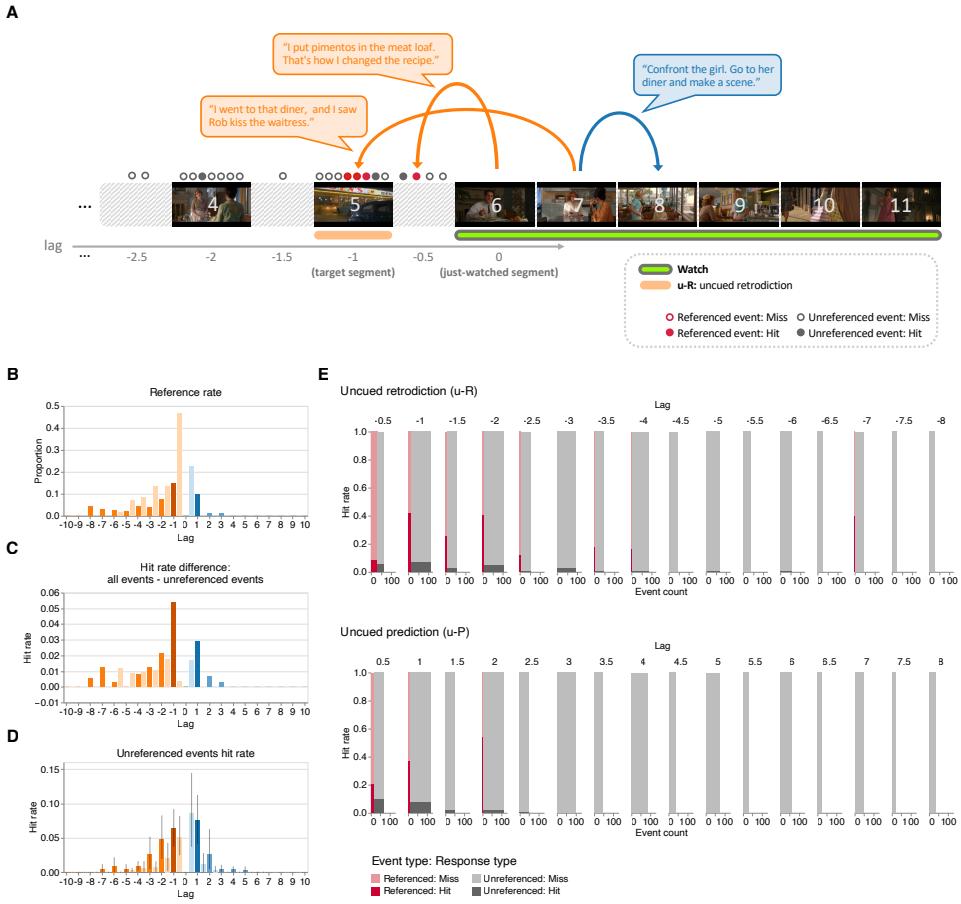


Figure 5: Characters' references drive participants' retrodiction and prediction performance. A. Illustration of annotation approach. We manually annotated references to events in past or future segments in characters' spoken conversations. We matched each such reference with its corresponding storyline 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 Fig. S7.

(Figs. 5C). This indicates that the asymmetries in participants' retrodictions versus predictions are also at least partially influenced by the characters' conversations. However, these temporal asymmetries in participants' retrodictions and predictions persisted even for events that characters never referenced in their conversations (hit rates of uncued retrodicted versus predicted unreferenced events: OR = 2.00, Z = 2.40, p = 0.02, CI: 1.14 to 3.51; Fig. 5D). When we further separated the unreferenced events into onscreen events and offscreen events, we found that these asymmetries held only for the onscreen events (onscreen: OR = 2.65, Z = 2.59, p = 0.01, CI: 1.27 to 5.54; offscreen: OR = 1.50, Z = 0.91, p = 0.36, CI: 0.63 to 3.62). We found similar patterns in our replication experiment (Fig. S7; hit rates of uncued retrodictions for referenced events: OR = XXX, Z = XXX, p = XXX, CI: XXX to XXX; uncued predictions for referenced events: OR = XXX, Z = XXX, p = XXX, CI: XXX to XXX; hit rates of uncued retrodcitions for *unreferenced* events: OR = XXX, Z = XXX, p = XXX, CI: XXX to XXX; for predicted events: OR = XXX, Z = XXX, p = XXX, CI: XXX to XXX). Taken together, these analyses suggest that asymmetries in the number of references characters make to past and future events partially (but not entirely) explain why participants tend to retrodict the past further and more accurately than they predict the future.

If characters' direct references cannot fully account for the temporal asymmetry in retrodicting the unobserved past versus predicting the unobserved future, what other factors might explain this phenomenon? The results above indicate that characters' references to specific unobserved events in the past or future boost participants' estimates of these events. But might characters' references have other effects on participants' responses *beyond* the referenced events? For example, real-world experiences and events in realistic narratives are often characterized by temporal autocorrelations (i.e., what is "happening now" will likely relate to what happens "a moment from now," and so on). Real-world experiences and realistic narratives are also often structured into "schemas" whereby experiences unfold according to a predictable pattern or formula that characterizes a particular situation, such as going to a restaurant or catching a flight at the airport (Baldassano et al., 2018). If there are associations or temporal dependencies between temporally nearby events in the television show participants watched, participants might be able to pick up on these patterns in forming their responses. This would be reflected in an inference "boost" for events that were

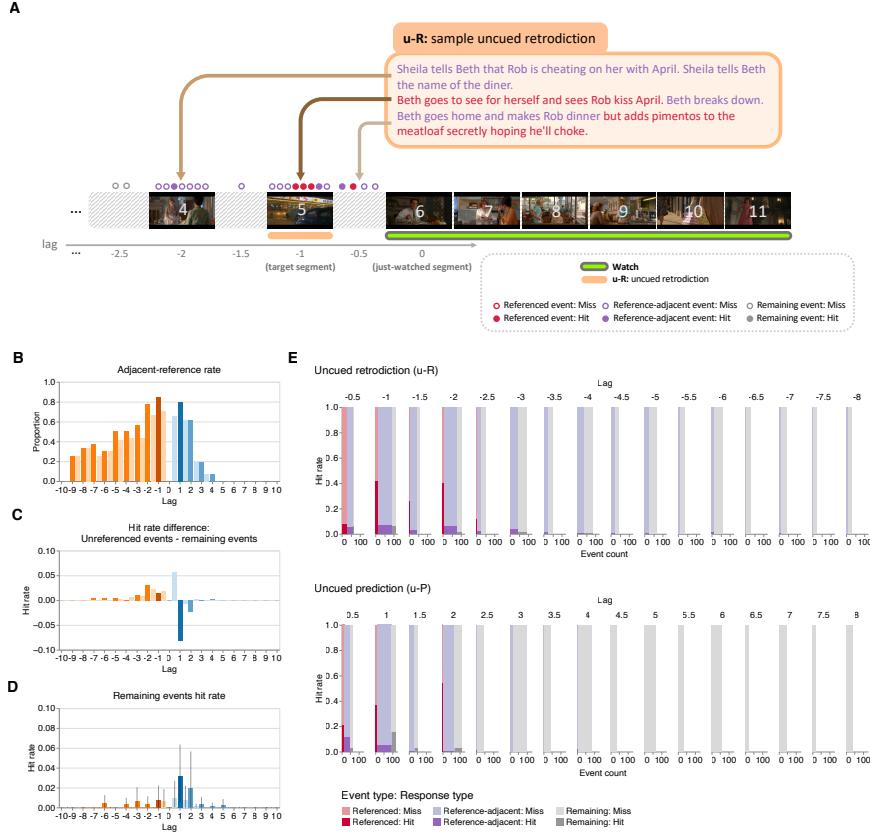


Figure 6: Reference-adjacent events are associated with higher hit rates (main experiment). **A. Illustration of annotation approach.** We extended the annotation procedure depicted in Figure 5A to also label unreferenced events that were either temporally adjacent to (i.e., immediately preceding or 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 nearby in time to events that characters referred to in their conversations, in addition to the referenced
343 events themselves (Fig. 6A).

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

367 The preceding analyses show that when characters reference past or future events, those refer-
368 enced events, and other events that are temporally adjacent to the referenced events, are more likely
369 to be retrodicted and predicted. In other words, referring to a past or future event in conversation



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. 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.

leads to a “boost” in that event’s hit rate. We wondered whether this boost was bi-directional. In particular: when a character refers (during a *referring event*) to another event (i.e., the *referenced event*), does this boost only the referenced event’s hit rate, or does the referring event also receive a boost? We labeled each event as a “referring event,” a “referenced event,” or a “other event” (i.e., not referring or referenced; Fig. 7A, B). We limited our analysis to references to onscreen (explicit) events. Consistent with our analysis of the proportions of referenced events (Fig. 5B), the proportions of *referring* events exhibited a *forward* temporal asymmetry (Fig. 7C). Controlling for absolute lag, we found that referring events were associated with lower hit rates than referenced events in our main experiment (uncued retrodiction: OR = 0.03, Z = -4.81, $p < 0.001$, CI: 0.01 to 0.11; uncued prediction: OR = 0.04, Z = -5.84, $p < 0.001$, CI: 0.01 to 0.12; Fig. 7D) and had no reliable differences in hit rates compared with other events (uncued retrodiction: OR = 0.37, Z = -1.46, $p = 0.15$, 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 observed this phenomenon in our replication experiment (referenced events, uncued retrodiction: OR = XXX, Z = XXX, $p = XXX$, CI: XXX to XXX; referenced events, uncued prediction: OR = XXX, Z = XXX, $p = XXX$, CI: XXX to XXX; other events, uncued retrodiction: OR = XXX, Z = XXX, $p = XXX$, CI: XXX to XXX; other events, uncued prediction: OR = XXX, Z = XXX, $p = XXX$, CI: XXX to XXX; Fig. S9). Taken together, this indicates that only referenced events received a hit rate boost (relative to other events), suggesting that the retrodictive and predictive benefits of references are directed (i.e., asymmetric).

The above analyses show that characters in the television shows we used as stimuli in our main experiment and replication experiment refer more often to the past than to the future. This appears to bias participants’ inferences about the past and future. But how universal is this pattern? For example, were the television shows we happened to select for our experiment representative of television shows more generally? Or perhaps media created for entertainment purposes 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 a total of over 440 million conversations from over 17 million documents. The data comprised transcripts from television shows and

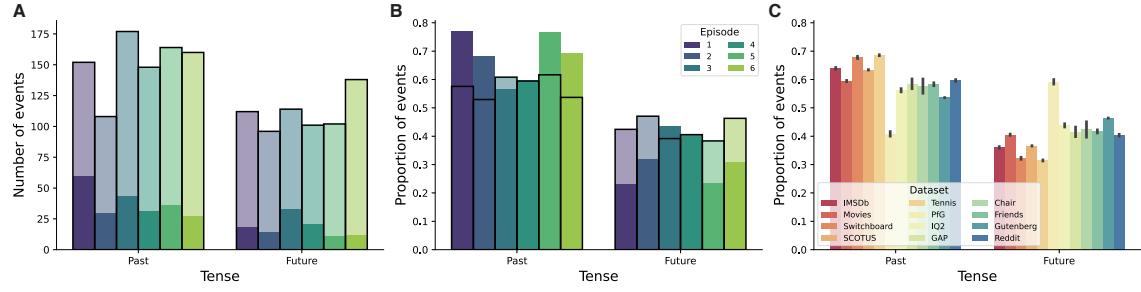


Figure 8: Meta analysis. We used natural language processing to automatically identify references to past or future events across a variety of sources. **A. Numbers of past and future events in *The Chair*, Season 1, Episodes 1–6.** The bar heights indicate the raw numbers of manually identified (lighter 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 S2. Error bars denote bootstrap-estimated 95% confidence intervals.

398 popular films, novels, and spoken and written utterances from natural conversations. A summary
 399 of the data we analyzed may be found in Table S2. As summarized in Figure 8, we used natural
 400 language processing to identify references to past or future events in each conversation (also see
 401 *Meta analysis of conversation data*).

402 To validate our basic approach, we compared the numbers (Fig. 8A) and proportions (Fig. 8B) of
 403 automatically and manually identified references to past and future events, across six episodes of
 404 the television show *The Chair*. (The first episode was used as the stimulus in our replication study.)
 405 In general, our automated tagging procedure tended to overcount the numbers of references. From
 406 manually inspecting hundreds of example tags, we believe this discrepancy follows from which
 407 criteria were used to generate the tags. The manually generated tags sought to identify references
 408 to specific events that occurred or were implied to occur in other parts of the narrative. In contrast,
 409 as a heuristic, we designed the automatic tagging procedure to identify uses of the past or future
 410 *tense* as a proxy for references to past or future *events*. We noticed that, a single conversation often
 411 contains multiple references to a given (past or future) event. Whereas the manually generated
 412 tags counted these as “single” references, our automated tagging procedure had no means of

413 differentiating between multiple references to the same event versus references to different events.
414 Nevertheless, this discrepancy did not appear to bias the balance of the overall *proportions* of past
415 or future references.

416 In all, across all of the datasets we examined in our meta analysis, we identified a total of
417 36,008,500 references to past or future events. A total of 19,464,741 (54.06%) of these were ref-
418 erences to past events, and the remaining 16,543,759 (45.94%) were references to future events.
419 We also computed the average proportions of references to past and future events across doc-
420 uments within each individual dataset. Across the 12 datasets we examined (Fig. 8, Tab. S2),
421 there were significantly more references to the past than the to the future (mean \pm standard de-
422 viation: $58.99\% \pm 7.28\%$; $t(11) = 4.28, p = 0.0013$). This bias towards the past also held for each
423 dataset individually ($ts \geq 5.14, ps < 0.01$) except for one dataset, "Persuasion for Good," which
424 comprised natural conversations between pairs of Amazon Mechanical Turk workers wherein
425 one participant tried to convince the other participant to donate to a charity in the future. In
426 that dataset, references to the future were significantly more common than references to the past
427 ($t(11438) = -22.65, p < 0.001$). This latter example provided a nice sanity check for verifying that
428 our general approach was not itself biased in favor of the past, e.g., even in conversations that were
429 actually biased towards the future. Taken together, the results from our meta analysis indicate
430 that people tend to refer to the past more than they refer to the future, across a wide variety of
431 situations (including in both fictional and real conversations). Although (as in the Persuasion for
432 Good dataset) there may be specific exceptions to this bias, it seems that a bias in favor of the past
433 is a common element of many (and perhaps even *most*) human conversations.

434 Discussion

435 We asked participants to watch sequences of movie segments from a character-driven television
436 drama and then either retrodict what had happened prior to a just-watched segment, predict what
437 would happen next, or recall what they had just watched. We found that participants tended
438 to more accurately and more readily retrodict the unobserved past than predict the unobserved

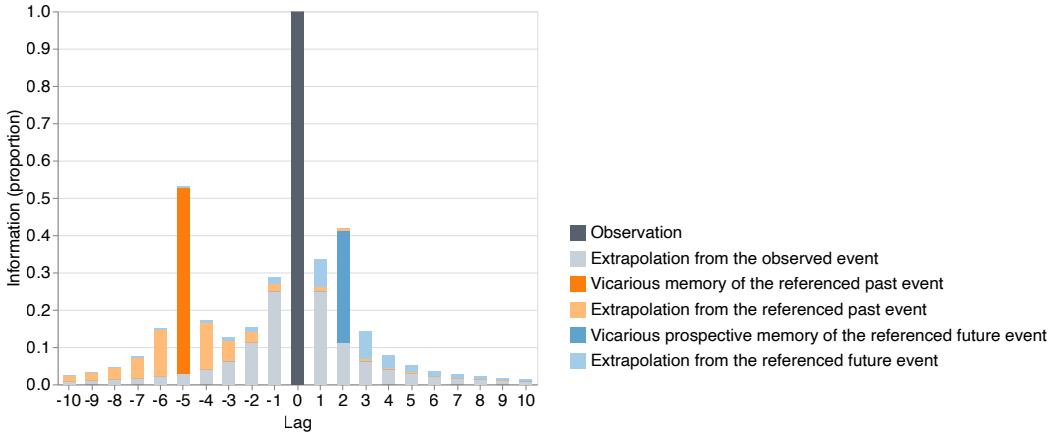


Figure 9: How much information about the past and future can be inferred by observing the present? By definition, let us say that the present moment (lag 0) contains all information about itself (dark gray). Given learned statistical regularities, one might extrapolate from the present moment into the 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.

439 future. We traced this temporal asymmetry to (a) characters' tendencies to refer to past events
 440 more than future events in their ongoing conversations, and (b) associations between temporally
 441 proximal events (Fig. 9). Essentially, associations between temporally proximal events serve to
 442 enhance asymmetries in inferences driven by conversational references (light orange and blue bars
 443 in Fig. 9). Our findings show that other peoples' psychological arrows of time can affect external
 444 observers' inferences about the unobserved past and future.

445 When people communicate through language or other observable behaviors, they can transmit
 446 their knowledge and memories to others (Hirst and Echterhoff, 2012; Mahr and Csibra, 2018;
 447 Dessalles, 2007; Zadbood et al., 2017). A consequence of this sharing across people is that biases or
 448 limitations in one person's knowledge and memories may also be transmitted to external observers.
 449 Although people *can* communicate their intentions and future plans (i.e., information about their
 450 future), because people know *more* about their pasts than their futures, the knowledge transmitted

451 to observers is inherently biased in favor of the past (Fig. 9; Demiray et al., 2018). Since observers
452 leverage communicated knowledge to reconstruct the unobserved past and future, this explains
453 why observers' inferences about observed people's lives also favor the past.

454 People's knowledge asymmetries are not always directly observable. For example, in a con-
455 versation where someone talks exclusively about their future plans, a passive observer might gain
456 more insight into the speaker's unobserved future than their unobserved past. However, because
457 the speaker is also guided by their own psychological arrow of time, the "upper limit" of knowledge
458 about their past is still higher than that of their future. Therefore, after accounting for knowledge
459 that *could* be revealed through active participation in the conversation, the seemingly future-biased
460 conversation masks an underlying knowledge asymmetry in favor of the past. This hypothesized
461 "unmasking" effect of interaction implies that the influence of other people's psychological arrows
462 of time should be more robust when the receiver is an active participant in the conversation. Other
463 social dimensions, such as trust, motivation or level of engagement, personal goals, and beliefs,
464 might serve to modulate the effective "gain" of the communication channel— i.e., how much the
465 speaker's knowledge influences the observer's knowledge.

466 In typical statistical sequences used in laboratory studies, there is no temporal asymmetry,
467 either theoretically (Cover, 1994; Bialek et al., 2001; Ellison et al., 2009), or empirically (Jones and
468 Pashler, 2007). What makes narratives and real-world event sequences time-asymmetric? Of
469 course there are many superficial differences between simple laboratory-manufactured sequences
470 and real-world experiences. As one example, real-world experiences often involve other people
471 who have their own memories and goals. At a deeper level, however, are our subjective experi-
472 ences essentially more complicated versions of laboratory-manufactured sequences? Or are there
473 fundamental differences? One possibility is that real-life event sequences are not stationary (i.e.,
474 not in equilibrium, Cover, 1994). For example, real-life events might start from a special initial
475 condition (Albert, 2000; Feynman, 1965; Cover, 1994) and proceed through a series of transitions
476 from more-ordered to less-ordered states, thus exhibiting an arrow time. When we retrodict, it is
477 possible that we only consider possible past events that are compatible with the highly-ordered
478 special initial state (Carroll, 2010, 2016). For example, when we see a broken egg we might infer

479 that the egg had been intact at some point in the past. But it would be difficult to guess at what
480 states or forms the broken egg might take in the future (Carroll, 2010, 2016). In other words, the
481 procession from order to disorder might result in better retrodiction performance compared with
482 that of (implicitly less-restricted) prediction tasks. The special initial state might also explain why
483 we remember the past, but not the future. Some recent work suggests that the psychological arrow
484 of time might be explained by a related concept in the statistical physics literature, termed the
485 “thermodynamic” arrow of time (Mlodinow and Brun, 2014; Rovelli, 2022). However, the relation
486 between the thermodynamic and psychological arrows of time is still under debate (Gołosz, 2021;
487 Hemmo and Shenker, 2019).

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

500 Methods

501 Participants

502 A total of 36 participants (25 female, mean age 21.47 years, range 19–50 years) were recruited from
503 the Dartmouth College community. All participants had self-reported normal or corrected-to-

504 normal vision, hearing, and memory, and had not watched any episodes of *Why Women Kill* before
505 the experiment. Participants gave written consent to enroll in the study under a protocol approved
506 by the Committee for the Protection of Human Subjects at Dartmouth College. Participants received
507 course credit or monetary compensation for their time. Two participants completed only the first
508 half of the study and one participant's data from the second half of their testing session was lost
509 due to a technical error. All available data were used in the analyses.

510 **Stimuli**

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

521 **Task design and procedure**

522 Our experimental paradigm was divided across two testing sessions. In each session, participants
523 performed a sequence of tasks on segments from one storyline (Fig. 2). For each storyline, there
524 were four different task sequences: two forward chronological order sequences and two backward
525 chronological order sequences. Participants completed one task sequence in forward chronological
526 order for one storyline, and one in backward chronological order for the other storyline. The order
527 of the two sessions (forward chronological order sequence first or backward chronological order
528 sequence first), and the pairing of task sequences with storylines, were counterbalanced across

529 participants.

530 Tasks in each sequence alternated between watching, recall, and retrodiction or prediction,
531 with the specific order of tasks differing across the four sequences. For example, in sequence A1,
532 participants first watched segment 1, followed by an immediate recall of segment 1. Then they
533 predicted what would happen in segment 2 (first uncued and then character-cued). Participants
534 then watched segment 3 and recalled segment 3. After that, participants guessed what happened in
535 segment 2 again, which we termed “updated prediction”. Then they watched segment 2, recalled
536 segment 2, and so on as depicted in Figure 2. This procedure was repeated to cover all possible
537 segments. We also note several edge cases at the start and end of the narrative sequences. Since
538 no segments precede the first segment, participants could never make “prediction” responses with
539 the first segment as their target. For analogous reasons, participants never made “retrodiction”
540 responses with the last segment as their target. Another edge case occurred in task sequences
541 B2 and A2 (Fig. 2). In the A1 and A2 sequences, participants experience the narrative in the
542 original (forward) order, predicting one segment ahead along the way. In the B1 and B2 sequences,
543 participants experience the narrative in the reverse order, retrodicting one segment ahead along
544 the way. However, because A2 and B2 are offset from A1 and B1 by one segment, the initial A2
545 responses are *retrodiction*, and the initial B2 responses are *predictions* (i.e., they conflict with the
546 temporal directions of the remaining responses in those conditions). We therefore excluded from
547 our analysis those initial retrodiction responses from the A2 condition, and the initial prediction
548 responses from the B2 condition.

549 Before watching each segment, participants were given the following task instructions. After
550 watching the video, participants were instructed to type their responses (retrodiction, prediction,
551 or recall) in 1–4 sentences. Participants were also asked to specify the characters’ names in their
552 responses, i.e., avoiding use of characters’ pronouns. For the recall task, the names of the characters
553 in the recall segment were displayed, and participants were asked to summarize the major plot
554 points in the present tense. For the retrodiction and prediction tasks, participants were instructed
555 to retrodict or predict the major plot points of the segment (also in the present tense), as though
556 they had watched the segment and were writing a plot synopsis. They were also instructed to

557 avoid speculation words (e.g., “I *think* Beth will...”). For the uncued retrodiction and prediction
558 tasks, participants made retrodictions or predictions without any cues provided, so they had to
559 guess which of the characters would be present in the segment. For character-cued retrodictions
560 and predictions, the characters in the target segment were revealed on the screen, alongside
561 participants’ previous responses. Participants were instructed to include or incorporate those
562 characters into their character-cued responses, if their previous responses did not contain all the
563 characters provided. They were also told that the characters were not necessarily listed in their
564 order of appearance in the segment, and that only the main characters would be given. Also, the
565 characters given did not necessarily interact with each other in that segment, and they could appear
566 in successive events in that segment. If participants’ previous responses included all the characters
567 given, then they could directly proceed to the next task without updating their responses. For
568 all of the prediction and retrodiction tasks, participants were instructed to provide at least one
569 response, but they were given the opportunity enter up to three responses if they felt that multiple
570 possibilities were more or less equally likely. Each response (including recall) was followed by a
571 confidence rating on a 1–5 point scale. However, these confidence data were not analyzed in the
572 present study.

573 Before their first testing session, participants were given a practice session, where they watched
574 the first segment of storyline 3 followed by a recall trial, an uncued prediction trial, and a character-
575 cued prediction trial. Participants’ responses were checked by the experimenter to ensure compli-
576 ance with the instructions. To provide participants with sufficient background information about
577 the storyline (especially for the backward chronological sequences), at the beginning of each ses-
578 sion, participants were shown the time, location, and the main characters (with pictures) of the
579 storyline. The first session was approximately 1.5 h long and the second session was approximately
580 1 h long. We allowed participants, at their own discretion and convenience, to sign up for two
581 consecutive testing time-slots (i.e., with their testing sessions occurring in immediate succession),
582 or for testing sessions on two different days. The mean inter-session interval was 0.73 days (range:
583 0–4 days). The experiment was conducted in a sound- and light-attenuated testing room. Videos
584 were displayed using a 27-inch iMac desktop computer (resolution: 5120 × 2880) and sound was

585 presented using the iMac's built-in speakers. The experiment was implemented using jsPsych (de
586 Leeuw, 2015) and JATOS (Lange et al., 2015).

587 **Video annotation**

588 Events in the first 11 segments of the two storylines were identified by the first author (X.X.),
589 corresponding to major plot points (total: 117; mean: 5.32 per segment; range 3–9). Additionally,
590 74 offscreen events were identified. Of these 74 offscreen events, 43 events were identified from
591 references in conversations during onscreen events. Another 16 events were identified based on
592 characters' implied movements and travels. For example, if in segment 1 character A was in place
593 A and in segment 2 she was in place B, then the transit from place A to B for character A would be
594 identified as an offscreen event. The remaining 15 offscreen events were identified based on logical
595 inferences. For example, if a photograph was shown in an onscreen event (but not the act of the
596 photograph being taken), then the action that someone took the photograph would be identified
597 as an offscreen event. Offscreen events always occurred between two contiguous segments, or
598 before the first segment. The purpose of identifying offscreen events was to match participants'
599 responses to video events; thus our identification of these offscreen events was not intended to be
600 exhaustive.

601 **Response analyses**

602 Participants' retrodiction, prediction, and recall responses were minimally processed to correct
603 obvious typos (e.g., in characters' names) and remove speculation descriptions (e.g., "I predict
604 that..."). All responses were manually coded and matched to events from the video annotations.
605 Retrodiction and prediction responses were coded by two coders (X.X. and Z.Z.). Recall responses
606 were coded by one coder (X.X.). While most responses were clearly identifiable as either matching
607 specific storyline events or as not matching any storyline events, several ambiguous cases arose.
608 First, some responses combined or summarized over several (distinct) storyline events. Second,
609 some responses lacked any specific detail (e.g., "character A and B talk" without describing the

610 specific topic(s) of conversation or providing other relevant details). Based on participants' re-
611 sponses, in addition to the original 117 onscreen events and 74 offscreen events, we added 25 new
612 events (23 onscreen, 2 offscreen) that either summarized across several events or partially matched
613 the annotated events. Whereas the original events were each assigned a value of one point, we
614 assigned these additional events a half point. This point system enabled us to directly match events
615 in participants' responses to the annotated events. In our analyses of retrodictions, predictions,
616 and recalls, we added up the number of points earned for each response to estimate participants'
617 event hit rates.

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

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

630 To estimate the semantic similarities between pairs of responses, we first transformed each
631 response into a 512-dimensional vector (embedding) using the Universal Sentence Encoder (Trans-
632 former USE, Cer et al., 2018). We defined *similarity* as the cosine of the angle formed by the
633 responses' vectors. Following Heusser et al. (2021), we defined the *precision* of participants' re-
634 sponds as the median similarity between that response's vector and the embedding vectors for
635 all other participants' recalls of the target segment. We defined the *convergence* of a given response
636 as the mean similarity between that response's vector and all other participants' responses to the
637 corresponding segment, in the same condition. To compute these median or mean similarities we

638 first applied the Fisher z -transformation to the similarity values, then took the median or mean
639 of the z -transformed similarities, and finally applied the inverse z -transformation to obtain the
640 precision or convergence score.

641 To test the validity and reliability of the USE embeddings, we performed a classification analysis
642 of recall responses using a leave-one-out approach. For each recall response, we calculated its
643 semantic similarity with all other recall responses for the same storyline. We took the segment
644 with the highest median semantic similarity (to the recall response) as the “predicted” segment.
645 Across all responses, the predicted segments matched the true recalled segments’ labels 98.6% of
646 the time (1088 out of 1103 predictions; chance level: 9%).

647 Reference coding

648 Two coders (X.X. and Z.Z.) identified character dialogues in the narrative that referred to past
649 events or future (onscreen or offscreen) events. Only references to events that occurred in a different
650 segment were included in this tagging procedure. For each reference, the source (referring) segment
651 and the referred event number were recorded. A total of 82 references were identified. Of these, 30
652 referred to onscreen events and 52 referred to offscreen events. For these referenced events, their
653 corresponding summary events or partial events were also labelled as referenced. In instances
654 where the coders disagreed about a given tag, disagreements were resolved through discussions
655 between the two coders. In our analyses, each storyline event was coded according to whether
656 or not it had been referenced in the segment(s) that the participant had viewed thus far in the
657 experiment.

658 In principle, a given event could receive multiple labels. For example, during event A , a
659 character might speak about another event, B , during which a reference to a third event (C) was
660 made. In this scenario, event B could be both a “referring event” ($B \rightarrow C$) and a referenced event
661 ($A \rightarrow B$). In practice, however, this scenario was quite rare, accounting for only one out of a total
662 of 30 onscreen events.

663 **Statistical analysis**

664 We used (generalized) linear mixed models to analyze the hit rates and numbers of events retrodicted,
665 predicted, and recalled, as well as the precisions and convergences of participants' responses.
666 Our models were implemented in R using the `aefex` package. We carried out comparisons or con-
667 trasts, and extracted *p*-values, using the `emmeans` package. Participants and stimuli (e.g., segment
668 identity) were modeled as crossed random effects (as specified below). Random effects were se-
669 lected as the maximal structure that allowed model convergence. All of our statistical tests were
670 two-sided.

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

```
673   cbind(thp, ttp - thp) ~ direction * level * seg_cnt * storyline +  
674   (direction * level | target) +  
675   (direction * level * seg_cnt | subject)
```

676 where `thp` was the number of points hit for the target segment, `ttp` was the total number of points
677 for the target segment (from its annotations), `direction` was either retrodiction or prediction, `level`
678 had four levels (uncued, character-cued, updated, and recall), `seg_cnt` represented the number of
679 segments in the storyline that had been watched (1–10, centered), `storyline` had two levels (1
680 or 2), and `target` had 22 levels according to the identity of the target segment. For our tests of
681 precision and convergence (Fig. 3C, D), we fit linear mixed models using the same formula. To
682 test the effect of `direction` (retrodiction or prediction) on target event hit rates, precision, and
683 convergence, we fit a (generalized) linear mixed model separately for each of the three levels
684 (uncued, character-cued, and recall).

685 For our tests comparing the numbers of hits for different types of events (Fig. 4B), we fit
686 generalized linear mixed models using the same formula, but with a Poisson link function. For
687 these models, we manually doubled the point counts to ensure that half points were mapped onto
688 integers, ensuring compatibility with the Poisson link function.

689 For our analyses of the numbers of events hit, controlling for lag (Fig. 4C), we fit a generalized

690 linear mixed model with a Poisson link function:

```
691 hp_lag ~ direction * full_stp * lag * storyline +
692 (direction | base_seg) + (1 | base_seg_pair) +
693 (direction * full_stp * lag * storyline | subject)
```

694 where `hp_lag` is the number of “points” earned (for each lag) in each trial (we manually doubled
695 the point counts to ensure that half points were mapped onto integers, for compatibility with the
696 Poisson link function), `full_stp` denoted whether the given events (of the given lag) were onscreen
697 (i.e., full step) or offscreen (i.e., half step), `lag` denotes the (centered) absolute lag, `base_seg` denotes
698 the identity of the just-watched segment (22 levels), and `base_seg_pair` denotes the pairing of the
699 just-watched segment and the segment at each lag (440 levels).

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

```
702 cbind(hp_lag, tp_lag - hp_lag) ~ direction * reference * full_stp +
703 lag + (direction | base_seg) +
704 (1 | base_seg_pair) +
705 (direction * reference * full_stp + lag | subject)
```

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

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

713 **Code and data availability**

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

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

716 **References**

- 717 Albert, D. Z. (2000). *Time and chance*. Harvard University Press, Cambridge, Mass.
- 718 Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas
719 during narrative perception. *The Journal of Neuroscience*, 38(45):9689–9699.
- 720 Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural
721 Computation*, 13(11):2409–2463.
- 722 Bordwell, D. (2008). *Poetics of cinema*, chapter Three dimensions of film narrative, pages 85–134.
723 Routledge.
- 724 Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*,
725 11(2):177–220.
- 726 Carroll, S. (2010). *From eternity to here: the quest for the ultimate theory of time*. Penguin.
- 727 Carroll, S. (2016). *The big picture: on the origins of life, meaning, and the universe itself*. Dutton.
- 728 Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,
729 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.
730 *arXiv*, 1803.11175.
- 731 Cover, T. M. (1994). Which processes satisfy the second law? In Halliwell, J. J., Pérez-Mercader,
732 J., and Zurek, W. H., editors, *Physical Origins of Time Asymmetry*, pages 98–107. Cambridge
733 University Press, Cambridge, UK.
- 734 de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web
735 browser. *Behavior Research Methods*, 47(1):1–12.
- 736 Demiray, B., Mehl, M. R., and Martin, M. (2018). Conversational time travel: evidence of a
737 retrospective bias in real life conversations. *Frontiers in Psychology*, 9:2160.

- 738 Dessalles, J.-L. (2007). Storing events to retell them. *Behavioral and Brain Sciences*, 30(3):321–322.
- 739 Ellison, C. J., Mahoney, J. R., and Crutchfield, J. P. (2009). Prediction, retrodiction, and the amount of
740 information stored in the present. *Journal of Statistical Physics*, 136(1005):doi.org/10.1007/s10955–
741 009–9808–z.
- 742 Feynman, R. (1965). *The character of physical law*. MIT Press.
- 743 Gołosz, J. (2021). Entropy and the direction of time. *Entropy*, 23(4):388.
- 744 Hawking, S. W. (1985). Arrow of time in cosmology. *Physical Review D*, 32(10):2489–2495.
- 745 Hemmo, M. and Shenker, O. (2019). The second law of thermodynamics and the psychological
746 arrow of time. *The British Journal for the Philosophy of Science*.
- 747 Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behavioral
748 and neural signatures of transforming naturalistic experiences into episodic memories. *Nature
749 Human Behavior*, 5:905–919.
- 750 Hirst, W. and Echterhoff, G. (2012). Remembering in conversations: the social sharing and reshap-
751 ing of memories. *Annual Review of Psychology*, 63(1):55–79.
- 752 Horwich, P. (1987). *Asymmetries in time: problems in the philosophy of science*. MIT Press.
- 753 Jones, J. and Pashler, H. (2007). Is the mind inherently forward looking? comparing prediction and
754 retrodiction. *Psychonomic Bulletin and Review*, 14(2):295–300.
- 755 Koster-Hale, J. and Saxe, B. (2013). Theory of mind: A neural prediction problem. *Neuron*,
756 79(5):836–848.
- 757 Lange, K., Kühn, S., and Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): an
758 easy solution for setup and management of web servers supporting online studies. *PLoS One*,
759 10(6):e0130834.
- 760 Maheu, M., Meyniel, F., and Dehaene, S. (2022). Rational arbitration between statistics and rules
761 in human sequence processing. *Nature Human Behaviour*, pages 1–17.

- 762 Mahr, J. B. and Csibra, G. (2018). Why do we remember? the communicative function of episodic
763 memory. *Behavioral and Brain Sciences*, 41:e1.
- 764 Mlodinow, L. and Brun, T. A. (2014). Relation between the psychological and thermodynamic
765 arrows of time. *Physical Review E*, 89(5):052102.
- 766 Radvansky, G. A. and Copeland, D. E. (2006). Walking through doorways causes forgetting:
767 situation models and experienced space. *Memory and Cognition*, 34(5):1150–1156.
- 768 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature
769 Reviews Neuroscience*, 13:713–726.
- 770 Rovelli, C. (2022). Memory and entropy. *Entropy*, 24(8):1022.
- 771 Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive
772 Sciences*, 22(3):201–212.
- 773 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit mem-
774 ories to other brains: constructing shared neural representations via communication. *Cerebral
775 Cortex*, 27(10):4988–5000.
- 776 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and
777 memory. *Psychological Bulletin*, 123(2):162–185.

778 **Acknowledgements**

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

⁷⁸⁵ **Author contributions**

⁷⁸⁶ Conceptualization: X.X. and J.R.M.; Methodology: X.X. and J.R.M.; Software: X.X.; Analysis: X.X.,
⁷⁸⁷ Z.Z., X.Z., and J.R.M.; Writing, Reviewing, and Editing: X.X., Z.Z., X.Z., and J.R.M.; Supervision:
⁷⁸⁸ J.R.M.

⁷⁸⁹ **Competing interests**

⁷⁹⁰ The authors declare no competing interests.