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

Xinming Xu[1], Ziyan Zhu[2], Xueyao Zheng[3], and Jeremy R. Manning[1,⋆]

[1]Dartmouth College, Hanover, NH, USA

[2]Peking University, Beijing, China

[3]Beijing Normal University, Beijing, China

⋆Address correspondence to jeremy.r.manning@dartmouth.edu

October 9, 2023

Abstract

~~How much can we infer about the past and future, given our current knowledge and observations in the present? Inferences about our own lives are time-asymmetric: we are better able to infer the past than the future, since we remember our past but not our future. The asymmetry in our memories is known as the *psychological arrow of time*. But when both the past and future are unexperienced and unobserved, for example when we make inferences about *other* people's lives, are our inferences about the past and future time-symmetric or asymmetric? To study these questions, we had participants view segments of a character-driven television drama. They used free-form responses to guess at what would happen just before or just after each just-watched segment. We found that participants' inferences were time-asymmetric, in that they inferred past events more accurately than future events after controlling for their exposure to past and future events in the narrative. This asymmetry was driven by *characters'* biases in conversational references to their own pasts. Our work reveals a temporal asymmetry in how observations of other peoples' behaviors can inform us about the past and future.~~

~~**Keywords: episodic memory, prediction, retrodiction, narratives, conversation**~~ How much can we infer about the past and future, given our knowledge of the present? Unlike temporally symmetric inferences about simple sequences, inferences about our own lives are asymmetric: we are better able to infer the past than the future, since we remember our past but not our future (i.e., the psychological arrow of time). What happens when both the past and future are unobserved, as when we make inferences about *other* people's lives? We had participants in two experiments view segments of two character-driven television dramas. They wrote out what would happen just before or after each just-watched segment. Participants were better at inferring past (versus future) events. This asymmetry was driven by participants' reliance on characters' conversational references in the narrative, which tended to favor the past. We also carried out a meta analysis to estimate the prevalence of temporal asymmetries in past versus future conversational references in hundreds of millions of dialogues from television shows, popular movies, novels, and written and spoken natural conversations. We found that, on average, references to the past are 1.45 times more prevalent in human conversations than references to the future. Our work reveals a temporal asymmetry in how observations of other people's behaviors can inform us about the past and future.

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

Introduction

What we experience in the current moment tells us about *now*– but what does it tell us about the past or future? And does the current moment tell us, as human observers, *more* about the past or about the future? One way of examining ~~this question~~ these questions is to consider highly simplified scenarios that are artificially constructed in the laboratory (e.g., Maheu et al., 2022). At one extreme, for ~~fully observable deterministic sequences , observing any part of the sequence provides~~ deterministic sequences with *known* rules, knowing the current state provides the observer with sufficient information to exactly reconstruct the entire past and future history of the stimulus. ~~For example, if we observe the partial sequence "5, 6, 7, 8, 9, 10," we might reasonably guess that the just-preceding number was 4, and the just-proceeding number will be 11. If we were given additional information through further observations, or by being told about the rules for generating the sequence, our confidence would increase. Eventually, with more observations or more knowledge about the underlying rules, our uncertainty might approach zero.~~ At another extreme, for purely random sequences, observing the current state provides no information about

the past ~~or~~ *or* future.

~~The statistical learning literature (Schapiro and Turk-Browne, 2015; Hasson, 2017; Maheu et al., 2022) has tended to focus on manufactured scenarios~~ Sequences generated by stochastic processes fall somewhere between these two extremes~~, whereby participants experience structured (but not fully deterministic) sequences. Often the sequences encountered in this domain are~~ . For Markov processes, ~~whereby (in the simplest framing, as in first-order Markov processes) each element of the sequence~~ where each state is solely dependent on the immediately preceding ~~element, plus some additional noise (McNealy et al., 2006; Cunillera et al., 2009; Furl et al., 2011; Daikoku et al., 2014, 2015; Koelsch et al., 2~~ ~~. In the typical setup, participants are instructed to adjust their behaviors or to form explicit predictions about what will come next, given their observations up to that point. In some studies, participants are also asked to generate inferences about the unobserved past~~ state, Shannon entropy may be used to quantify the uncertainty of the past and future states, given the present state. Cover (1994) showed that, for any stationary process (i.e., ~~what came *earlier*) in a sequence, prior to their first observation (Jones and Pashler, 2007).~~

~~Markov sequences are, by definition, symmetric in time: after learning the "rules"~~ processes in equilibrium), Markov or otherwise, the present state provides equal information (i.e., ~~transition probabilities) that govern how the sequences are generated, observing any part of the sequence provides equal information about its past and future states (Cover, 1994; Bialek et al., 2001; Ellison et al., 2009) . By contrast, real-world experiences are often asymmetric, since our memories provide us with information about our~~ mutual information) about past and future states (also see Bialek et al., 2001; Ellison et al., 2009) . Further, there is some evidence that humans are similarly adept at inferring the most likely previous and next items in sequences governed by stochastic Markov processes (Jones and Pashler, 2007) .

Deterministic, random, and probabilistic sequences (in equilibrium) are all symmetric: the present state of these sequences is equally informative about past ~~experiences but not~~ versus future states. In contrast, our subjective experience in everyday life is that we know more about our own past than our future (e.g., Horwich, 1987). We have memories of our past that we carry with us into the present moment, but we do not have memories of our yet-to-be-experienced future.

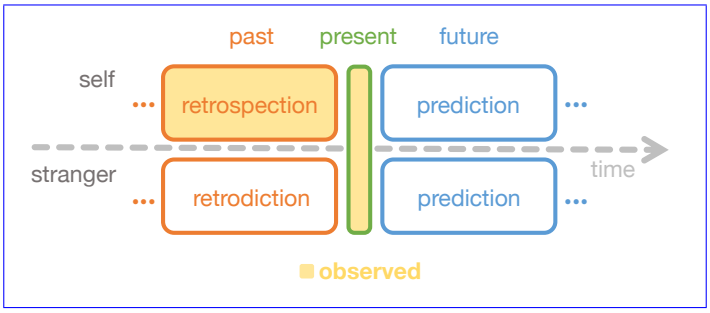

Figure 1: Retrodiction, retrospection, and prediction. In one's own life, one may draw on memory to retrospect (i.e., review or re-evaluate) the past or predict the future. This process is time-asymmetric, since our own past is (typically) observed whereas our future is not. When we make inferences about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

This ~~memory-driven asymmetry in our knowledge is~~ temporal asymmetry imposes an "arrow of time" on our subjective experience, known as the *psychological arrow of time* (e.g., Hawking, 1985). ~~Although the notion of the~~

Although the psychological arrow of time ~~helps to formalize why we know more about our *own*~~ implies that we should be better able to infer our past than our future, how generally does this ~~hold?~~ temporal asymmetry hold? And does the asymmetry hold only for our own experiences (due to our memories), or is the asymmetry a general property 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, ~~is our knowledge~~ are our inferences about the past and future symmetric or asymmetric~~(Fig. ??)~~? For example, imagine that you are meeting a stranger for the first time. ~~Would you~~ At the moment of your meeting, you lack both memories of their past and knowledge about what they might do in the future. After that first encounter with the stranger, would you be able to more accurately or easily form inferences about what had happened in their past ~~versus their future (Tamir and Thornton, 2018; Koster-Hale and Saxe, 2013)~~ (*retrodiction*) or what will happen in their future (*prediction*; Fig. 1)? Or suppose you started watching a movie partway through. Again, ~~given~~ 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?

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.

If we hope to connect prior work on statistical learning tasks that employ sequences generated using relatively simple rules with the sorts of inferences we make in everyday life, several considerations apply. One set of considerations concerns how similar the underlying neurocognitive processes are when we encounter simple sequences in the laboratory versus more complex sequences in our real-world experiences. In other words, if our day-to-day experiences violate the Markov assumption by displaying interactions between temporally separated occurrences or events, does this change anything meaningful in terms of how our brains process, respond to, leverage, or remember those experiences? Another set of considerations concerns how similar the experiences themselves are. For example, to what extent are real-world experiences deterministic versus stochastic? To what extent are real-world experiences structured versus random? And to what extent do real-world experiences satisfy the Markov assumption? Narrative stimuli, such as stories and movies, can provide a useful testbed for exploring several of these questions.

Although narratives are unlikely to be confused with one's own experiences, narratives are most compelling when they mirror some of the structure of real-world experiences. Character behaviors and interactions are often designed in a way that helps the audience connect with or relate to the characters. Events in narratives also unfold in ways that are intended to build rapport or engagement with the audience. This might be accomplished by having events follow a believable structure that is reminiscent of real-world experiences, or by designing the audience's experiences in ways that communicate clear "rules" or "features" that help to immerse the audience in the narrative's universe. The characters in a realistic narrative can also

be written to behave in ways reminiscent of real-world people. These same aspects of narratives that authors use to drive engagement with ~~characters and events~~ events and characters can lead narratives to replicate some core aspects of real-world experiences that are typically lost or overlooked in traditional sequence learning paradigms. Narratives can drive the audience to ~~form a theory of mind of the characters (Tamir and Thornton, 2018), or to~~ build situation models ~~(Ranganath and Ritchey, 2012; Bower et al., 1979)~~ (Radvansky and Copeland, 2006; Zwaan and Radvansky, 1998) of the narrative's universe, or to form a theory of mind of and make predictions about the characters (Tamir and Thornton, 2018; Koster-Hale and Saxe, 2013). Events in narratives may unfold in a consistent or logical way, but they also ~~often incorporate stochastic elements and events that violate the Markov assumption, such as when the consequences of an action or event early on in the narrative are realized much later on.~~

~~If inferences formed for~~ exhibit complex and meaningful interactions across events reminiscent of real-world ~~events and narratives are like inferences for the Markov sequences often studied~~ experiences (but not necessarily the simple sequences traditionally used in the statistical learning literature~~, people should be equally adept at inferring the past and future. Alternatively, perhaps important differences between naturalistic versus Markov sequences might lead to an advantage for the past or future. In the above examples of meeting a stranger or starting a movie partway through, perhaps the *stranger's* (or characters') asymmetric knowledge about their own pasts might influence their observable behaviors~~).

One key difference between simple artificial sequences and more naturalistic (real or narrative) sequences is that naturalistic sequences often incorporate other people. Despite the past and future being equally unknown to ~~*you*~~*the observer* prior to the current moment, other people, and realistic characters in narratives, have their own psychological arrows of time. Specifically, they have memories of their own pasts. Other people's asymmetric knowledge about their ~~own observed pasts might give you an advantage at inferring the unobserved past versus future (Pillemer et al., 2015). Similarly, if a conversation you joined partway through was focused on planning a future event, that might advantage your inferences of the future (versus the past~~*own* pasts and futures might affect their behaviors (e.g., conversations). In ~~general, what sorts of behaviors might reflect other~~

~~individuals' asymmetric knowledge about their past, or their plans for the future? Are there particular types of information that people are especially sensitive to, or that are particularly informative or useful? Could we quantify or characterize those factors in a systematic way? Or might they be too subtle or nuanced to formally define?~~

~~At a high level, these questions all relate to how people form inferences about the pastand future, given their observations. But these questions also often relate to issues of temporal symmetry– i.e., whether~~ turn, this might provide time-asymmetric clues that favor the past (e.g., other people might talk more about thei . If observers leverage these clues from other people's asymmetric knowledge, then observers should also be better at inferring the past ~~is *easier* or *harder* than inferring the future. Presumably, the overall amount of information present will determine something about how *much* may be inferred before or after the present. In other words, a tiny amount of information, or an unreliable observation, and so on, will provide less insight into other moments than a large amount of reliable information. In a sense, this is trivial: more information is "better" if our goal is to form inferences . However, *controlling* for the amount of information, can we learn anything meaningful about inherent biases in either our memory systems (including related systems for forming inferences) or in the sort of information we are likely to encounter? Studying these questions often entails comparing inferences~~ (versus the future) of other people's lives. Alternatively, inferences about other people's lives may be more like inferences about artificial statistical sequences (e.g., perhaps solely relying on statist . If so, then the accuracy of inferences about the past ~~versus the future by characterizing any apparent~~ and the future of others' lives should be approximately equal. We note that the aforementioned authors make no specific claims about temporal symmetries or asymmetries. Rather, we claim that statistical regularities might *imply* symmetry (e.g., if you are on step n of an unfolding schema, this suggests you may have just completed step $n-1$ and that you may next encounter step $n+1$).

We designed a naturalistic paradigm for exposing participants to scenarios where the past and future were equally unobserved. We asked our participants to watch a series of movie segments drawn from a character-driven dramatic television show. Across the conditions and trials in the experiment, ~~we carefully controlled for how much of the past or future of the narrative the~~

~~participants had experienced prior to viewing the target segment. We then asked the participants to guess about what might have happened before the target segment(*retrodiction*), what might happen after the target segment(*prediction*), or to recount what they had observed in the target segmentitself (*recall*)~~participants 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. We used manual annotations and sentence-level natural language processing models to characterize participants' responses. To foreshadow our results, we found that participants were overall better at retrodicting the past than predicting the future. This appeared to be driven by two main factors. First, characters more often referred to past events than future (e.g., planned) events, and this influenced participants' responses. Second, associations and dependencies between temporally adjacent events enabled participants to form estimates about nearby events (e.g., to a just-watched scene or a past or future event referenced in an observed conversation). We also ran a pre-registered replication study to confirm that these findings generalized to another television show and group of participants. Finally, we ran a meta analysis using natural language processing to estimate the prevalence of references to past and future events in hundreds of millions of dialogues drawn from television shows, popular movies, novels, and written and spoken natural conversations. Taken together, our work reveals a temporal asymmetry in how observations of other humans' behaviors inform us about the past versus the future.

Results

Participants in our ~~study~~ main experiment (n = 36) watched segments from two storylines, drawn from the CBS television show ~~,~~ *Why Women Kill*. Each storyline comprised 11 segments (mean duration: 2.05 min; range: 0.97–3.87 min, Table S1). We asked participants to use free-form (typed) text responses to retrodict what had happened prior to a just-watched segment, predict what would happen next, or recall what they had just watched ~~.~~ (Fig. 2, *Task design*). We referred to the to-be-retrodicted, to-be-predicted, or to-be-recalled segment as the ~~target segment~~ *target segment* for

each response. We systematically varied whether participants watched the segments in forward or reverse chronological order, and how many segments ~~preceded or proceeded the target segment~~ they had seen prior to making a response (~~Fig. 2, *Task design*~~see *Task design and procedure*). We also ran a pre-registered replication study with a similar design, where participants (n = 37) watched segments from the Netflix television show *The Chair*, comprising 13 segments (mean duration: 1.97 min; range: 0.58–4.30 min, Table S2).

We asked participants in our main experiment to generate four types of responses after watching each video segment: uncued responses, character-cued responses, updated responses, and recalls (Fig. 2, *Data overview*). To generate *uncued* responses, we asked participants to either retrodict (uncued retrodiction; *u-R*) what happened shortly before or predict (uncued prediction; *u-P*) what happened shortly after the just-watched segment. To generate *character-cued* responses, we asked participants to retrodict (character-cued retrodiction; *c-R*) or predict (character-cued prediction; *c-P*) what came before or after the just-watched segment, but we provided additional information to the participant about which character(s) would be present in the target (to-be-retrodicted or to-be-predicted) segment. We hypothesized that character-cued responses should be more accurate than uncued responses, to the extent that participants incorporate the character information we provided to them into their retrodictions and predictions. To generate updated responses, we asked participants to watch an additional segment that came just prior to or just after the target segment, and then to update their retrodiction (*c-RP*) or prediction (*c-PR*) about the target segment. Results on updated responses are not reported in this paper. Finally, we also asked participants to *recall* what happened in the just-watched segment. We labeled these responses according to which other segments participants had watched prior to the just-watched target. Retrodiction-matched recall (*re(R)*) responses were made during the retrodiction sequences (B1 and B2; Fig. 2), whereas prediction-matched recall (*re(P)*) responses were made during the prediction sequences (A1 and A2~~). Participants' recalls provided us with a benchmark for examining information about the participants' experiences and memories, without asking the participants to explicitly speculate about the unobserved.~~

~~For each~~ ; Fig. 2). Whereas retrodiction and prediction ~~, participants were asked to generate at~~

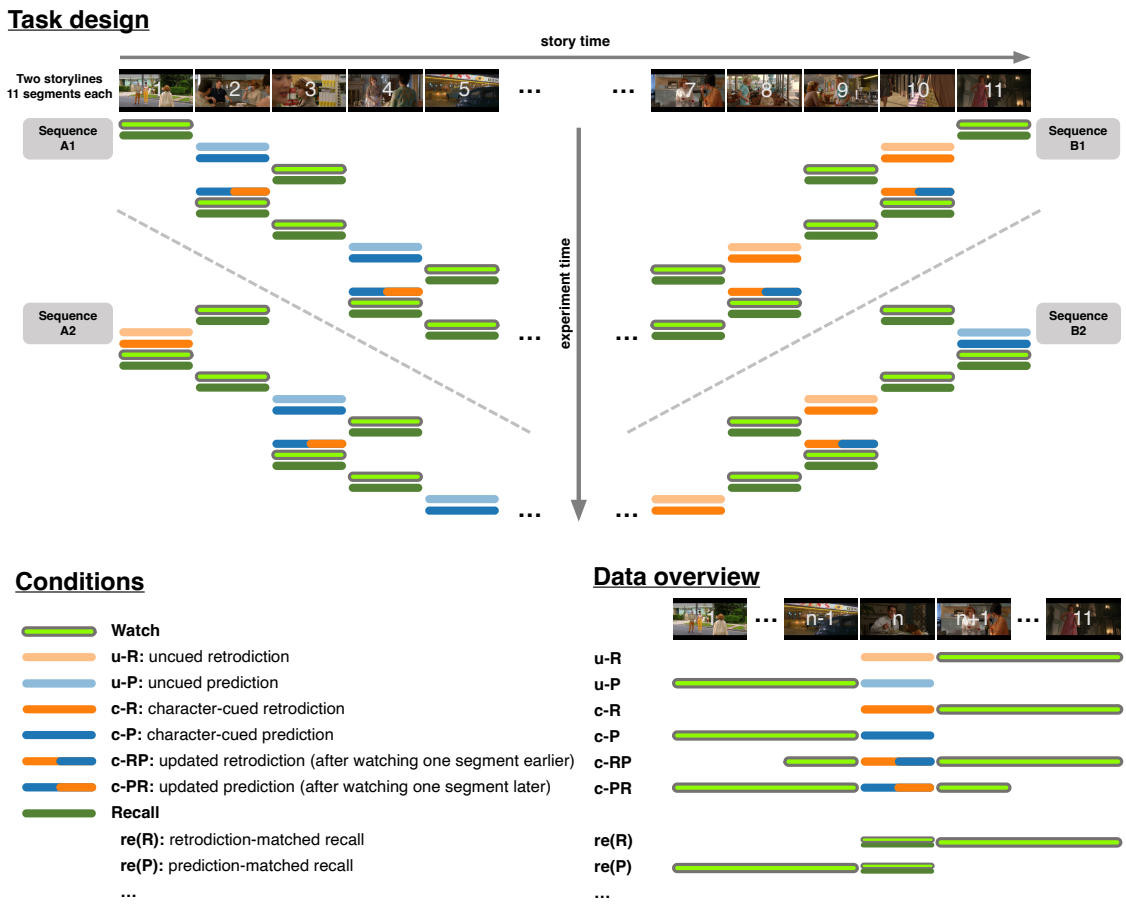

Figure 2: Task overview. Participants in our main experiment watched segments of two storylines from the television series *Why Women Kill*. They made free-form text responses to either retrodict what had happened in the previous segment, predict what would happen in the next segment, or recall what 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 segment orders 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. S1.

~~least one, and not more than three, responses that constituted "the sorts of things the participant wouldexpect to have remembered if they had watched the target segment. " They were asked to generate multiple responses only if those additional responses were (in their judgement) of equal likelihood to occur.On average, participants generated 1.08 responses per prompt; therefore we chose to consider only participants' first ("most probable" or "most important") responses to each prompt. We also discarded a small number (n = 20) of character-cued responses that did not contain references to all cued characters, along with one additional response due to the participant's misunderstanding of the task instructions during that trial. We carried out our analyses on the remaining 2084 retrodiction, prediction, and~~ responses reflect what participants *estimate* they would remember after watching the (inferred) target segment, recall responses provide a benchmark for comparison by measuring what they *actually* remember about the target segment. Our replication experiment (Fig. S1) used a similar design, but did not have participants generate recall responses.

We used two general approaches to assess the quality of participants' responses (see ~~*Methods*~~*Response analyses*, *Text embeddings of participants' responses*, Fig. 3A). One approach entailed manually annotating events in the video and counting the number of matched events in participants' responses. We identified a total of 117 unique events reflected across the 22 video segments in our main experiment (range: 3–9 per segment~~; see~~), and a total of 71 events across the 13 segments in our replication experiment (range 1–16; see ~~*Methods*~~*Video annotation*, ~~Table S1~~Tables S1, S2). We assigned one "point" to each of these video events. We also identified ~~23 additional events~~ a number of additional events (main experiment: 23; replication experiment: 17) in participants' responses that were either summaries of several events or that were partial matches to the manually identified video events. We assigned 0.5 point to each of these additional events. This point system enabled us to compute the numbers and proportions (*hit rates*) of correctly retrodicted, predicted, and recalled events contained in each response. Our second approach entailed using a natural language processing model (Cer et al., 2018) to embed annotations and responses in a 512-dimensional feature space. This approach was designed to capture conceptual overlap between responses that were not necessarily tied to specific events. To quantify this conceptual overlap, we computed the similarities between the embeddings of different sets of responses. Following Heusser et al. (2021), we defined

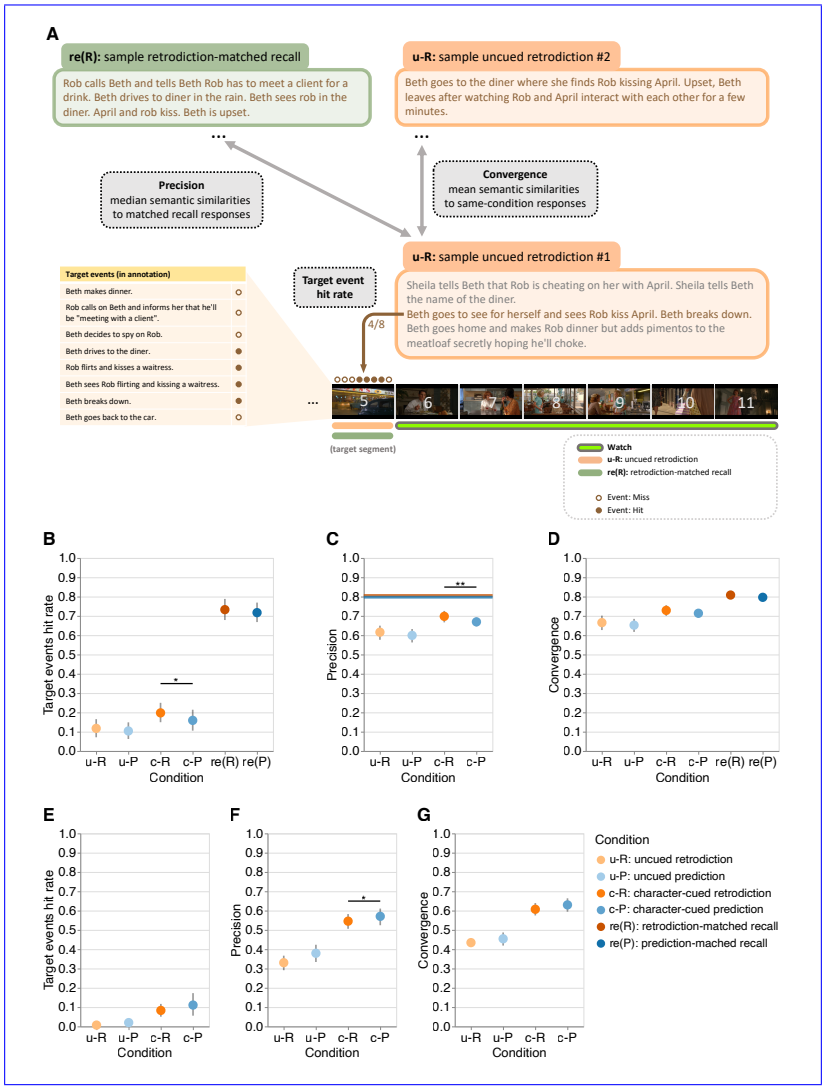

Figure 3: ~~Retrodiction, prediction, and recall performance by experimental condition.~~ Retrodiction, prediction, and recall performance by experimental condition in our main and replication experiments. **A. Methods schematic.** For each retrodiction, prediction, and recall response, we calculated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

the *precision* of ~~each participants' retrodictions or predictions about a given segment~~ participants' responses as the median cosine ~~similarities between~~ similarity between that response's vector and the embedding vectors for all other participants' recalls of the target segment (main experiment), or the similarity between that response's vector and the embedding vector for an online plot synopsis (obtained via Screen Spy; `www.screenspy.com/the-chair-season-1-episode-1`) of the ~~embeddings of (a)the participant's retrodiction or prediction response for the given segment and (b) each *other* participant's recalls of the same segment~~ target segment (replication experiment). 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 ~~given~~ 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*~~*Statistical analysis*).

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 in our main experiment (odds ratio (OR): 2.65, $Z = 4.24$, $p < 0.001$, 95% confidence interval (CI): 1.69 to 4.16; Fig. 3B). These character-cued responses were also more precise ($b = 0.13$, $t(18.1) = 9.43$, $p < 0.001$, CI: 0.10 to 0.16; Fig. 3C) and convergent across individuals ($b = 0.11$, $t(18.6) = 6.21$, $p < 0.001$, CI: 0.07 to 0.15; Fig. 3D). Relative to character-cued responses, participants' recalls showed higher target event hit rates (OR = 21.83, $Z = 10.61$, $p < 0.001$, CI: 12.35 to 38.59) and were more convergence 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 (target event hit rates of character-cued vs. uncued responses: OR: 18.63, $Z = 4.26$, $p < 0.001$, CI: 4.85 to 71.58, Fig. 3E; precisions of character-cued vs. uncued responses: $b = 0.26$, $t(11.70) = 9.87$, $p < 0.001$, CI: 0.20 to 0.31, Fig. 3F; convergence of character-cued vs. uncued responses: $b = 0.25$, $t(11.98) = 8.93$, $p < 0.001$, CI: 0.19 to 0.31, Fig. 3G).

Next we carried out a series of analyses specifically aimed at characterizing temporal direction effects— i.e, the relative quality of retrodictions versus predictions across different types of responses. We hoped that these analyses might provide insights into our central question about whether ~~the present is equally informative~~ inferences about the past and future are equally accurate. Across both uncued and character-cued responses in our main experiment (Fig. 2), retrodictions had numerically higher hit rates than predictions (Fig. 3B). However, these differences were only statistically reliable for character-cued responses (uncued responses: OR = 1.17, $Z = 0.35$, $p = 0.73$, CI: 0.47 to 2.92; character-cued responses: OR = 1.93, $Z = 2.15$, $p = 0.03$, CI: 1.06 to 3.52). We observed a similar pattern of results for the precisions of participants' responses (Fig. 3C). Specifically, their responses tended to be numerically more precise for retrodictions versus predictions, but the differences were only statistically reliable for character-cued responses (uncued responses: $b = 0.03$, $t(20.9) = 1.09$, $p = 0.29$, CI: -0.03 to 0.10; character-cued responses: $b = 0.06$, $t(20.8) = 3.01$, $p = 0.007$, CI: 0.02 to 0.11). We also consistently 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). ~~Because the retrodiction versus prediction performance differences we observed were only statistically reliable when participants were cued with the target segments' characters, this suggests that information about the unobserved past versus the unobserved future may differently affect retrodictions versus predictions. 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 (~~In our replication experiment, as in our main experiment, most of these differences were not statistically reliable (target event hit rates for uncued responses: OR = 0.11, $Z = -1.92$, $p = 0.05$, CI: 0.01 to 1.04; target event hit ~~rate~~rates for character-cued responses: OR = ~~1.12, $Z = 1.07$, $p = 0.29$~~1.42, $Z = 0.62$, $p = 0.53$, CI: ~~0.91 to 1.39; convergence: $b = 0.03$, $t(19.3) = 1.89$, $p = 0.07$~~0.47 to 4.23, Fig. 3E; precision for uncued response: $b = -0.06$, $t(15.86) = -1.85$, $p = 0.08$, CI: -0.12 to 0.01; precision for character-cued responses: $b = -0.04$, $t(25.02) = -2.28$, $p = 0.03$, CI: -0.08 to 0.00, Fig. 3F; convergence for uncued responses: $b = -0.03$, $t(12.15) = -0.55$, $p = 0.59$, CI: -0.13 to 0.07~~).~~; convergence for character-cued responses: $b = -0.05$, $t(13.68) = -1.78$, $p = 0.10$, CI: -0.11 to 0.01, Fig. 3G). Taken together, our results from both experiments suggest that participants' inferences about the immediate past and future do *not* show reliable asymmetries.

The above analyses were focused solely on the target segment (i.e., retrodiction of segment ~~$i - 1$~~ n after watching segments ~~i...11~~$(n + 1)$...N, or prediction of segment ~~$i + 1$~~ n after watching segments ~~1...i~~1...$(n - 1)$). 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 in our main experiment's stimuli 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 applied the same procedure to our replication experiment's stimuli and identified 66 implicit offscreen events, plus two additional partial events and one additional summary event. We defined the just-watched segment as having a *lag* of 0.

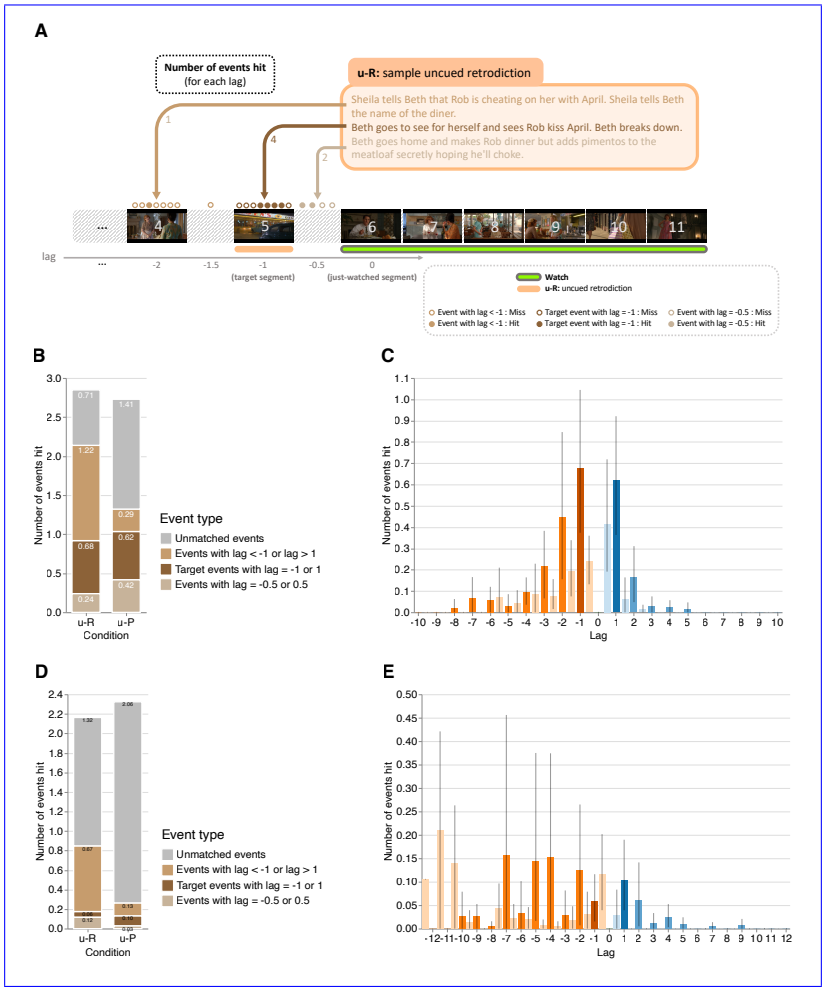

Figure 4: Retrodictions and predictions of temporally near and distant events. A. Illustration of annotation approach. For each uncued retrodiction and prediction response, we calculated the number of (retrodicted or predicted) events as a function of temporal distance from the target segment, or *lag*. 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.**~~ **B. Number of events hit in participants' uncued retrodictions and predictions for each event type (main experiment).** 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.**~~ **C. Number of events hit as a function of temporal distance (main experiment).** 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). **D. Number of events hit in participants' uncued retrodictions and predictions for each event type (replication experiment).** Same format as Panel B. **E. Number of events hit as a function of temporal distance (replication experiment).** Same format as Panel C. 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).

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, we asked how many events from segment $i + lag$ they retrodicted or predicted, on average, given that they were aiming to retrodict or predict events at lags of ±1. We also counted the numbers of *unmatched* events in participants' responses that did not correspond to any events in the relevant segments of the narrative. We focused specifically on *uncued* retrodictions and predictions, which we hypothesized would provide the cleanest characterizations of participants' initial estimates of the unobserved past and future (i.e., without potential biases introduced by additional character information, as in the character-cued responses). ~~The~~ For participants in our main experiment, the numbers of uncued retrodicted and predicted target (lag = ~~±1~~±1) events were not reliably different (~~OR~~ Ratio = 0.92, $Z = -0.15$, $p = 0.88$, CI: 0.30 to 2.84; Fig. 4B). In other words, uncued retrodictions and predictions over short timescales did not exhibit reliable asymmetries. This "null result" also held in our replication experiment (Ratio = 0.44, $Z = -1.38$, $p = 0.17$, CI: 0.14 to 1.41; Fig. 4D). However, when retrodicting, participants in both experiments mentioned events from the distant past (lag < -1) more often than participants predicted events from the distant future (lag > 1; ~~OR~~ main experiment: Ratio = 9.10, $Z = 3.80$, $p < 0.001$, CI: 2.92 to 28.39; Fig. 4B, C; replication experiment:Ratio = 7.98, $Z = 5.50$, $p < 0.001$, CI: 3.81 to 16.74; Fig. 4D, E; for results from the character-cued conditions, see ~~Fig.~~ Figs. S4, S5). Despite this asymmetry in the ~~accuracies~~ accuracy of participants' long-range retrodictions versus predictions, there were no reliable differences in the ~~*numbers*~~*total* numbers of uncued retrodicted versus predicted events (~~across all lags; OR~~ main experiment: Ratio = 1.05, $Z = 0.75$, $p = 0.45$, CI: 0.93 to 1.18~~). Nor did we~~ ; replication experiment: Ratio = 0.93, $Z = -0.67$, $p = 0.50$, CI: 0.75 to 1.15). We did not

find any reliable differences in the numbers of offscreen events immediately before or after the just-watched segment in our main experiment (*lag* = ±0.5; ~~OR~~ main experiment: Ratio = 0.75, $Z = -0.36$, $p = 0.72$, CI: 0.15 to 3.59), but participants in our replication experiment responded with more prior (versus future) immediate offscreen events (Ratio = 26.46, $Z = 2.45$, $p = 0.01$, CI: 1.93 to 362.29). The apparent discrepancy between participants' asymmetric accuracy but symmetric (overall) event counts was due to participants' tendencies to reference "unmatched" events (i.e., events that did not correspond to any explicit or implicit event in the story) ~~more in their predictions than retrodictions (OR~~ less in their retrodictions than predictions (main experiment: Ratio = 0.36, $Z = -4.53$, $p < 0.001$, CI: 0.23 to 0.56~~). We confirmed that the~~ ; replication experiment: Ratio = 0.66, $Z = -3.26$, $p = 0.001$, CI: 0.51 to 0.85). This retrodiction advantage held when controlling for absolute lag ~~(OR~~ in our main experiment (Ratio = 34.31, $Z = 3.28$, $p = 0.001$, CI: 4.16 to 283.20), although it did not hold up in our replication experiment (Ratio = > 9999, $Z = 0.00$, $p > 0.99$), as participants in the replication experiment almost *never* referenced offscreen events in their predictions. The retrodiction advantage also held for onscreen events alone ~~(OR~~ in our main experiment (Ratio = 47.54, $Z = 3.74$, $p < 0.001$, CI: 6.27 to 360.60), marginally in our replication experiment (Ratio = 3.86, $Z = 1.86$, $p = 0.06$, CI: 0.93 to 15.98), and marginally for offscreen events alone ~~(OR~~ in our main experiment (main experiment: Ratio = 24.76, $Z = 1.71$, $p = 0.09$, CI: 0.63 to 975.27~~). ~~; replication experiment: Ratio > 9999, $Z = 0.00$, $p > 0.99$). Again, the lack of "effect" in our replication experiment is due to the lack of *any* offscreen event responses in participants' predictions. Taken together, these analyses show that (in generating uncued responses) participants tend to reach "further" into the unobserved past, and with greater accuracy, than the unobserved future.

~~**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).

What might be driving participants to retrodict further and more accurately into the unobserved past, compared with their predictions of the unobserved future? By inspecting the video content, we noticed that characters in the television show frequently referenced both past events and (planned or predicted) future events in their spoken conversations, which might provide clues about past and future events. We wondered whether the characters' references might show temporal asymmetries that might explain participants' behaviorsparticipants' responses might be influenced by characters' conversational references. Across all of the characters' conversations, and across all of the video segments from our main experiment, we manually identified a total of 82 references to past or future events (i.e., that occurred onscreen or offscreen before or after the events depicted in the current segment; FigFigs. 5A,A S6A, also see *Reference coding*). Characters in our main experiment's stimulus tended to reference the past (52 references) more than the future (30 references), consistent with previous work (Demiray et al., 2018). References to the past were also skewed to more temporally distant events compared with references to the future (Figs. 5B, S6B). These asymmetries also held for characters in the replication experiment's stimulus

(46 past references versus 7 future references, Figs. S8A, S7B). These observations indicate that the characters in the ~~stimulus display a preference~~ stimuli display a "preference" for the past (versus future) in their conversations. Might this asymmetry be driving the asymmetries in participants' retrodictions versus predictions?

Controlling for temporal distance (lag), past and future events that story characters referenced in their conversations were associated with higher hit rates than unreferenced events in our main experiment (uncued retrodiction: OR = 12.70, $Z = 10.94$, $p < 0.001$, CI: 8.06 to 20.03; uncued prediction: OR = 8.29, $Z = 6.83$, $p < 0.001$, CI: 4.52 to 15.20; Fig. 5E). ~~This indicates that~~ In our replication study this result held for past events (uncued retrodiction: OR = 5.57, $Z = 5.88$, $p < 0.001$, CI: 3.14 to 9.89) but not for future events (uncued prediction: OR = 1.54, $Z = 0.22$, $p = 0.83$, CI: 0.03 to 73.36; Fig. S8D). The failure to replicate the "prediction" result appeared to follow from the fact that references to future events in characters' conversations were very rare in our replication experiment's stimulus. These findings suggest that participants' responses are at least partially influenced by the characters' conversations. To estimate the contributions of characters' references on hit rates, we computed the difference in hit rates between all events (which comprised both referenced and unreferenced events) and unreferenced events, as a function of lag. These differences exhibited a temporal asymmetry in favor of retrodiction (~~Fig~~Figs. 5C, S8B). 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 in both our main experiment (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) ~~.~~ and replication experiment (OR = 3.67, $Z = 2.61$, $p = 0.01$, CI: 1.38 to 9.74; Fig. S8C). When we further separated the unreferenced events into onscreen events and offscreen events, we found that these asymmetries held only for the onscreen events in our main experiment (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), and only for offscreen events in our replication experiment (onscreen: OR = 0.97, $Z = -0.06$, $p = 0.95$, CI: 0.37 to 2.57; offscreen: OR = 13.88, $Z = 3.06$, $p = 0.002$, CI: 2.58 to 7.48). Taken together, these analyses suggest that asymmetries

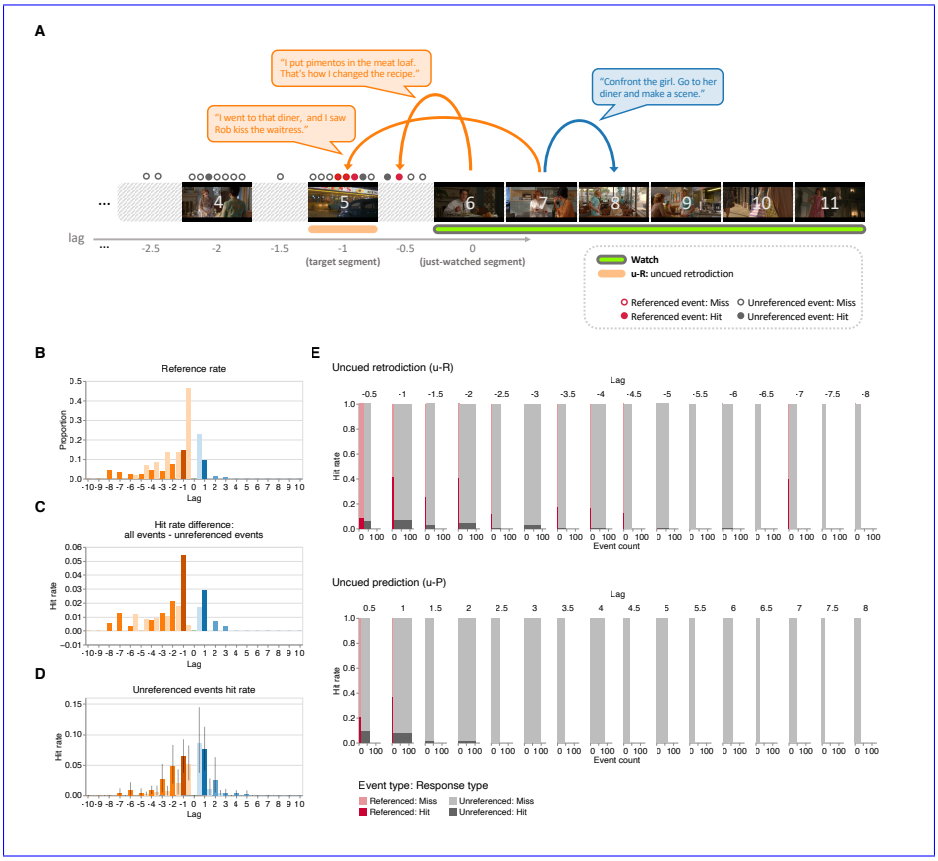

Figure 5: **Characters' references drive participants' retrodiction and prediction performance (main experiment). A. Illustration of annotation approach.** We manually annotated references to events in past or future segments in characters' spoken conversations. We matched each such reference with its 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 or future 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). Intuitively, the widths of the rectangles at each lag denote the total number of events at each possible lag. The darker shading denotes the proportions of events that participants retrodicted or predicted, and the lighter shading denotes the proportions of events that participants "missed" in their responses. For an analogous presentation of results from the replication experiment, see Figure S8.

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 ~~and~~ or temporal dependencies between temporally ~~adjacent events , might characters' references to specific events also boostparticipants' estimates of other~~ nearby events in the television shows 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 ~~temporally *adjacent*~~*nearby in time* to events that characters referred to in their conversations, in addition to the referenced events themselves (Fig. 6A)~~?~~ .

Because characters tended to refer to past events more often than future events, the proportions of unreferenced events that were adjacent to referenced events should show a similar temporal asymmetry in favor of the past. We ~~confirmed~~ tested this intuition by computing the proportions of unreferenced events in the stimulus that were temporally adjacent to past or future events referenced by the characters during a given segment. Here we defined *temporally adjacent* as any event within an absolute lag of one relative to a referenced onscreen event, or within an absolute lag of 0.5 to a referenced offscreen event. We also defined *remaining* events as unreferenced events that were not temporally adjacent to any referenced events. ~~As shown in Figure 6B,~~ In our main experiment we observed higher proportions of unreferenced past than future events

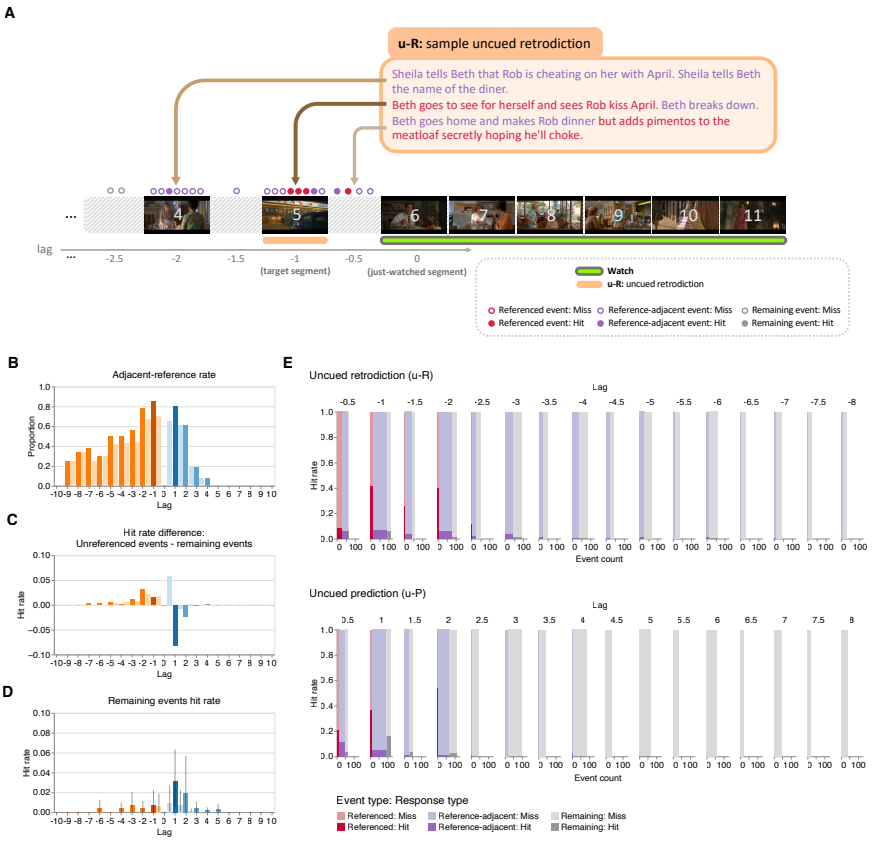

Figure 6: ~~**Reference-adjacent events are associated with higher hit rates.**~~ **Reference-adjacent events are associated with higher hit rates (main experiment). A. Illustration of annotation approach.** We extended the annotation procedure depicted in Figure 5A to also label unreferenced events 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 Figs. S9 and S10.

that were temporally adjacent to referenced events ~~.~~ (Fig. 6B). Further, these reference-adjacent events had higher hit rates than remaining events after controlling for absolute lag (uncued retrodiction: OR = 7.15, Z = 2.40, p = 0.02, CI: 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 the contributions of reference adjacency on hit rates, we computed the ~~difference~~ differences in hit rates between unreferenced events (which comprised both reference-adjacent and remaining events) and remaining events, as a function of lag. These differences exhibited a temporal asymmetry in favor of retrodiction ~~.~~ (Fig. 6C). This suggests that reference-adjacent events also contribute to participants' retrodiction advantage. This reference-adjacency effect did not hold in our replication experiment (uncued retrodiction: OR = 6.46, Z = 1.58, p = 0.13, CI: 0.64 to 65.04; uncued prediction: OR = 0.002, Z = 0.007, p = 0.99, CI: < 0.001 to > 9999; Fig. S9D, B). Upon further examination of the stimulus we used in our replication experiment, along with participants' responses, we noticed that the television episode appears to comprise several interleaved storylines (Fig. S10A). This meant that what we had originally labeled as "reference-adjacent" events (based solely on the temporal order in the *episode*) did not necessarily correspond to chronological order in the *story*. For example, if (across successive segments) the narrative focuses on character A at time t in segment n, and on character B at time t in segment $n + 1$, then we reasoned that watching segment n might not provide much insight into what would happen in segment $n + 1$. However, watching segment n *could* provide clues about what would happen to character A at time $t + 1$, which might have been shown later on in the episode. When we "corrected" the reference-adjacency labels in the replication experiment stimulus to correspond to individual storylines, rather than solely with respect to the episode segment orders, we recovered the reference adjacent effect for uncued retrodiction (OR = 7.55, Z = 2.93, p = 0.003, CI: 1.95 to 29.20; Fig. S10E, C). We did not find a significant reference adjacent effect in uncued prediction (OR = 1176.66, Z = 0.04, p = 0.97, CI: < 0.001 to > 9999), again likely due to the limited number of future references in the narrative. Remaining events did *not* exhibit a reliable temporal asymmetry (main experiment: OR = 0.75, Z = 0.33, p = 0.74, CI: 0.14 to 4.08; Fig. 6D; replication experiment: OR = 889.48, Z = 0.03, p = 0.97, CI: < 0.001 to > 9999; Fig. S10D), suggesting that, after accounting for temporal adjacency, character's references to past and future events can explain participants'

retrodiction advantage.

The preceding analyses show that when characters reference past or future events, those referenced events, and other events that are temporally adjacent to the referenced events, are more likely to be retrodicted and predicted. In other words, referring to a past or future event in conversation 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 ,~~" or an "other event" (i.e., not referring or referenced; ~~Fig~~Figs. 7A, B). We limited our analysis to references to onscreen (explicit) events. Consistent with our analysis of ~~character's references to other events (Fig~~the proportions of referenced events (Figs. 5B~~),~~ , S8A), the proportions of *referring* events exhibited a *forward* temporal asymmetry (Fig. 7C, S11A). 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. ~~5~~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). ~~This~~ In our replication experiment, because there were very few referenced events during prediction, which also resulted in limited referring events during retrodiction, we had insufficient data to reliably compare between referenced, referring, and other events (all ps > 0.99; Fig. S11). Taken together, this indicates that only referenced events received a hit rate boost (relative to ~~referring~~ other events), suggesting that the retrodictive and predictive benefits of references are directed (i.e., asymmetric).

~~Discussion~~

~~We asked participants~~

The above analyses show that participants leveraged characters' references to make inferences

Figure 7: ~~**Referenced events are associated with higher hit rates, but referring events are not.**~~ **Referenced events are associated with higher hit rates, but referring events are not (main experiment). A. Illustration of annotation approach.** We extended the annotation procedure depicted in Figure 5A to also label which events in our main experiment's stimuli *contained* references to events in other segments. **B. Referenced versus referring events.** During event *i*, when a character makes a reference to another event (*j*), we define *i* as the *referring* event and *j* as the *referenced* event. **C. Referring rate as a function of lag.** Across all possible just-watched segments (lag 0), the bar heights denote the across-segment mean proportions of events containing references to events in other past ~~(orange, negative lags)~~ or future ~~(blue, positive lags)~~ segments in our main experiment's stimuli. The bar colors are described in the Figure 4 caption. **D. Hit rates and counts of referenced, referring, and other events.** As a function of temporal distance to the just-watched segment, the sub-panels display the numbers (*x*-axes) and hit rates (*y*-axes) of referenced (red), referring (green), and other (gray) events that participants hit (darker shading) or missed (lighter shading) in their uncued retrodictions (top sub-panel) and uncued predictions (bottom sub-panel). For a display of analogous results from our replication experiment see Figure S11.

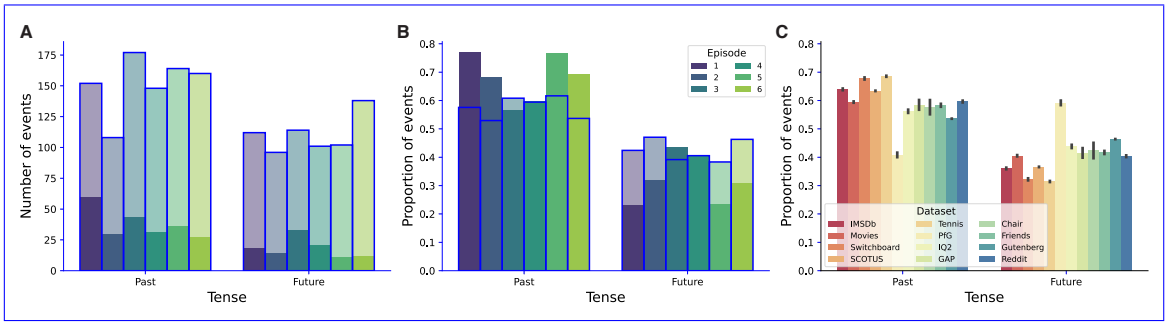

Figure 8: Meta analysis. We used natural language processing to automatically identify references to past or future events across a variety of sources. **A. Numbers of past and future events in *The Chair*, Season 1, Episodes 1–6.** The bar heights indicate the raw numbers of manually identified (lighter 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 S3. Error bars denote bootstrap-estimated 95% confidence intervals.

about the past and the future, and the retrodiction advantage could be attributed to the fact 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. 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 narratives created for entertainment purposes tend to have biases towards the past in order to keep the stories engaging and unpredictable. To better understand temporal biases in conversations, we carried out a meta analysis using extracted conversation data from several large datasets, comprising over 17 million documents. The data comprised transcripts from television shows and popular films, novels, and spoken and written utterances from natural conversations. A summary of the data we analyzed may be found in Table S3. As summarized in Figure 8, we used natural language processing to identify references to past or future events in each conversation (also see *Meta analysis*).

To validate our basic approach, we manually identified references to past and future events, across six episodes of the television show *The Chair* (the first episode was used as the stimulus in our replication study). We then compared the numbers (Fig. 8A) and proportions (Fig. 8B) of

automatically and manually identified references. In general, our automated tagging procedure tended to over-count the numbers of references. From manually "spot checking" hundreds of example tags, we noticed that our automated tagging procedure often counts the "same" references multiple times. Specifically, the manually generated tags sought to identify references to specific events that occurred or were implied to occur in other parts of the narrative. In contrast, as a heuristic, we designed the automated tagging procedure to identify uses of the past or future *tense* as a proxy for references to past or future *events*. Individual conversations often contains multiple references to a given (past or future) event. Whereas the manually generated tags counted these as "single" references, our automated tagging procedure had no means of differentiating between several references to the same event versus the same number of references to different events. This leads the automated tagging procedure to overestimate the numbers of distinct events being referenced. Nevertheless, this discrepancy did not appear to bias the balance of the overall *proportions* of past or future references.

In all, across all of the datasets we examined in our meta analysis, we identified a total of 36,008,500 references to past or future events. A total of 19,464,741 (54.06%) of these were references to past events, and the remaining 16,543,759 (45.94%) were references to future events. We also computed the average proportions of references to past and future events across documents within each individual dataset. Across the 12 datasets we examined (Fig. 8, Tab. S3), there were significantly more references to the past than the to the future (mean ± standard deviation proportion of references to past events: 58.99% ± 7.28%; $t(11) = 4.28, p = 0.0013$). We used the numbers of past references divided by the numbers of future references to quantify the effect size of temporal biases (see *Meta analysis*). Specifically, effect sizes greater than 1 reflect a bias towards the past, whereas effect sizes of less than 1 reflect a bias towards the future. Across all 12 datasets, we observed an average effect size of 1.45 (standard error: 0.40); this indicates that references to the past are 1.45 times more prevalent than references to the future. This bias towards the past also held for each dataset individually (range of effect sizes: 1.16 – 2.18) except for one dataset, "Persuasion for Good," which comprised natural conversations between pairs of Amazon Mechanical Turk workers wherein one participant tried to convince the other participant to donate

to a charity in the future. In that dataset, references to the future were significantly more common than references to the past (effect size: 0.68 ± 0.10; CI: 0.59 to ~~watch sequences of movie segments from a character-driven television drama. Across trials and participants, we controlled for how many segments preceding or proceeding the target segment the participants had seen, prior to watching the target segment. Then we asked participants~~ 0.65). This latter example provided a nice sanity check for verifying that our general approach was not itself biased in favor of the past, e.g., even in conversations that were actually biased towards the future. All of these effect sizes were significant at the $p < 0.001$ level or lower, except for the dataset containing transcripts of the six episodes of Season 1 of "The Chair" (average effect size: 1.50 ± 0.20; CI: 0.43 to ~~either~~ 4.39; $p = 0.60$), which contained relatively few observations (therefore our confidence interval for that dataset was particularly large). Taken together, the results from our meta analysis indicate that people tend to refer to the past more than they refer to the future, across a wide variety of situations (including in both fictional and real conversations). Although (as in the Persuasion for Good dataset) there may be specific exceptions to this bias, it seems that a bias in favor of the past is a common element of many (and perhaps even *most*) human conversations.

Discussion

We asked participants in our main experiment to watch sequences of movie segments from a character-driven television drama and then either retrodict what had happened prior to a just-watched segment, predict what would happen next, ~~retrodict what had happened before, or recount what had happened in the just-watched target segment~~or recall what they had just watched. We found that participants tended to more accurately and more readily retrodict the unobserved past than predict the unobserved future. We traced this temporal asymmetry to (a) characters~~'~~' tendencies to refer to past events more than future events in their ongoing conversations, and (b) associations between temporally proximal events (Fig. 9). ~~Our findings provide a number of important insights into how we make~~ Essentially, associations between temporally proximal events serve to enhance asymmetries in inferences driven by conversational references

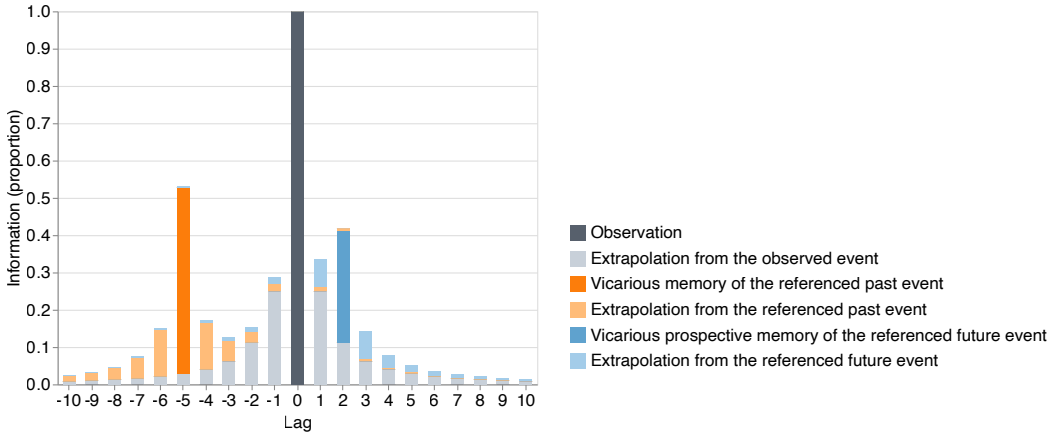

Figure 9: ~~How much information about the past and future can be extracted by observing the present?~~ **How much information about the past and future can be inferred by observing the present?** By definition, let us say that the present moment (lag 0) contains all information about itself (dark gray). 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.

(light orange and blue bars in Fig. 9). Our findings show that other peoples' psychological arrows of time can affect external observers' inferences about the unobserved past and future~~, and how we can study and quantify these processes.~~ . We confirmed our main behavioral findings in a pre-registered replication study. We also carried out a meta analysis of tens of millions of utterances from television shows, movies, novels, and natural spoken and written conversations. We found that the tendency to refer more often to the past than the future appears to be a widespread characteristic of human conversation.

~~In free recall of random word lists, when the participant recalls the word studied at position x, they are likely to next recall the word studied at positions $x \pm 1$. This phenomenon is termed the *contiguity effect* (Kahana, 1996). The contiguity effect has a well-characterized forward asymmetry (Kahana et al., 2022) , whereby the probability of next recalling the word from position $x + 1$ is more likely than recalling~~

~~the word from position $x-1$. This forward asymmetry suggests that we move through our memories with greater ease in the forward (versus reverse) temporal direction. Although our memory systems likely play a role in retrodiction and prediction (Momennejad and Howard, 2018; Barron et al., 2020; Ch , it is important to draw a distinction between our current paradigm and the circumstances leading to the forward asymmetry in the free recall contiguity effect. In free recall, for example, relative to the moment the word from position x was studied, the response period is nearer in timeto the study of word $x+1$ than to $x-1$. In our paradigm, however, the past and future are equidistant from the present moment~~There exists a fundamental knowledge asymmetry such that we know more about our own past than the future, since we remember our past but not our future. A number of prior studies have examined other temporal biases, such as how much people focus on the past, present, and future in their spontaneous thoughts (Grant and Walsh, 2016; Song and Wang, 2012; Shipp and Aeon, 2019) , everyday conversations (Demiray et al., 2018), and social media messages (Park et al., 2017). Several of these studies found that, on average, people's spontaneous *internal* thoughts tend to be more future-oriented than past-oriented (Grant and Walsh, 2016; Song and Wang, 2012). In contrast, people's external *communications* tend to focus more on the past (Park et al., 2017; Demiray et al., 2018) .

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 e . 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. 9; 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 more insight into the speaker's unobserved future than their unobserved past. However,

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

~~Several prior studies have examined retrodiction and prediction in statistical learning paradigms that use Markov sequences as stimuli. In these paradigms, both infants (Tummeltshammer et al., 2016) and adults (Jones and Pashler, 2007)show a (numerical) prediction advantage over retrodiction. During paired associates learning, when A–B pairs of items are presented simultaneously, participants generally show similar forward (generating B when cued with A) and backward (generating A when cued with B) performance (Asch and Ebenholtz, 1962; Kahana, 2002). Taken together, these classic studies do not show the advantage for retrodiction over prediction that we observe here. What accounts for these apparent discrepancies? Why do people display a forward temporal asymmetry in free recall and for Markov sequences, temporal *symmetry* in paired associates learning, but a backward asymmetry in our task? Our worksuggests that three main factors determine how readily participants infer one event from another in more complex~~ In typical statistical sequences used in laboratory studies, there is no temporal asymmetry, either theoretically (Cover, 1994; Bialek et al., 2001; Ellison et al., 2009), or empirically (Jones and Pashler, 2007) . What makes narratives and real-world event sequences time-asymmetric? Of course there are many superficial differences between simple laboratory-manufactured sequences and real-world experiences. As one example, real-world experiences often involve other people who have their own memories and goals. Some recent work (e.g., Tamir and Mitchell, 2013; Meyer et al., 2019) suggests that people might gain insights into other people using "mental simulations" of how they might respond in particular situations (e.g., ~~"naturalistic") circumstances. The first factor is~~

~~that nearby events are associated. This may follow from the notion that events in narratives, as in real-world experiences , tend to change gradually. In other words, what is happening~~ in the future), or of which sorts of prior experiences might have led someone to behave a particular way in the present~~moment (where you are located, what you are doing, who you are with, what you are thinking or feeling, etc.) will, in general, be similar to what was happening a moment ago, and also to what will happen in the next moment. These similarities are roughly symmetric: the past and present are (on average) as similar as the present and future, after controlling for temporal distance. In our study, this symmetry provides a balanced forward *and* backward advantage for events that are nearby in time to the present moment (Fig. 9, gray). The second factor is that, when characters in the narrative refer to other (temporally distant) moments, participants incorporate those references into their retrodictions and predictions (Fig. 5; dark orange and blue in Fig. 9). Other work has tended to focus on how memories can be shared across people through conversational references to *prior* experiences (Hirst and Echterhoff, 2012; Mahr and Csibra, 2018; Dessalles, 2007) . Our study extends this notion by showing that "memory sharing" need not be limited solely to past events. Rather, conversations can also provide information about the *future*, in the form of shared prospective memories . Presumably due to inferred associations between temporally adjacent events , events that~~ . But at a higher level, are our subjective experiences essentially more complicated versions of laboratory-manufactured sequences? Or are there fundamental differences? One possibility is that real-life event sequences are not stationary (i.e., not in equilibrium; Cover, 1994) . For example, real-life events might start from a special initial condition (Albert, 2000; Feynman, 1965; Cover, 1994) and proceed through a series of transitions from more-ordered to less-ordered states, thus exhibiting an arrow time. When we retrodict, it is possible that we only consider possible past events that are compatible with the highly-ordered special initial state (Carroll, 2010, 2016). For example, when we see a broken egg we might infer that the egg had been intact at some point in the past. But it would be difficult to guess at what states or forms the broken egg might take in the future (Carroll, 2010, 2016) . In other words, the procession from order to disorder might result in better retrodiction performance compared with that of (implicitly less-restricted) prediction tasks. The special initial state might also explain why we remember the past, but not the future. Some recent work suggests

that the psychological arrow of time might be explained by a related concept in the statistical physics literature, termed the "thermodynamic" arrow of time (Mlodinow and Brun, 2014; Rovelli, 2022). However, the relation between the thermodynamic and psychological arrows of time is still under debate (Gołosz, 2021; Hemmo and Shenker, 2019).

Beyond forming inferences about unobserved past and future events, our work also relates to prior studies of how people perceive time (Block and Gruber, 2014; Howard, 2018; Eagleman, 2008; Ivry and Schlerf, 200 , and how we "move" through time in our memories of our past experiences (Manning, 2021; Manning et al., 2011; Howa or in our imagined (past or future) experiences (Schacter, 2012; Josselyn and Tonegawa, 2020; Schacter et al., 1998; Mome . For example, a well-studied phenomenon in the episodic memory literature concerns how remembering a given event cues our memories of other events that we experienced nearby in time~~to a referenced eventalso receive a boost (Fig. 6, light orange and blue in Fig. 9) . We expect that these first two factors will hold across different narratives and real-world experiences, as associations between temporally proximal items and experiences have been widely documented (for review, see Kahana et al., 2022; M . The third factor is that characters in the narrative used in our study tend to refer to past events more than future events (Fig. 5). This temporal asymmetry in the characters' conversations is what~~ (i.e., the *contiguity effect*; Kahana, 1996). Across a large number of studies there appears to be a nearly universal tendency for people to move *forwards* in time in their memories, whereby recalling an "event" (e.g., a word on a previously studied list) is about twice as likely to be followed by recalling the event that immediately followed as compared with the event immediately preceding the just-recalled event (Healey and Kahana, 2014). Superficially our current study appears to ~~drive the advantage for retrodiction over prediction in our study. At least one prior study suggests that people are also more likely to refer to past (versus future) events in natural conversations (Demiray et al., 2018). To the extent that this holds in general across narratives and experiences, we expect that the retrodiction advantage we observe here will apply in general. However, we also hypothesize that in scenarios where people discuss the future *more* than the past, participants should show a *prediction* advantage. Similarly, when the past and future are balanced in conversations, we hypothesize that the asymmetry should disappear.~~report the *opposite* pattern, whereby participants display a *backwards* temporal bias. However, the two sets of findings may be

reconciled when one considers the frame of reference (and current mental context; e.g., Howard and Kahana, 2002) of the participant at the moment they make their response. In our study, participants observe an event in the present, and they make guesses about what happened in the unobserved past or future, relative to the just-observed event. (Our findings imply that participants are more facile at moving backwards in time than forwards in time, relative to "now.") In contrast, the classic contiguity effect in episodic memory studies refers to how people move through time relative to a just *remembered* event. The forward asymmetry in the contiguity effect follows from the notion that the moment of remembering has greater contextual overlap with events *after* the remembered event from the past, including the moment of remembering, than events that happened before it (for review also see Manning et al., 2015; Manning, 2020). In other words, our current frame of reference appears to exhibit a sort of "pull" on our thoughts, such that thoughts about recent experiences still lingering in our minds drag us towards the recent past, but after thinking about the more distant past we are dragged (relatively) forward in time back to "now."

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

Methods

Participants

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

Replication experiment. A total of 37 participants (21 female, mean age 22.24 years, range 19–30 years) were recruited from the Dartmouth College community for our pre-registered replication experiment. All participants had self-reported normal or corrected-to-normal vision, hearing, and memory, and had not watched any episodes of *The Chair* before the experiment. Participants gave written consent to enroll in the study under a protocol approved by the Committee for the Protection of Human Subjects at Dartmouth College. Participants received monetary compensation for their time. For two participants, one segment was not played due to a technical error, resulting in four unregistered trials. All available data were used in the analyses. The replication experiment was pre-registered (`https://aspredicted.org/blind.php?x=LV6_953`).

Stimuli

~~The stimulus used in the study~~

Main experiment. The stimuli used in our main experiment were segments of the CBS television series *Why Women Kill* Season 1. The TV series contained three distinct storylines depicting three

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

Replication experiment. The stimuli used in our replication experiment were segments of the first episode of the Netflix television show *The Chair*, Season 1. The TV series depicts the life of a professor who is the English department chair at a major university. The first episode was used in our study and was divided into 13 segments. Segments were divided based on major scene cuts and were minimally edited. The mean length of the segments was 1.97 min (range 0.58–4.30 min). We chose this TV series based on its strictly linear storytelling and its realistic settings where most events depicted everyday life on a college campus.

Task design and procedure

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

Tasks in each sequence alternated between watching, recall, and retrodiction or prediction,

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

Before watching each segment, participants were given the following task instructions. After watching the video, participants were instructed to type their responses (retrodiction, prediction, or recall) in 1–4 sentences. Participants were also asked to specify the characters' names in their responses, i.e., avoiding use of characters' pronouns. For the recall task, the names of the characters in the recall segment were displayed, and participants were asked to summarize the major plot points in the present tense. For the retrodiction and prediction tasks, participants were instructed to retrodict or predict the major plot points of the segment (also in the present tense), as though they had watched the segment and were writing a plot synopsis. They were also instructed to avoid speculation words (e.g., "I *think* Beth will..."). For the uncued retrodiction and prediction tasks, participants made retrodictions or predictions without any cues provided, so they had to

guess which of the characters would be present in the segment. For character-cued retrodictions and predictions, the characters in the target segment were revealed on the screen, alongside participants' previous responses. Participants were instructed to include or incorporate those characters into their character-cued responses, if their previous responses did not contain all the 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 ~~response. For all of the prediction and retrodiction tasks~~responses. For each retrodiction and prediction, participants were ~~instructed to provide~~ asked to generate at least one~~response, but they were given the opportunity enter up to threeresponses if they felt that multiple possibilities were more or less equally likely~~, and not more than three, responses that constituted "the sorts of things [the participant would] expect to have remembered if [they] had watched the [target] segment." They were asked to generate multiple responses only if those additional responses were (in their judgement) of equal likelihood to occur. On average, participants in our main experiment generated 1.08 responses per prompt; therefore we chose to consider only participants' first ("most probable" or "most important") responses to each prompt. 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).

Replication experiment. The design and procedure of the replication experiment were similar to the main experiment, other than the following differences. In the replication experiment, we used only one storyline, and therefore participants performed only one task sequence (either chronological or backward chronological), in one session (Fig. S1). Tasks alternated between watching, and retrodiction or prediction. Some segments contained multiple scenes with different characters. For these segments, characters for each scene were shown in the cued conditions, and participants were asked to guess what would happen in each scene between these characters. For each retrodiction and prediction, participants were asked to generate only one response. No confidence ratings were collected. No practice sessions were provided. At the beginning of the experiment, participants were shown the main characters (with pictures) in the TV show. The experiment was approximately 1 h long, and was implemented using Qualtrics.

Video annotation

Main experiment. 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 references in conversations during onscreen events. Another 16 events were identified based on characters' ~~transits between two places~~implied movements and travels. For example, if in segment 1 character A was in place A and in segment 2 she was in place B, then the transit from place A to B for character A would be identified as an offscreen event. The remaining 15 offscreen events were identified based on logical inferences. For example, if a ~~photo~~

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

Replication experiment. Events in the 13 segments were identified by the authors (X.X. and X.Z.), corresponding to major plot points (total: 71; mean: 5.46 per segment; range 1–14). Additionally, 66 offscreen events were identified. Of these 66 offscreen events, 47 events were identified from references in conversations during onscreen events. Another one event was identified based on characters' implied movements and travels. The remaining 18 offscreen events were identified based on logical inferences.

Response analyses

Participants' retrodiction, prediction, and recall responses were minimally processed to correct obvious typos (e.g., in characters' names) and remove speculation descriptions (e.g., "I predict that..."). We discarded a small number (main experiment: $n = 20$, replication experiment: $n = 6$) of character-cued responses that did not contain references to all cued characters, along with one additional response due to the participant's misunderstanding of the task instructions during that trial in the main experiment. We carried out our analyses on the remaining 1781 retrodiction, prediction, and recall responses in the main experiment, and 878 retrodiction and prediction responses in the replication experiment.

All responses were manually coded and matched to events from the video annotations. Retrodiction and prediction responses were coded by two coders (main experiment: X.X. and Z.Z.; replication experiment: X.X. and X.Z.). Recall responses were coded by one coder (X.X.). While ~~most~~ many responses were clearly identifiable as either matching specific storyline events or as not matching any storyline events, several ambiguous cases arose. First, some responses combined or summarized over several (distinct) storyline events. Second, some responses lacked any specific

detail (e.g., "character A and B talk" without describing the specific topic(s) of conversation or providing other relevant details). Based on participants' responses, in addition to the original 117 onscreen events and 74 offscreen events in the main experiment's stimulus, we added 25 new events (23 onscreen, 2 offscreen) that either summarized across several events or partially matched the annotated events. In our replication study, in addition to the original 71 onscreen events and 66 offscreen events, we added 20 new events (17 onscreen, 3 offscreen). Whereas the original events were each assigned a value of one point, we assigned these additional events a half point. This point system enabled us to directly match events in participants' responses to the annotated events. In our analyses of retrodictions, predictions, and recalls, we added up the number of points earned for each response to estimate participants' event hit rates.

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

~~An additional ambiguous case arose in one participant's responses pertaining to segment 12, storyline 2, whereby the participant correctly identified an onscreen event that had not been included in our original annotations. To account for this participant's response, we retroactively added that event to our annotations of that segment~~We also retroactively added events to the annotations that were mentioned by participants that matched events in future episodes of the TV show. We also identified and counted unmatched events in participants' responses (i.e., events that did not match any annotated events). ~~In several cases, the~~

Resolving ambiguities and estimating inter-rater reliability

We used Jaccard similarity to quantify the inter-rater reliabilities of the annotations, defined as the size of the intersection divided by the size of the union of the two coders' ~~independent scoring disagreed. These cases were resolved through discussions between the two coders.~~ event labels for participants' responses. The Jaccard similarities were calculated for each experiment (across

all trials in the uncued and cued conditions), and unmatched event labels were excluded. We observed a Jaccard similarity of 0.42 for both the main and replication experiments.

This low inter-rater reliability appeared to follow from difficulties related to setting criteria for determining whether a given response counted as a "hit" for a specific event. Whereas we had initially expected that manually matching up participants' responses with events in the narrative would be obvious, empirically we found substantial ambiguities in this process. As one example, during one scene in our replication experiment's stimulus, the main character (Ji-Yoon) chaired a meeting for her department. One participant made a retrodiction response "Ji-Yoon chaired a department meeting" and another participant wrote "All faculty had a meeting." If a given rater's "match" criteria included specifically mentioning that *Ji-Yoon* was *leading* the meeting, only the first participant's response would count as a "hit" for this event. However, a more lenient scorer might consider both responses to be "hits." After reviewing the scores across raters and discussing each scene on a case-by-case basis, the raters decided to re-score the responses using strict criteria (e.g., in the above example, only the first participant's response would be counted as a hit).

Another pattern we observed was that participants' guesses sometimes contained some events that actually happened (or would happen) alongside other incorrect events or details. For example, in another scene in our replication experiment's stimulus, one character (Dafna) gives another character (Bill) a ride in her car. One participant predicted that "Dafna bails Bill out and drives him back to Pembroke or helps him sober up." In one sense, if incorrect or extraneous details are ignored, this response would be considered a "hit" because the participant mentions that Dafna gives Bill a ride. However, if incorrect or extraneous details are factored into the scoring procedure (for example, Dafna never bails Bill out, nor does she help Bill sober up), the same response would be considered a miss. After reviewing the scores across raters and discussing each scene on a case-by-case basis, the raters decided to re-score the responses using the former "ignore incorrect or extraneous details" approach.

The raters repeated this general process of developing scoring criteria, comparing and discussing differences, and re-scoring the responses following those discussions until consensus was reached about every response in both experiments (i.e., Jaccard similarities of 1).

Text embeddings of participants' responses

To estimate the semantic similarities between pairs of responses, we first transformed each response into a 512-dimensional vector (embedding) using ~~*Universal Sentence Encoder*~~the Universal Sentence Encoder (Transformer USE, Cer et al., 2018). We defined *similarity* as the cosine of the angle formed by the responses' vectors. Following Heusser et al. (2021), we defined the *precision* of participants' responses as the median similarity between that response's vector and the embedding vectors for all other participants' recalls of the target segment (main experiment), or the similarity between that response's vector and the embedding vector for the plot synopsis of the target segment (replication experiment). We defined the *convergence* of a given response as the mean similarity between that response's vector and all other participants' responses to the corresponding segment, in the same condition. To compute these median or mean similarities we first applied the Fisher z-transformation to the similarity values, then took the median or mean of the z-transformed similarities, and finally applied the inverse z-transformation to obtain the precision or convergence score.

To test the validity and reliability of the USE embeddings, we performed a classification analysis of recall responses using a leave-one-out approach. For each recall response, we calculated its semantic similarity with all other recall responses for the same storyline. We took the segment with the highest median semantic similarity (to the recall response) as the "predicted" segment. 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%). We note that this validation analysis could only be carried out with data from our main experiment, since we did not collect recall responses in our replication experiment.

Reference coding

Two coders (main experiment: X.X. and Z.Z.; replication experiment: X.X. and X.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 in the main experiment stimulus, and 53 were identified in the replication experiment stimulus. Of these references in the main experiment, 30 referred to onscreen events and 52 referred to offscreen events. In the replication experiment, 13 referred to onscreen events and 40 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 in our main experiment and one out of 13 onscreen events in our replication experiment.

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 ~~participant's~~ 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.

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

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

where for analyses of our main experiment `thp` was the number of points hit for the target segment, `ttp` was the total number of points for the target segment (from its annotations), `direction` was either retrodiction or prediction, `level` had four levels (uncued, character-cued, updated, and recall), `seg_cnt` represented the number of segments in the storyline that had been watched (1–10, centered), `storyline` had two levels (1 or 2), and `target` had 22 levels according to the identity of the target segment. For our analyses of our replication experiment, `level` had two levels (uncued and character-cued), `seg_cnt` ranged from 1–12, the `storyline` parameter was omitted since there was only a single storyline, and `target` had 13 levels according to the identity of the target segment. In the replication experiment, we did not include random slopes of `direction` effect in the participant level in all analyses, as participants either made retrodictions or predictions (i.e., participants and tasks were nested).

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

For our tests comparing the numbers of hits for different types of events (Fig. 4B, S6), we fit generalized linear mixed models using the same formula, but with a ~~poisson~~ Poisson link function. For these models, we manually doubled the point counts to ensure that half points were mapped onto integers, ensuring compatibility with the ~~poisson~~ Poisson link function.

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

```
hp_lag ~ direction * full_stp * lag * storyline +
(direction | base_seg) + (1 | base_seg_pair) +
```

~~(direction * full_stp | lag * storyline | subject)~~

(direction * full_stp * lag * storyline | participant)

where `hp_lag` is the ~~numbers~~ number of "points" earned (for each lag) in each trial (again, we manually doubled the point counts to ensure that half points were mapped onto integers, for compatibility with the ~~poisson~~ Poisson link function), `full_stp` denoted whether the given events (of the given lag) were onscreen (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 (main experiment: 22 levels; replication experiment: 13 levels), and `base_seg_pair` denotes the pairing of the just-watched segment and the segment at each lag (main experiment: 440 levels; replication experiment: 324 levels).

For our analyses of the proportions of events hit for referenced versus unreferenced events (~~Fig~~Figs. 5D, E, S7), 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)~~
(direction * reference * full_stp + lag | participant)
```

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~~Figs. 6D, E, S9, S10; or referenced, referring, and other: ~~Fig~~Figs. 7D, S11), we fit a generalized linear mixed model using the same formula as above, but with three (rather than two) `reference` levels.

Several of our analyses entailed comparing the relative hit rates or probabilities of two different conditions or outcomes. We used the `emmeans` package to compute the odds ratios given the generalized linear mixed models we fit for the given analysis. These odds ratios reflect the odds

(calculated as $\frac{p}{1-p}$), where p is the probability that the outcome occurs) of a particular outcome (e.g., making a response about a particular event) given a scenario (e.g., the event occurred *prior* to the just-watched segment) divided by the odds of the outcome occurring in the alternative scenario (e.g., the event occurred *after* the just-watched segment).

Meta analysis

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

First, we used the `nltk` package (Bird et al., 2009) to segment each document into individual sentences. Each sentence was processed independently of the others. Second, we handled contractions using the `contractions` package (e.g., "we'll" was split into "we will," and so on). Third, we defined two sets of "keywords" (words and phrases) that tended to be indicative of referring to the past (Tab. S6) or future (Tab. S7). We used ChatGPT (OpenAI, 2023) to generate each list, with exactly 50 templates per list, using the following prompt:

```
I'm designing a heuristic algorithm for identifying references (in text) to
past and future events.  Part of the algorithm will involve looking for specific
keywords or phrases that suggest that the text is referring to something that
happened (or will happen) in the past and/or future.  Could you help me generate
a list of 50 keywords or phrases to include in each list (one list for identifying
references to the past and a second list for identifying references to the
future)?  I'd like to be able to paste the lists you generate into two plain
text documents with one row per keyword or phrase, and no other content.  Please
output the lists as a "code" block (enclosed by '''...''').
```

Fourth, we used part-of-speech tagging (again, using the `nltk` package) to look for verbs or verb phrases that were in past or future tenses. After the words were tagged with their predicted parts of speech, we used regular expressions (applied to the sequences of tags) to label each verb or verb phrase with a human readable verb form (e.g., "future perfect continuous passive," "conditional perfect continuous passive," and so on). The regular expressions we used to generate these labels are shown in Table S4, and the part of speech tags are defined in Table S5.

We treated each keyword match (of past or future keywords) as a single "reference" (to a past or future event, respectively), and if any past or future verb forms were detected we treated those as (up to) one additional reference. We then tallied up the numbers of past and/or future references across sentences within the given document. The meta analysis results reported in Figure 8C display the average numbers of references aggregated across all documents within each dataset we analyzed (described in Tab. S3).

Finally, we used a bootstrap procedure to quantify the magnitude of temporal imbalances. For each dataset, we sampled (with replacement) a set of n observations (where n was the number of observations in the given dataset). We computed the effect size for that sample as the total number of references to past events divided by the total number of references to future events. We repeated this process across 100 iterations to obtain a distribution of effect sizes. If the 95% confidence interval of a dataset's distribution does not contain 1 (i.e., an equal balance of past and future references), this indicates that there are significantly more past than future references (if the lower 2.5% is greater than 1), or significantly more future than past references (if the upper 2.5% is less than 1), at the $p < 0.05$ threshold. We estimated p-values as 2 times the minimum between (a) the proportion of bootstrapped effect sizes greater than 1 and (b) the proportion of bootstrapped effect sizes less than 1. In the main text, we report these average effect sizes within and across datasets (computed as $\exp(\mathbb{E}[\log(x)])$, where x is the bootstrap-estimated distribution of effect sizes for the given dataset).

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~~`https://github.com/ContextLab/prediction-retrodiction-paper`

References

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

Asch, S. E. and Ebenholtz, S. M. (1962). The principle of associative symmetry. *Proceedings of the American Philosophical Society*, 106:135–163.

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

Barron, H. C., Auksztulewicz, R., and Friston, K. (2020). Prediction and memory: a predictive coding account. *Progress in Neurobiology*, 192:101821–101834.

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

Bird, S., Klein, E., and Loper, E. (2009). *Nature language processing with Python: analyzing text with the natural language toolkit*. Reilly Media, Inc.

Block, R. A. and Gruber, R. P. (2014). Time perception, attention, and memory: a selective review. *Acta Psychologica*, 149:129–133.

Bordwell, D. (2008). *Poetics of cinema*, chapter Three dimensions of film narrative, pages 85–134. Routledge.

Bower, G. H., Black, J. B., and Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, 11(2):177–220.

Carroll, S. (2010). *From eternity to here: the quest for the ultimate theory of time*. Penguin.

Carroll, S. (2016). *The big picture: on the origins of life, meaning, and the universe itself*. Dutton.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. *arXiv*, 1803.11175.

Chow, W.-Y., Momma, S., Smith, C., Lau, E., and Phillips, C. (2016). Prediction as memory retrieval: timing and mechanisms. *Language, Cognition and Neuroscience*, 31(5):617–627.

Cover, T. M. (1994). Which processes satisfy the second law? In Halliwell, J. J., Pérez-Mercader, J., and Zurek, W. H., editors, *Physical Origins of Time Asymmetry*, pages 98–107. Cambridge University Press, Cambridge, UK.

Cunillera, T., Càmara, E., Toro, J. M., Marco-Pallares, J., Sebastián-Galles, N., Ortiz, H., Pujol, J., and Rodrígues-Fornells, A. (2009). Time course and functional neuroanatomy of speech segmentation in adults. *NeuroImage*, 48(3):541–553.

Daikoku, T., Yatomi, Y., and Yumoto, M. (2014). Implicit and explicit statistical learning of tone sequences across spectral shifts. *Neuropsychologia*, 63:194–204.

Daikoku, T., Yatomi, Y., and Yumoto, M. (2015). Statistical learning of music- and language-like sequences and tolerance for spectral shifts. *Neurobiology of Learning and Memory*, 118:8–19.

de Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1):1–12.

Demiray, B., Mehl, M. R., and Martin, M. (2018). Conversational time travel: evidence of a retrospective bias in real life conversations. *Frontiers in Psychology*, 9:2160.

Dessalles, J.-L. (2007). Storing events to retell them. *Behavioral and Brain Sciences*, 30(3):321–322.

Eagleman, D. M. (2008). Human time perception and its illusions. *Current Opinion in Neurobiology*, 18(2):131–136.

1174 Ellison, C. J., Mahoney, J. R., and Crutchfield, J. P. (2009). Prediction, retrodiction, and the amount of
1175 information stored in the present. *Journal of Statistical Physics*, 136(1005):doi.org/10.1007/s10955–
1176 009–9808–z.

1177 Feynman, R. (1965). *The character of physical law*. MIT Press.

1178 Furl, N., Kumar, S., Alter, K., Durrant, S., Shawe-Taylor, J., and Griffiths, T. D. (2011). Neural
1179 prediction of higher-order auditory sequence statistics. *NeuroImage*, 54(3):2267–2277.

1180 Gołosz, J. (2021). Entropy and the direction of time. *Entropy*, 23(4):388.

1181 Grant, J. B. and Walsh, E. (2016). Exploring the use of experience sampling to assess episodic
1182 thought. *Applied Cognitive Psychology*, 30(3):472–478.

1183 Hasson, U. (2017). The neurobiology of uncertainty. *Philosophical Transactions of the Royal Society of*
1184 *London Series B: Biological Sciences*, 372(1711):20160048.

1185 Hawking, S. W. (1985). Arrow of time in cosmology. *Physical Review D*, 32(10):2489–2495.

1186 Healey, M. K. and Kahana, M. J. (2014). Is memory search governed by universal principles or
1187 idiosyncratic strategies? *Journal of Experimental Psychology: General*, 143(2):575–596.

1188 Hemmo, M. and Shenker, O. (2019). The second law of thermodynamics and the psychological
1189 arrow of time. *The British Journal for the Philosophy of Science*.

1190 Heusser, A. C., Fitzpatrick, P. C., and Manning, J. R. (2021). Geometric models reveal behavioral and
1191 neural signatures of transforming experiences into memories. *Nature Human Behavior*, 5:905–919.

1192 Hirst, W. and Echterhoff, G. (2012). Remembering in conversations: the social sharing and reshap-
1193 ing of memories. *Annual Review of Psychology*, 63(1):55–79.

1194 Horwich, P. (1987). *Asymmetries in time: problems in the philosophy of science*. MIT Press.

1195 Howard, M. W. (2018). Memory as perception of the past: compressed time in mind and brain.
1196 *Trends in Cognitive Sciences*, 22(2):124–136.

Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46:269–299.

Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL neurons "jump back in time" in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.

Ivry, R. B. and Schlerf, J. E. (2008). Dedicated and intrinsic models of time perception. *Trends in Cognitive Sciences*, 12(7):273–280.

Jones, J. and Pashler, H. (2007). Is the mind inherently forward looking? comparing prediction and retrodiction. *Psychonomic Bulletin and Review*, 14(2):295–300.

Josselyn, S. A. and Tonegawa, S. (2020). Memory engrams: recalling the past and imagining the future. *Science*, 367(6473):eaaw4325.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory and Cognition*, 24:103–109.

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory and Cognition*, 30:823–840.

Kahana, M. J., Diamond, N. B., and Aka, A. (2022). Laws of human memory. In Kahana, M. J. and Wagner, A. D., editors, *Handbook of Human Memory*. Oxford University Press.

Koelsch, S., Busch, T., Jentschke, S., and Rohrmeier, M. (2016). Under the hood of statistical learning: a statistical MMN reflects the magnitude of transitional probabilities in auditory sequences. *Scientific Reports*, 6(19741):doi.org/10.1038/srep19741.

Koster-Hale, J. and Saxe, B. (2013). Theory of mind: A neural prediction problem. *Neuron*, 79(5):836–848.

Lange, K., Kühn, S., and Filevich, E. (2015). "Just Another Tool for Online Studies" (JATOS): an easy solution for setup and management of web servers supporting online studies. *PLoS One*, 10(6):e0130834.

Maheu, M., Meyniel, F., and Dehaene, S. (2022). Rational arbitration between statistics and rules in human sequence processing. *Nature Human Behaviour*, pages 1–17.

Mahr, J. B. and Csibra, G. (2018). Why do we remember? the communicative function of episodic memory. *Behavioral and Brain Sciences*, 41:e1.

Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors, *Handbook of Human Memory*. Oxford University Press.

Manning, J. R. (2021). Episodic memory: mental time travel or a quantum "memory wave" function? *Psychological Review*, 128(4):711–725.

Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory. In Gazzaniga, M., editor, *The Cognitive Neurosciences*, pages 557–566. MIT Press.

Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences, USA*, 108(31):12893–12897.

Manns, J. R., Howard, M. W., and Eichenbaum, H. (2007). Gradual changes in hippocampal activity support remembering the order of events. *Neuron*, 56(3):530–540.

McNealy, K., Mazziotta, J. C., and Dapretto, M. (2006). Cracking the language code: neural mechanisms underlying speech parsing. *The Journal of Neuroscience*, 26(29):7629–7639.

Meyer, M. L., Zhao, Z., and Tamir, D. I. (2019). Simulating other people changes the self. *Journal of Experimental Psychology: General*, 148(11):1898–1914.

Mlodinow, L. and Brun, T. A. (2014). Relation between the psychological and thermodynamic arrows of time. *Physical Review E*, 89(5):052102.

Momennejad, I. and Howard, M. W. (2018). Predicting the future with multi-scale successor representations. *bioRxiv*, page doi.org/10.1101/449470.

OpenAI (2023). ChatGPT. Personal communication.

Park, G., Schwartz, H. A., Sap, M., Kern, M. L., Weingarten, E., Eichstaedt, J. C., Berger, J., Stillwell, D. J., Kosinski, M., Ungar, L. H., and Seligman, M. E. P. (2017). Living in the past, present, and future: measuring temporal orientation with language. *Journal of Personality*, 85(2):270–280.

Pillemer, D. B., Steiner, K. L., Kuwabara, K. J., Thomsen, D. K., and Svob, C. (2015). Vicarious memories. *Consciousness and Cognition*, 36:233–245.

Polyn, S. M. and Kahana, M. J. (2008). Memory search and the neural representation of context. *Trends in Cognitive Sciences*, 12:24–30.

Radvansky, G. A. and Copeland, D. E. (2006). Walking through doorways causes forgetting: situation models and experienced space. *Memory and Cognition*, 34(5):1150–1156.

Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature Reviews Neuroscience*, 13:713–726.

Rovelli, C. (2022). Memory and entropy. *Entropy*, 24(8):1022.

Schacter, D. L. (2012). Constructive memory: past and future. *Dialogues in Clinical Neurosciences*, 1:7–18.

Schacter, D. L., Norman, K. A., and Koutstaal, W. (1998). The cognitive neuroscience of constructive memory. *Annual Review of Psychology*, 49:289–318.

Schacter, D. L. and Tulving, E. (1994). *Memory systems 1994*. MIT Press, Cambridge, MA.

Schapiro, A. and Turk-Browne, N. (2015). Statistical learning. *Brain Mapping: An Encyclopedic Reference*, 3:501–506.

Shankar, K. H. and Howard, M. W. (2012). A scale-invariant internal representation of time. *Neural Computation*, 24:134–193.

Shipp, A. J. and Aeon, B. (2019). Temporal focus: thinking about the past, present, and future. *Current Opinion in Psychology*, 26:37–43.

Song, X. and Wang, X. (2012). Mind wandering in Chinese daily lives – an experience sampling study. *PLoS One*, 7(9):e44423.

Tamir, D. I. and Mitchell, J. P. (2013). Anchoring and adjustment during social inferences. *Journal of Experimental Psychology: General*, 142(1):151–162.

Tamir, D. I. and Thornton, M. A. (2018). Modeling the predictive social mind. *Trends in Cognitive Sciences*, 22(3):201–212.

Tummeltshammer, K., Amso, D., French, R. M., and Kirkham, N. Z. (2016). Across space and time: infants learn from backward and forward visual statistics. *Developmental Science*, 20(5):e12474.

Wearden, J. (2016). *The psychology of time perception*. Springer.

Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit memories to other brains: constructing shared neural representations via communication. *Cerebral Cortex*, 27(10):4988–5000.

Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162–185.

Acknowledgements

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

Author contributions

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

Competing interests

The authors declare no competing interests.