

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

3 Xinming Xu¹, Ziyang Zhu², and Jeremy R. Manning^{1, *}

4 ¹Dartmouth College, Hanover, NH, USA

5 ²Peking University, Beijing, China

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

7 January 7, 2023

8 **Abstract**

9 How much can we infer about the past and future, given our knowledge of the present?
10 Unlike temporally symmetric inferences about simple sequences, inferences about our own lives
11 are asymmetric: we are better able to infer the past than the future, since we remember our past
12 but not our future (i.e., the psychological arrow of time). What happens when both the past
13 and future are unobserved, as when we make inferences about *other* people's lives? We had
14 participants view segments of a character-driven television drama. They wrote out what would
15 happen just before or after each just-watched segment. Participants were better at inferring
16 past (versus future) events. This asymmetry was driven by participants' reliance on characters'
17 conversational references in the narrative, which tended to favor the past. Our work reveals a
18 temporal asymmetry in how observations of other people's behaviors can inform us about the
19 past and future.

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

21 Introduction

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

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

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

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

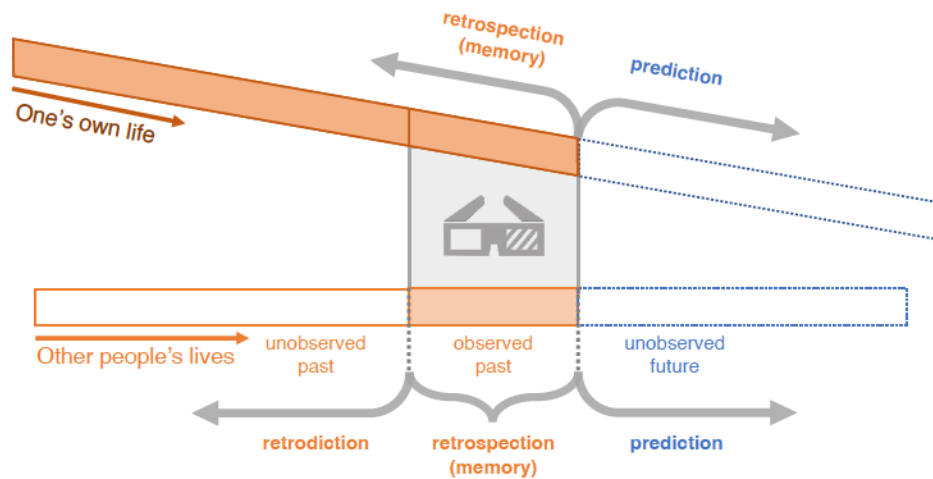


Figure 1: Retrodiction, retrospection, and prediction. In one's own life, one may draw on memory to retrospect (i.e., review or re-evaluate) the past or predict the future. This process is time-asymmetric, since our own past is (typically) observed whereas our future is not. When we make inferences about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

61 these questions.

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

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

89 about artificial statistical sequences (e.g., perhaps solely relying on statistical regularities like event
90 schemas, scripts, or situation models Radvansky and Copeland, 2006; Zwaan and Radvansky,
91 1998; Bower et al., 1979; Ranganath and Ritchey, 2012), then the accuracy of inferences about the
92 past and the future of others’ lives should be approximately equal.

93 We designed a naturalistic paradigm for exposing participants to scenarios where the past and
94 future were equally unobserved. We asked our participants to watch a series of movie segments
95 drawn from a character-driven dramatic television show. Across the conditions and trials in the
96 experiment, participants made free-form text responses to either retrodict what had happened in
97 the previous segment, predict what would happen in the next segment, or recall what happened
98 in the just-watched segment. We used manual annotations and sentence-level natural language
99 processing models to characterize participants’ responses. To foreshadow our results, we found
100 that participants were overall better at retrodicting the past than predicting the future. This
101 appeared to be driven by two main factors. First, characters more often referred to past events than
102 future (e.g., planned) events, and this influenced participants’ responses. Second, associations and
103 dependencies between temporally adjacent events enabled participants to form estimates about
104 nearby events (e.g., to a just-watched scene or a past or future event referenced in an observed
105 conversation). Taken together, our work reveals a temporal asymmetry in how observations of
106 other humans’ behaviors inform us about the past versus the future.

107 Results

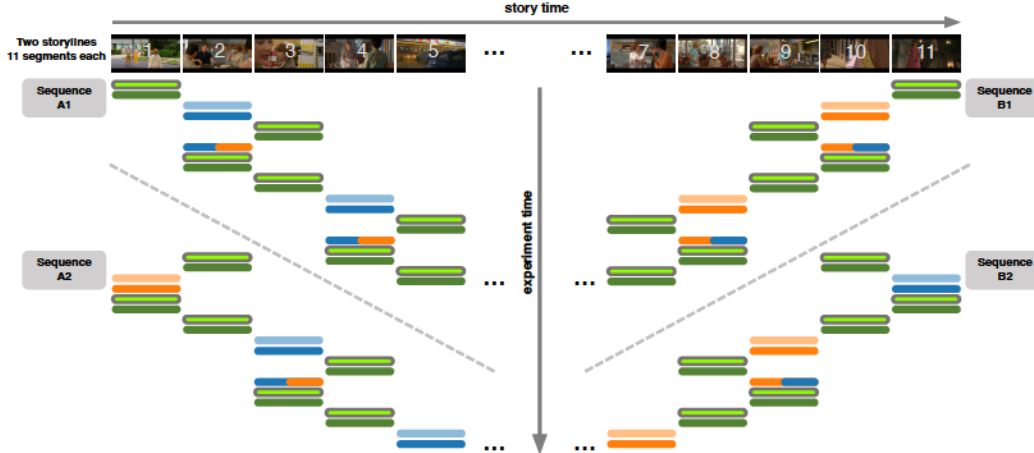
108 Participants in our study ($n = 36$) watched segments from two storylines, drawn from the CBS
109 television show *Why Women Kill*. Each storyline comprised 11 segments (mean duration: 2.05 min;
110 range: 0.97–3.87 min, Table S1). We asked participants to use free-form (typed) text responses to
111 retrodict what had happened prior to a just-watched segment, predict what would happen next,
112 or recall what they had just watched (Fig. 2, *Task design*). We referred to the to-be-retrodicted, to-
113 be-predicted, or to-be-recalled segment as the *target segment* for each response. We systematically
114 varied whether participants watched the segments in forward or reverse chronological order, and

115 how many segments they had seen prior to making a response (see *Methods*).

116 We asked participants to generate four types of responses after watching each video segment:
117 uncued responses, character-cued responses, updated responses, and recalls (Fig. 2, *Data overview*).
118 To generate *uncued* responses, we asked participants to either retrodict (uncued retrodiction; *u-R*)
119 what happened shortly before or predict (uncued prediction; *u-P*) what happened shortly after
120 the just-watched segment. To generate *character-cued* responses, we asked participants to retrodict
121 (character-cued retrodiction; *c-R*) or predict (character-cued prediction; *c-P*) what came before or
122 after the just-watched segment, but we provided additional information to the participant about
123 which character(s) would be present in the target (to-be-retrodicted or to-be-predicted) segment.
124 We hypothesized that character-cued responses should be more accurate than uncued responses,
125 to the extent that participants incorporate the character information we provided to them into their
126 retrodictions and predictions. To generate updated responses, we asked participants to watch an
127 additional segment that came just prior to or just after the target segment, and then to update their
128 retrodiction (*c-RP*) or prediction (*c-PR*) about the target segment. Results on updated responses are
129 not reported in this paper. Finally, we also asked participants to *recall* what happened in the just-
130 watched segment. We labeled these responses according to which other segments participants had
131 watched prior to the just-watched target. Retrodiction-matched recall (*re(R)*) responses were made
132 during the retrodiction sequences (B1 and B2; Fig. 2), whereas prediction-matched recall (*re(P)*)
133 responses were made during the prediction sequences (A1 and A2; Fig. 2). Whereas retrodiction
134 and prediction responses reflect what participants *estimate* they would remember after watching
135 the (inferred) target segment, recall responses provide a benchmark for comparison by measuring
136 what they *actually* remember about the target segment.

137 For each retrodiction and prediction, participants were asked to generate at least one, and not
138 more than three, responses that constituted “the sorts of things [the participant would] expect
139 to have remembered if [they] had watched the [target] segment.” They were asked to generate
140 multiple responses only if those additional responses were (in their judgement) of equal likelihood
141 to occur. On average, participants generated 1.08 responses per prompt; therefore we chose to
142 consider only participants’ first (“most probable” or “most important”) responses to each prompt.

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 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.

143 We also discarded a small number ($n = 20$) of character-cued responses that did not contain
144 references to all cued characters, along with one additional response due to the participant's
145 misunderstanding of the task instructions during that trial. We carried out our analyses on the
146 remaining 2084 retrodiction, prediction, and recall responses.

147 We used two general approaches to assess the quality of participants' responses (see *Methods*,
148 Fig. 3A). One approach entailed manually annotating events in the video and counting the number
149 of matched events in participants' responses. We identified a total of 117 unique events reflected
150 across the 22 video segments (range: 3–9 per segment; see *Methods*, Table S1). We assigned
151 one “point” to each of these video events. We also identified 23 additional events in participants'
152 responses that were either summaries of several events or that were partial matches to the manually
153 identified video events. We assigned 0.5 point to each of these additional events. This point
154 system enabled us to compute the numbers and proportions (*hit rates*) of correctly retrodicted,
155 predicted, and recalled events contained in each response. Our second approach entailed using
156 a natural language processing model (Cer et al., 2018) to embed annotations and responses in
157 a 512-dimensional feature space. This approach was designed to capture conceptual overlap
158 between responses that were not necessarily tied to specific events. To quantify this conceptual
159 overlap, we computed the similarities between the embeddings of different sets of responses.
160 Following Heusser et al. (2021), we defined the *precision* of each participants' retrodictions or
161 predictions about a target segment as the median cosine similarities between the embeddings
162 of (a) the participant's retrodiction or prediction response for the target segment and (b) each
163 *other* participant's recalls of the same segment. In other words, precision is designed to measure
164 the extent to which retrodictions and predictions captured the conceptual content that (other)
165 participants remembered. We also developed a related measure, which we call *convergence*, to
166 characterize response similarities across participants. In particular, we defined convergence as the
167 mean cosine similarity between the embeddings of a participant's responses to a target segment
168 and all other participants' responses (of the same type) to the same segment. We analyzed the
169 data using generalized linear mixed models, with participant and stimulus (e.g., target segment)
170 identities as crossed random effects (see *Methods*).

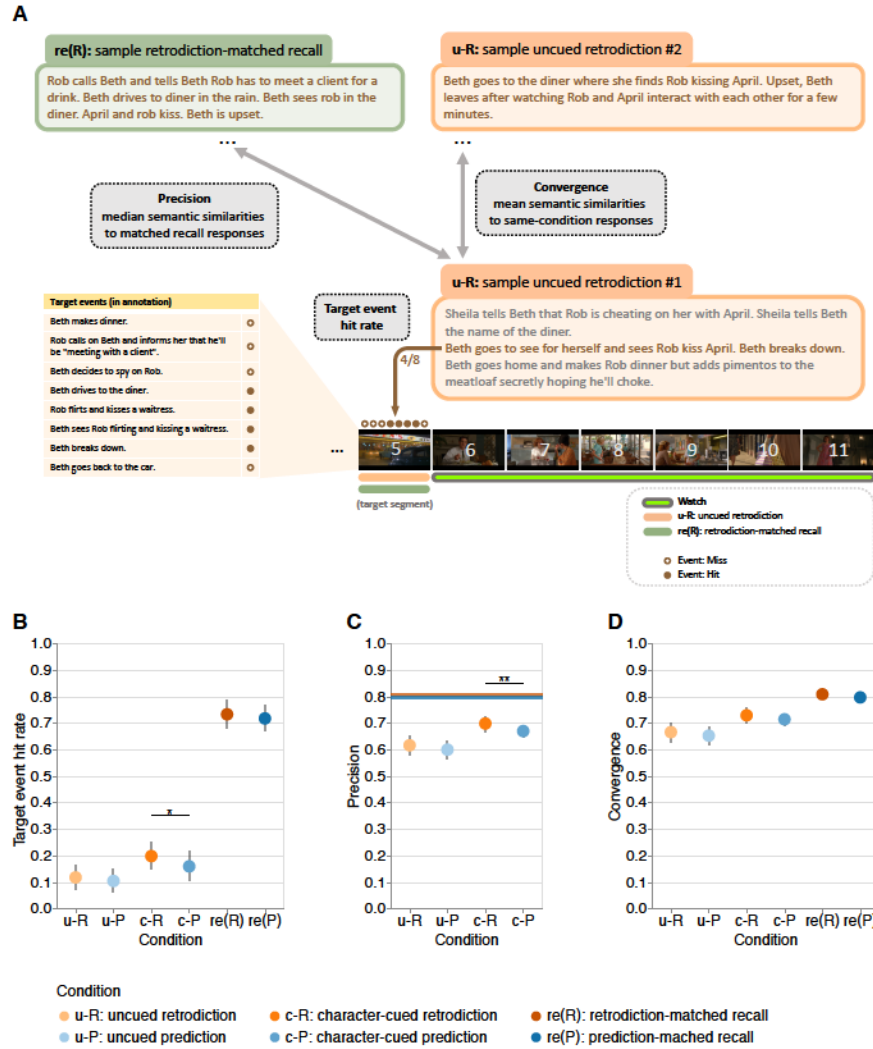


Figure 3: Retrodiction, prediction, and recall performance by experimental condition. **A. Methods schematic.** For each retrodiction, prediction, and recall response, we calculated the hit rate for events in the target segment, the response precision (see *Methods*), and the response convergence across 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 $p < 0.01$.

171 First we sought to validate a main effect of response type (i.e., uncued responses, character-
 172 cued responses, and recalls), irrespective of the temporal direction (retrodiction versus prediction).
 173 Across these three types of responses, participants have access to increasing amounts of infor-
 174 mation about the target segment. Therefore, across these response types, we hypothesized that
 175 participants' responses should become both more accurate and more convergent across individ-
 176 uals. Consistent with this hypothesis, participants' character-cued retrodictions and predictions
 177 were associated with higher target event hit rates than uncued retrodictions and predictions (odds
 178 ratio (OR): 2.65, $Z = 4.24$, $p < 0.001$, 95% confidence interval (CI): 1.69 to 4.16; Fig. 3B). These
 179 character-cued responses were also more precise ($b = 0.13$, $t(18.1) = 9.43$, $p < 0.001$, CI: 0.10 to
 180 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;
 181 Fig. 3D). Relative to character-cued responses, participants' recalls showed higher target event hit
 182 rates (OR = 21.83, $Z = 10.61$, $p < 0.001$, CI: 12.35 to 38.59) and were more convergence across
 183 individuals ($b = 0.20$, $t(19.4) = 9.10$, $p < 0.001$, CI: 0.16 to 0.25). These results are consistent with
 184 the common-sense notion that access to more information about a target segment yields better
 185 performance (i.e., higher hit rates, precision, and convergence across individuals).

186 Next we carried out a series of analyses specifically aimed at characterizing temporal direc-
 187 tion effects— i.e, the relative quality of retrodictions versus predictions across different types of
 188 responses. We hoped that these analyses might provide insights into our central question about
 189 whether inferences about the past and future are equally accurate. Across both uncued and
 190 character-cued responses (Fig. 2), retrodictions had numerically higher hit rates than predictions
 191 (Fig. 3B). However, these differences were only statistically reliable for character-cued responses
 192 (uncued responses: OR = 1.17, $Z = 0.35$, $p = 0.73$, CI: 0.47 to 2.92; character-cued responses: OR =
 193 1.93, $Z = 2.15$, $p = 0.03$, CI: 1.06 to 3.52). We observed a similar pattern of results for the precisions
 194 of participants' responses (Fig. 3C). Specifically, their responses tended to be numerically more
 195 precise for retrodictions versus predictions, but the differences were only statistically reliable for
 196 character-cued responses (uncued responses: $b = 0.03$, $t(20.9) = 1.09$, $p = 0.29$, CI: -0.03 to 0.10;
 197 character-cued responses: $b = 0.06$, $t(20.8) = 3.01$, $p = 0.007$, CI: 0.02 to 0.11). We also consistently
 198 observed numerically higher convergence across participants for retrodictions versus predictions

(Fig. 3D), but neither of these differences were statistically reliable (uncued responses: $b = 0.03$, $t(17.9) = 0.75$, $p = 0.46$, CI: -0.05 to 0.11; character-cued responses: $b = 0.04$, $t(17.4) = 1.46$, $p = 0.16$, CI: -0.02 to 0.09). Taken together, these results suggest that participants are generally better at making retrodictions than predictions. We also verified that this was not solely a consequence of how participants' memory performance might have been affected by watching different segments (or making different responses to other segments) across conditions by comparing recall responses in the retrodiction-matched recall ($re(R)$) and prediction-matched recall ($re(P)$) conditions. Recall performance was similar in both conditions (target event hit rate: OR = 1.12, $Z = 1.07$, $p = 0.29$, CI: 0.91 to 1.39; convergence: $b = 0.03$, $t(19.3) = 1.89$, $p = 0.07$, CI: 0.00 to 0.07).

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

Because there is no "ground truth" number of offscreen events, we could not compute the hit rates for offscreen events. Instead, we counted up the absolute *number* of retrodicted or predicted events as a function of lag. In other words, given that the participant had just watched segment i ,

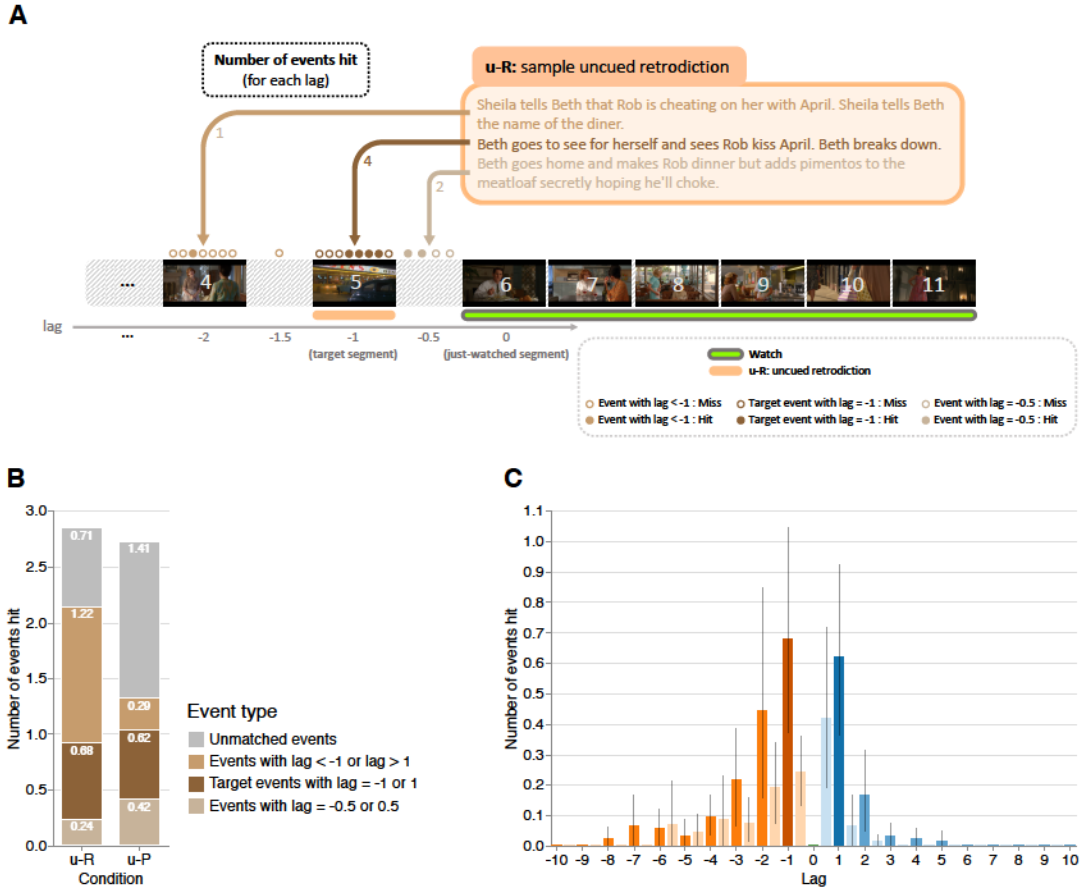


Figure 4: Retrodictions and predictions of temporally near and distant events. **A.** Illustration of annotation approach. For each uncued retrodiction and prediction response, we calculated the number of (retrodicted or predicted) events as a function of temporal distance from the target segment, or *lag*. Onscreen (explicit) events are tagged using integer-valued lags, whereas offscreen (implicit) events are tagged using half-step lags (± 0.5 , ± 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 ($|\text{lag}| > 1$), 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).

227 we asked how many events from segment $i + lag$ they retrodicted or predicted, on average, given
 228 that they were aiming to retrodict or predict events at lags of ± 1 . We also counted the numbers of
 229 *unmatched* events in participants' responses that did not correspond to any events in the relevant
 230 segments of the narrative. We focused specifically on *uncued* retrodictions and predictions, which
 231 we hypothesized would provide the cleanest characterizations of participants' initial estimates of
 232 the unobserved past and future (i.e., without potential biases introduced by additional character
 233 information, as in the character-cued responses). The numbers of uncued retrodicted and predicted
 234 target ($lag = \pm 1$) events were not reliably different ($OR = 0.92, Z = -0.15, p = 0.88, CI: 0.30$ to 2.84).
 235 In other words, uncued retrodictions and predictions over short timescales did not exhibit reliable
 236 asymmetries. However, when retrodicting, participants mentioned events from the distant past
 237 ($lag < -1$) more often than participants predicted events from the distant future ($lag > 1$; $OR =$
 238 $9.10, Z = 3.80, p < 0.001, CI: 2.92$ to 28.39 ; Fig. 4B, C; for results from the character-cued conditions,
 239 see Fig. S2). Despite this asymmetry in the accuracies of participants' long-range retrodictions
 240 versus predictions, there were no reliable differences in the *numbers* of uncued retrodicted versus
 241 predicted events (across all lags; $OR = 1.05, Z = 0.75, p = 0.45, CI: 0.93$ to 1.18). Nor did we find any
 242 reliable differences in the numbers of offscreen events immediately before or after the just-watched
 243 segment ($lag = \pm 0.5$; $OR = 0.75, Z = -0.36, p = 0.72, CI: 0.15$ to 3.59). The apparent discrepancy
 244 between participants' asymmetric accuracy but symmetric event counts was due to participants'
 245 tendencies to reference "unmatched" events (i.e., events that did not correspond to any explicit
 246 or implicit event in the story) more in their predictions than retrodictions ($OR = 0.36, Z = -4.53,$
 247 $p < 0.001, CI: 0.23$ to 0.56). We confirmed that the retrodiction advantage held when controlling
 248 for absolute lag ($OR = 34.31, Z = 3.28, p = 0.001, CI: 4.16$ to 283.20), for onscreen events alone (OR
 249 $= 47.54, Z = 3.74, p < 0.001, CI: 6.27$ to 360.60), and marginally for offscreen events alone ($OR =$
 250 $24.76, Z = 1.71, p = 0.09, CI: 0.63$ to 975.27). Taken together, these analyses show that (in generating
 251 uncued responses) participants tend to reach "further" into the unobserved past, and with greater
 252 accuracy, than the unobserved future.

253 What might be driving participants to retrodict further and more accurately into the unob-
 254 served past, compared with their predictions of the unobserved future? By inspecting the video

255 content, we noticed that characters in the television show frequently referenced both past events
 256 and (planned or predicted) future events in their spoken conversations. We wondered whether the
 257 characters' references might show temporal asymmetries that might explain participants' behav-
 258 iors. Across all of the characters' conversations, and across all of the video segments, we manually
 259 identified a total of 82 references to past or future events (i.e., that occurred onscreen or offscreen
 260 before or after the events depicted in the current segment; Fig. 5A, S3A). Characters tended to
 261 reference the past (52 references) more than the future (30 references), consistent with previous
 262 work (Demiray et al., 2018). References to the past were also skewed to more temporally distant
 263 events compared with references to the future (Figs. 5B, S3B). These observations indicate that the
 264 characters in the stimulus display a preference for the past (versus future) in their conversations.
 265 Might this asymmetry be driving the asymmetries in participants' retrodictions versus predictions?
 266 Controlling for temporal distance (lag), past and future events that story characters referenced
 267 in their conversations were associated with higher hit rates than unreferenced events (uncued
 268 retrodiction: $OR = 12.70$, $Z = 10.94$, $p < 0.001$, $CI: 8.06$ to 20.03 ; uncued prediction: $OR = 8.29$,
 269 $Z = 6.83$, $p < 0.001$, $CI: 4.52$ to 15.20 ; Fig. 5E). This indicates that participants' responses are at least
 270 partially influenced by the characters' conversations. To estimate the contributions of characters'
 271 references on hit rates, we computed the difference in hit rates between all events (which comprised
 272 both referenced and unreferenced events) and unreferenced events, as a function of lag. These
 273 differences exhibited a temporal asymmetry in favor of retrodiction (Fig. 5C). This indicates that the
 274 asymmetries in participants' retrodictions versus predictions are also at least partially influenced by
 275 the characters' conversations. However, these temporal asymmetries in participants' retrodictions
 276 and predictions persisted even for events that characters never referenced in their conversations
 277 (hit rates of uncued retrodicted versus predicted unreferenced events: $OR = 2.00$, $Z = 2.40$, $p = 0.02$,
 278 $CI: 1.14$ to 3.51 ; Fig. 5D). When we further separated the unreferenced events into onscreen events
 279 and offscreen events, we found that these asymmetries held only for the onscreen events (onscreen:
 280 $OR = 2.65$, $Z = 2.59$, $p = 0.01$, $CI: 1.27$ to 5.54 ; offscreen: $OR = 1.50$, $Z = 0.91$, $p = 0.36$, $CI: 0.63$
 281 to 3.62). Taken together, these analyses suggest that asymmetries in the number of references
 282 characters make to past and future events partially (but not entirely) explain why participants tend

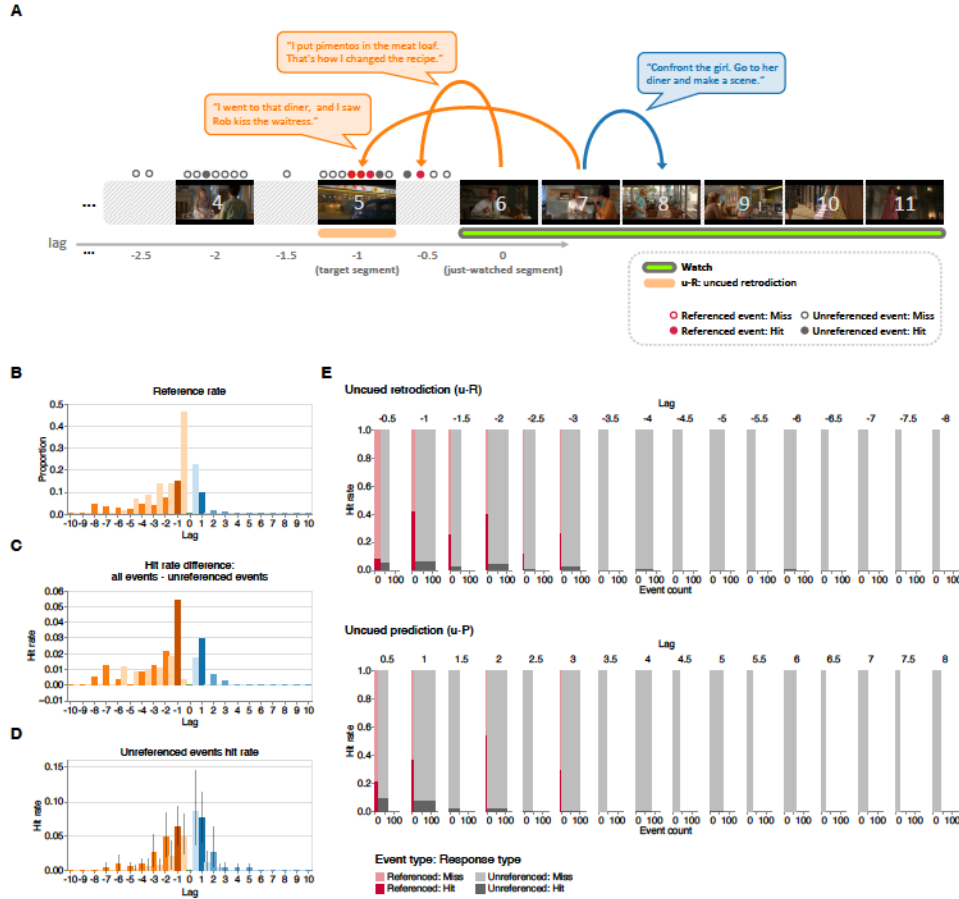


Figure 5: Characters' references drive participants' retrodiction and prediction performance. **A.** Illustration of annotation approach. We manually annotated references to events in past or future segments in characters' spoken conversations. We matched each such reference with its corresponding storyline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

283 to retrodict the past further and more accurately than they predict the future.

284 If characters' direct references cannot fully account for the temporal asymmetry in retrodicting
285 the unobserved past versus predicting the unobserved future, what other factors might explain this
286 phenomenon? The results above indicate that characters' references to specific unobserved events
287 in the past or future boost participants' estimates of these events. If there are associations and
288 dependencies between temporally adjacent events, might characters' references to specific events
289 also boost participants' estimates of other events that were temporally *adjacent* to the referenced
290 events (Fig. 6A)? Because characters tended to refer to past events more often than future events,
291 the proportions of unreferenced events that were adjacent to referenced events should show a
292 similar temporal asymmetry in favor of the past. We tested this intuition by computing the
293 proportions of unreferenced events in the stimulus that were temporally adjacent to past or future
294 events referenced by the characters during a given segment. Here we defined *temporally adjacent*
295 as any event within an absolute lag of one relative to a referenced onscreen event, or within an
296 absolute lag of 0.5 to a referenced offscreen event. We also defined *remaining* events as unreferenced
297 events that were not temporally adjacent to any referenced events. As shown in Figure 6B, we
298 observed higher proportions of unreferenced past than future events that were temporally adjacent
299 to referenced events. Further, these reference-adjacent events had higher hit rates than remaining
300 events after controlling for absolute lag (uncued retrodiction: $OR = 7.15$, $Z = 2.40$, $p = 0.02$, CI:
301 1.44 to 35.58; uncued prediction: $OR = 3.11$, $Z = 2.30$, $p = 0.02$, CI: 1.18 to 8.21; Fig. 6E). To estimate
302 the contributions of reference adjacency on hit rates, we computed the difference in hit rates
303 between unreferenced events (which comprised both reference-adjacent and remaining events)
304 and remaining events, as a function of lag. These differences exhibited a temporal asymmetry in
305 favor of retrodiction. This suggests that reference-adjacent events also contribute to participants'
306 retrodiction advantage. Remaining events did *not* exhibit a reliable temporal asymmetry ($OR =$
307 0.75, $Z = 0.33$, $p = 0.74$, CI: 0.14 to 4.08; Fig. 6D), suggesting that, after accounting for temporal
308 adjacency, character's references to past and future events can explain participants' retrodiction
309 advantage.

310 The preceding analyses show that when characters reference past or future events, those refer-

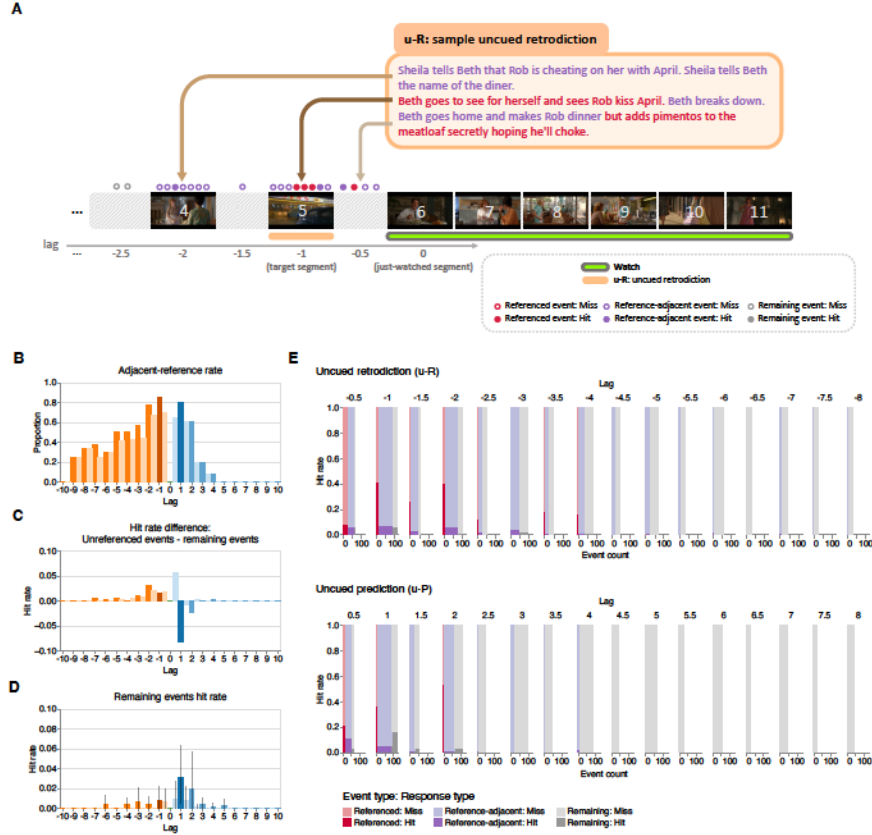


Figure 6: Reference-adjacent events are associated with higher hit rates. **A.** Illustration of annotation approach. We extended the annotation procedure depicted in Figure 5A to also label unreferenced events that were either temporally adjacent to (i.e., immediately preceding or proceeding) a referenced 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).

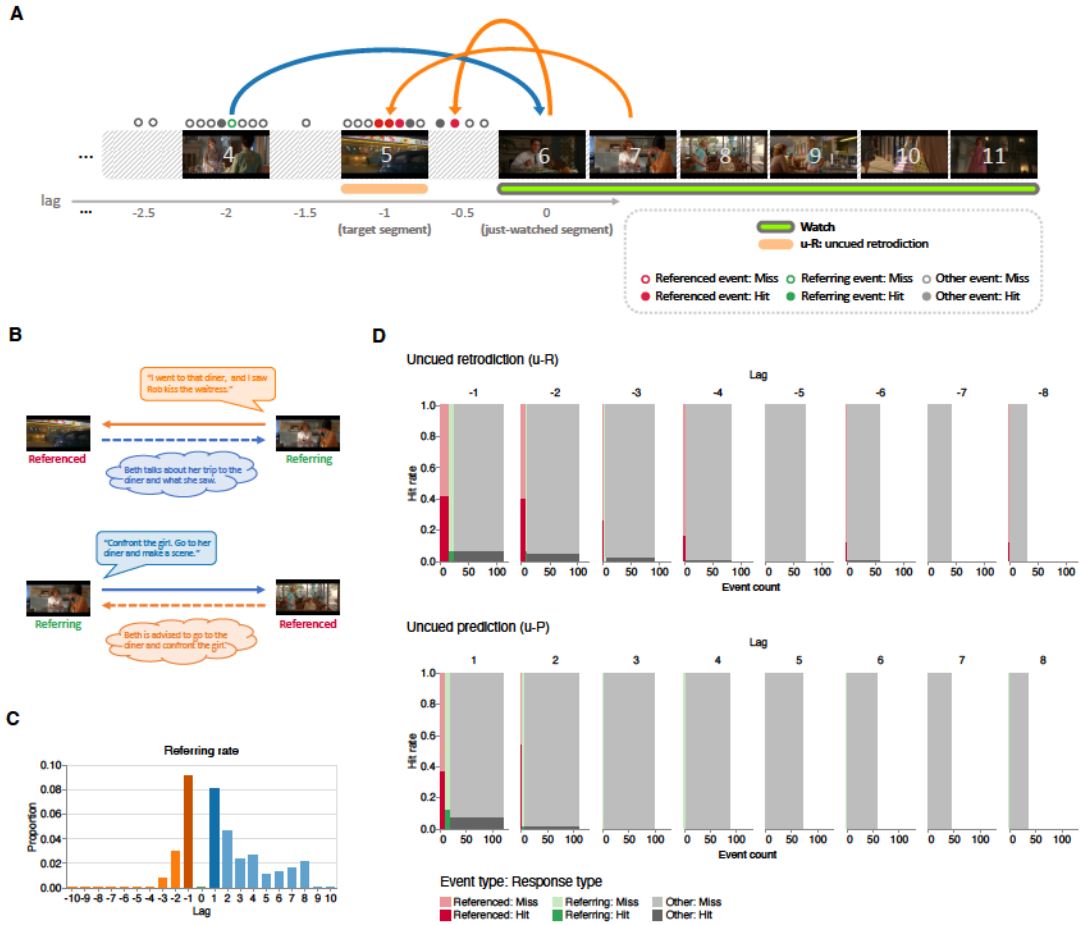


Figure 7: Referenced events are associated with higher hit rates, but referring events are not. **A.** Illustration of annotation approach. We extended the annotation procedure depicted in Figure 5A to also label which events *contained* references to events in other segments. **B.** Referenced versus 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).

311 enced events, and other events that are temporally adjacent to the referenced events, are more likely
 312 to be retrodicted and predicted. In other words, referring to a past or future event in conversation
 313 leads to a “boost” in that event’s hit rate. We wondered whether this boost was bi-directional. In
 314 particular: when a character refers (during a *referring event*) to another event (i.e., the *referenced*
 315 *event*), does this boost only the referenced event’s hit rate, or does the referring event also receive
 316 a boost? We labeled each event as a “referring event,” a “referenced event,” or a “other event”
 317 (i.e., not referring or referenced; Fig. 7A, B). We limited our analysis to references to onscreen
 318 (explicit) events. Consistent with our analysis of the proportions of referenced events (Fig. 5B), the
 319 proportions of *referring* events exhibited a *forward* temporal asymmetry (Fig. 7C). Controlling for
 320 absolute lag, we found that referring events were associated with lower hit rates than referenced
 321 events (uncued retrodiction: $OR = 0.03$, $Z = -4.81$, $p < 0.001$, CI: 0.01 to 0.11; uncued prediction:
 322 $OR = 0.04$, $Z = -5.84$, $p < 0.001$, CI: 0.01 to 0.12; Fig. 7D) and had no reliable differences in hit
 323 rates compared with other events (uncued retrodiction: $OR = 0.37$, $Z = -1.46$, $p = 0.15$, CI: 0.10 to
 324 1.41; uncued prediction: $OR = 2.16$, $Z = 1.68$, $p = 0.09$, CI: 0.88 to 5.30). This indicates that only
 325 referenced events received a hit rate boost (relative to other events), suggesting that the retrodictive
 326 and predictive benefits of references are directed (i.e., asymmetric).

327 Discussion

328 We asked participants to watch sequences of movie segments from a character-driven television
 329 drama and then either retrodict what had happened prior to a just-watched segment, predict what
 330 would happen next, or recall what they had just watched. We found that participants tended
 331 to more accurately and more readily retrodict the unobserved past than predict the unobserved
 332 future. We traced this temporal asymmetry to (a) characters’ tendencies to refer to past events
 333 more than future events in their ongoing conversations, and (b) associations between temporally
 334 proximal events (Fig. 8). Essentially, associations between temporally proximal events serve to
 335 enhance asymmetries in inferences driven by conversational references (light orange and blue bars
 336 in Fig. 8). Our findings show that other peoples’ psychological arrows of time can affect external

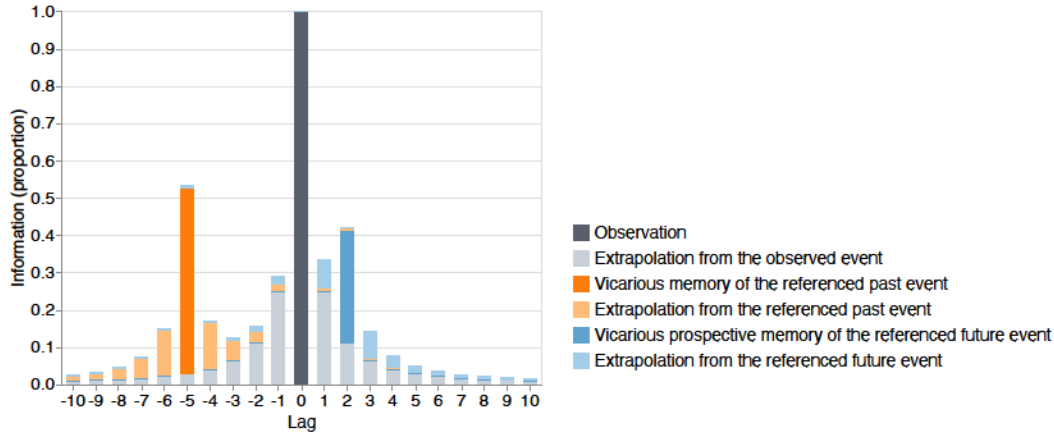


Figure 8: How much information about the past and future can be inferred by observing the present? By definition, let us say that the present moment (lag 0) contains all information about itself (dark gray). Given learned statistical regularities, one might extrapolate from the present moment into the 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).

observers' inferences about the unobserved past and future.

When people communicate through language or other observable behaviors, they can transmit their knowledge and memories to others (Hirst and Echterhoff, 2012; Mahr and Csibra, 2018; Dessalles, 2007; Zadbood et al., 2017). A consequence of this sharing across people is that biases or limitations in one person's knowledge and memories may also be transmitted to external observers. Although people *can* communicate their intentions and future plans (i.e., information about their future), because people know *more* about their pasts than their futures, the knowledge transmitted to observers is inherently biased in favor of the past (Fig. 8; Demiray et al., 2018). Since observers leverage communicated knowledge to reconstruct the unobserved past and future, this explains why observers' inferences about observed people's lives also favor the past.

People's knowledge asymmetries are not always directly observable. For example, in a conversation where someone talks exclusively about their future plans, a passive observer might gain

349 more insight into the speaker's unobserved future than their unobserved past. However, because
350 the speaker is also guided by their own psychological arrow of time, the "upper limit" of knowledge
351 about their past is still higher than that of their future. Therefore, after accounting for knowledge
352 that *could* be revealed through active participation in the conversation, the seemingly future-biased
353 conversation masks an underlying knowledge asymmetry in favor of the past. This hypothesized
354 "unmasking" effect of interaction implies that the influence of other people's psychological arrows
355 of time should be more robust when the receiver is an active participant in the conversation. Other
356 social dimensions, such as trust, motivation or level of engagement, personal goals, and beliefs,
357 might serve to modulate the effective "gain" of the communication channel– i.e., how much the
358 speaker's knowledge influences the observer's knowledge.

359 In typical statistical sequences used in laboratory studies, there is no temporal asymmetry,
360 either theoretically (Cover, 1994; Bialek et al., 2001; Ellison et al., 2009), or empirically (Jones and
361 Pashler, 2007). What makes narratives and real-world event sequences time-asymmetric? Of
362 course there are many superficial differences between simple laboratory-manufactured sequences
363 and real-world experiences. As one example, real-world experiences often involve other people
364 who have their own memories and goals. At a deeper level, however, are our subjective experi-
365 ences essentially more complicated versions of laboratory-manufactured sequences? Or are there
366 fundamental differences? One possibility is that real-life event sequences are not stationary (i.e.,
367 not in equilibrium). For example, real-life events might start from a special initial condition (Albert,
368 2000; Feynman, 1965) and proceed through a series of transitions from more-ordered to less-ordered
369 states, thus exhibiting an arrow time. When we retrodict, it is possible that we only consider possi-
370 ble past events that are compatible with the highly-ordered special initial state (Carroll, 2010, 2016).
371 For example, when we see a broken egg we might infer that the egg had been intact at some point in
372 the past. But it would be difficult to guess at what states or forms the broken egg might take in the
373 future (Carroll, 2010, 2016). In other words, the procession from order to disorder might result in
374 better retrodiction performance compared with that of (implicitly less-restricted) prediction tasks.
375 The special initial state might also explain why we remember the past, but not the future. Some
376 recent work suggests that the psychological arrow of time might be explained by a related concept

377 in the statistical physics literature, termed the “thermodynamic” arrow of time (Mlodinow and
378 Brun, 2014; Rovelli, 2022). However, the relation between the thermodynamic and psychological
379 arrows of time is still under debate (Gołosz, 2021; Hemmo and Shenker, 2019).

380 In our study, we explicitly designed participants’ experiences such that both the past and future
381 were unobserved. How representative is this scenario of everyday life? For example, we might
382 try to speculate about the unobserved future when making plans or goals, but when might we
383 encounter situations where the past is unobserved but still useful for us to speculate about? Real-life
384 events have long-range dependencies. In general, because the future depends on what happened
385 in the past, discovering or estimating information about the unobserved past can help us form
386 predictions about the future. We illustrate this point in Figure 8 by showing that the additional
387 information contributed by a referenced past event can also extend into the future (light orange bars
388 at lags > 0). This might explain why humans devote substantial effort and resources to attempting
389 to figure out what happened in the unobserved past: history, anthropology, geology, detective and
390 forensic science, and other related fields are each primarily focused on understanding, retrodicting,
391 or reconstructing unobserved past events.

392 **Methods**

393 **Participants**

394 A total of 36 participants (25 female, mean age 21.47 years, range 19–50 years) were recruited from
395 the Dartmouth College community. All participants had self-reported normal or corrected-to-
396 normal vision, hearing, and memory, and had not watched any episodes of *Why Women Kill* before
397 the experiment. Participants gave written consent to enroll in the study under a protocol approved
398 by the Committee for the Protection of Human Subjects at Dartmouth College. Participants received
399 course credit or monetary compensation for their time. Two participants completed only the first
400 half of the study and one participant’s data from the second half of their testing session was lost
401 due to a technical error. All available data were used in the analyses.

402 Stimuli

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

413 Task design and procedure

414 Our experimental paradigm was divided across two testing sessions. In each session, participants
415 performed a sequence of tasks on segments from one storyline (Fig. 2). For each storyline, there
416 were four different task sequences: two forward chronological order sequences and two backward
417 chronological order sequences. Participants completed one task sequence in forward chronological
418 order for one storyline, and one in backward chronological order for the other storyline. The order
419 of the two sessions (forward chronological order sequence first or backward chronological order
420 sequence first), and the pairing of task sequences with storylines, were counterbalanced across
421 participants.

422 Tasks in each sequence alternated between watching, recall, and retrodiction or prediction,
423 with the specific order of tasks differing across the four sequences. For example, in sequence A1,
424 participants first watched segment 1, followed by an immediate recall of segment 1. Then they
425 predicted what would happen in segment 2 (first uncued and then character-cued). Participants
426 then watched segment 3 and recalled segment 3. After that, participants guessed what happened in
427 segment 2 again, which we termed "updated prediction". Then they watched segment 2, recalled

428 segment 2, and so on as depicted in Figure 2. This procedure was repeated to cover all possible
429 segments. We also note several edge cases at the start and end of the narrative sequences. Since
430 no segments precede the first segment, participants could never make “prediction” responses with
431 the first segment as their target. For analogous reasons, participants never made “retrodiction”
432 responses with the last segment as their target. Another edge case occurred in task sequences
433 B2 and A2 (Fig. 2). In the A1 and A2 sequences, participants experience the narrative in the
434 original (forward) order, predicting one segment ahead along the way. In the B1 and B2 sequences,
435 participants experience the narrative in the reverse order, retrodicting one segment ahead along
436 the way. However, because A2 and B2 are offset from A1 and B1 by one segment, the initial A2
437 responses are *retrodictions*, and the initial B2 responses are *predictions* (i.e., they conflict with the
438 temporal directions of the remaining responses in those conditions). We therefore excluded from
439 our analysis those initial retrodiction responses from the A2 condition, and the initial prediction
440 responses from the B2 condition.

441 Before watching each segment, participants were given the following task instructions. After
442 watching the video, participants were instructed to type their responses (retrodiction, prediction,
443 or recall) in 1–4 sentences. Participants were also asked to specify the characters’ names in their
444 responses, i.e., avoiding use of characters’ pronouns. For the recall task, the names of the characters
445 in the recall segment were displayed, and participants were asked to summarize the major plot
446 points in the present tense. For the retrodiction and prediction tasks, participants were instructed
447 to retrodict or predict the major plot points of the segment (also in the present tense), as though
448 they had watched the segment and were writing a plot synopsis. They were also instructed to
449 avoid speculation words (e.g., “I *think* Beth will...”). For the uncued retrodiction and prediction
450 tasks, participants made retrodictions or predictions without any cues provided, so they had to
451 guess which of the characters would be present in the segment. For character-cued retrodictions
452 and predictions, the characters in the target segment were revealed on the screen, alongside
453 participants’ previous responses. Participants were instructed to include or incorporate those
454 characters into their character-cued responses, if their previous responses did not contain all the
455 characters provided. They were also told that the characters were not necessarily listed in their

order of appearance in the segment, and that only the main characters would be given. Also, the characters given did not necessarily interact with each other in that segment, and they could appear in successive events in that segment. If participants' previous responses included all the characters given, then they could directly proceed to the next task without updating their responses. For all of the prediction and retrodiction tasks, participants were instructed to provide at least one response, but they were given the opportunity enter up to three responses if they felt that multiple possibilities were more or less equally likely. Each response (including recall) was followed by a confidence rating on a 1–5 point scale. However, these confidence data were not analyzed in the present study.

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

Video annotation

Events in the first 11 segments of the two storylines were identified by the first author (X.X.), corresponding to major plot points (total: 117; mean: 5.32 per segment; range 3–9). Additionally, 74 offscreen events were identified. Of these 74 offscreen events, 43 events were identified from

483 references in conversations during onscreen events. Another 16 events were identified based on
484 characters' implied movements and travels. For example, if in segment 1 character A was in place
485 A and in segment 2 she was in place B, then the transit from place A to B for character A would be
486 identified as an offscreen event. The remaining 15 offscreen events were identified based on logical
487 inferences. For example, if a photograph was shown in an onscreen event (but not the act of the
488 photograph being taken), then the action that someone took the photograph would be identified
489 as an offscreen event. Offscreen events always occurred between two contiguous segments, or
490 before the first segment. The purpose of identifying offscreen events was to match participants'
491 responses to video events; thus our identification of these offscreen events was not intended to be
492 exhaustive.

493 **Response analyses**

494 Participants' retrodiction, prediction, and recall responses were minimally processed to correct
495 obvious typos (e.g., in characters' names) and remove speculation descriptions (e.g., "I predict
496 that..."). All responses were manually coded and matched to events from the video annotations.
497 Retrodiction and prediction responses were coded by two coders (X.X. and Z.Z.). Recall responses
498 were coded by one coder (X.X.). While most responses were clearly identifiable as either matching
499 specific storyline events or as not matching any storyline events, several ambiguous cases arose.
500 First, some responses combined or summarized over several (distinct) storyline events. Second,
501 some responses lacked any specific detail (e.g., "character A and B talk" without describing the
502 specific topic(s) of conversation or providing other relevant details). Based on participants' re-
503 sponses, in addition to the original 117 onscreen events and 74 offscreen events, we added 25 new
504 events (23 onscreen, 2 offscreen) that either summarized across several events or partially matched
505 the annotated events. Whereas the original events were each assigned a value of one point, we
506 assigned these additional events a half point. This point system enabled us to directly match events
507 in participants' responses to the annotated events. In our analyses of retrodictions, predictions,
508 and recalls, we added up the number of points earned for each response to estimate participants'
509 event hit rates.

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

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

522 To estimate the semantic similarities between pairs of responses, we first transformed each
523 response into a 512-dimensional vector (embedding) using the Universal Sentence Encoder (Trans-
524 former USE, Cer et al., 2018). We defined *similarity* as the cosine of the angle formed by the
525 responses’ vectors. Following Heusser et al. (2021), we defined the *precision* of participants’ re-
526 sponses as the median similarity between that response’s vector and the embedding vectors for
527 all other participants’ recalls of the target segment. We defined the *convergence* of a given response
528 as the mean similarity between that response’s vector and all other participants’ responses to the
529 corresponding segment, in the same condition. To compute these median or mean similarities we
530 first applied the Fisher z-transformation to the similarity values, then took the median or mean
531 of the z-transformed similarities, and finally applied the inverse z-transformation to obtain the
532 precision or convergence score.

533 To test the validity and reliability of the USE embeddings, we performed a classification analysis
534 of recall responses using a leave-one-out approach. For each recall response, we calculated its
535 semantic similarity with all other recall responses for the same storyline. We took the segment
536 with the highest median semantic similarity (to the recall response) as the “predicted” segment.
537 Across all responses, the predicted segments matched the true recalled segments’ labels 98.6% of

the time (1088 out of 1103 predictions; chance level: 9%).

Reference coding

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

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

Statistical analysis

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

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

```
565 cbind(thp, ttp - thp) ~ direction * level * seg_cnt * storyline +  
566 (direction * level | target) +  
567 (direction * level * seg_cnt | subject)
```

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

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

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

```
583 hp_lag ~ direction * full_stp * lag * storyline +  
584 (direction | base_seg) + (1 | base_seg_pair) +  
585 (direction * full_stp | lag * storyline | subject)
```

586 where hp_lag is the number of “points” earned (for each lag) in each trial (we manually doubled
587 the point counts to ensure that half points were mapped onto integers, for compatibility with the
588 Poisson link function), full_stp denoted whether the given events (of the given lag) were onscreen
589 (i.e., full step) or offscreen (i.e., half step), lag denotes the (centered) absolute lag, base_seg denotes

the identity of the just-watched segment (22 levels), and `base_seg_pair` denotes the pairing of the just-watched segment and the segment at each lag (440 levels).

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

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

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

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

Code and data availability

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

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

References

- Albert, D. Z. (2000). *Time and chance*. Harvard University Press, Cambridge, Mass.
- Bialek, W., Nemenman, I., and Tishby, N. (2001). Predictability, complexity, and learning. *Neural Computation*, 13(11):2409–2463.

- 612 Bordwell, D. (2008). *Poetics of cinema*, chapter Three dimensions of film narrative, pages 85–134.
613 Routledge.
- 614 Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*,
615 11(2):177–220.
- 616 Carroll, S. (2010). *From eternity to here: the quest for the ultimate theory of time*. Penguin.
- 617 Carroll, S. (2016). *The big picture: on the origins of life, meaning, and the universe itself*. Dutton.
- 618 Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,
619 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.
620 *arXiv*, 1803.11175.
- 621 Cover, T. M. (1994). Which processes satisfy the second law? In Halliwell, J. J., Pérez-Mercader,
622 J., and Zurek, W. H., editors, *Physical Origins of Time Asymmetry*, pages 98–107. Cambridge
623 University Press, Cambridge, UK.
- 624 de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web
625 browser. *Behavior Research Methods*, 47(1):1–12.
- 626 Demiray, B., Mehl, M. R., and Martin, M. (2018). Conversational time travel: evidence of a
627 retrospective bias in real life conversations. *Frontiers in Psychology*, 9:2160.
- 628 Dessalles, J.-L. (2007). Storing events to retell them. *Behavioral and Brain Sciences*, 30(3):321–322.
- 629 Ellison, C. J., Mahoney, J. R., and Crutchfield, J. P. (2009). Prediction, retrodiction, and the amount of
630 information stored in the present. *Journal of Statistical Physics*, 136(1005):doi.org/10.1007/s10955–
631 009–9808–z.
- 632 Feynman, R. (1965). *The character of physical law*. MIT Press.
- 633 Gołosz, J. (2021). Entropy and the direction of time. *Entropy*, 23(4):388.
- 634 Hawking, S. W. (1985). Arrow of time in cosmology. *Physical Review D*, 32(10):2489–2495.

- 635 Hemmo, M. and Shenker, O. (2019). The second law of thermodynamics and the psychological
636 arrow of time. *The British Journal for the Philosophy of Science*.
- 637 Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behavioral
638 and neural signatures of transforming naturalistic experiences into episodic memories. *Nature*
639 *Human Behavior*, 5:905–919.
- 640 Hirst, W. and Echterhoff, G. (2012). Remembering in conversations: the social sharing and reshaping
641 of memories. *Annual Review of Psychology*, 63(1):55–79.
- 642 Horwich, P. (1987). *Asymmetries in time: problems in the philosophy of science*. MIT Press.
- 643 Jones, J. and Pashler, H. (2007). Is the mind inherently forward looking? comparing prediction and
644 retrodiction. *Psychonomic Bulletin and Review*, 14(2):295–300.
- 645 Koster-Hale, J. and Saxe, B. (2013). Theory of mind: A neural prediction problem. *Neuron*,
646 79(5):836–848.
- 647 Lange, K., Kühn, S., and Filevich, E. (2015). “Just Another Tool for Online Studies” (JATOS): an
648 easy solution for setup and management of web servers supporting online studies. *PLoS One*,
649 10(6):e0130834.
- 650 Maheu, M., Meyniel, F., and Dehaene, S. (2022). Rational arbitration between statistics and rules
651 in human sequence processing. *Nature Human Behaviour*, pages 1–17.
- 652 Mahr, J. B. and Csibra, G. (2018). Why do we remember? the communicative function of episodic
653 memory. *Behavioral and Brain Sciences*, 41:e1.
- 654 Mlodinow, L. and Brun, T. A. (2014). Relation between the psychological and thermodynamic
655 arrows of time. *Physical Review E*, 89(5):052102.
- 656 Radvansky, G. A. and Copeland, D. E. (2006). Walking through doorways causes forgetting:
657 situation models and experienced space. *Memory and Cognition*, 34(5):1150–1156.

- 658 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature*
659 *Reviews Neuroscience*, 13:713–726.
- 660 Rovelli, C. (2022). Memory and entropy. *Entropy*, 24(8):1022.
- 661 Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive*
662 *Sciences*, 22(3):201–212.
- 663 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit mem-
664 ories to other brains: constructing shared neural representations via communication. *Cerebral*
665 *Cortex*, 27(10):4988–5000.
- 666 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and
667 memory. *Psychological Bulletin*, 123(2):162–185.

668 **Acknowledgements**

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

675 **Author contributions**

676 Conceptualization: X.X. and J.R.M.; Methodology: X.X. and J.R.M.; Software: X.X.; Analysis: X.X.
677 and Z.Z.; Writing, Reviewing, and Editing: X.X., Z.Z., and J.R.M.; Supervision: J.R.M.

⁶⁷⁸ **Competing interests**

⁶⁷⁹ The authors declare no competing interests.