Behavior Research Methods (2025) 57:163
https://doi.org/10.3758/s13428-025-02690-7

ORIGINAL MANUSCRIPT

Assessing autobiographical memory consistency: Machine and human approaches

Victoria Wardell[1] · Taylyn Jameson[1] · Peggy L. St. Jacques[2] · Christopher R. Madan[3] · Daniela J. Palombo[1]

Accepted: 11 April 2025

Abstract

Memory is far from a stable representation of what we have encountered. Over time, we can forget, modify, and distort the details of our experiences. How autobiographical memory—the memories we have for our personal past—changes has important ramifications in both personal and public contexts. However, methodological challenges have hampered research in this area. Here, we introduce a standardized manual scoring procedure for systematically quantifying the consistency of narrative autobiographical memory recall and review advancements in natural language processing models that might be applied to examine changes in memory narratives. We compare the performance of manual and automated approaches on a large dataset of memories recalled at two time points placed approximately 2 months apart (N(memory pairs) = 1,026). We show that human and automated approaches are moderately correlated (r = .21-.46), though numerically human scorers provide conservative measures of consistency, while machines provide a liberal measure. We conclude by highlighting the strengths and limitations of both manual and automated approaches and recommend that human scoring be employed when the *types* of mnemonic details that are consistent over time and/or what drives *in*consistencies in memory are of interest.

Keywords Autobiographical memory · Memory consistency · Memory accuracy · Narrative data · Episodic memory · Memory distortion

Since Ebbinghaus famously demonstrated forgetting curves in the first empirical examination of memory (1885), the malleability of memory has been of central interest in memory research (also see Linton, 1975; Bartlett, 1932). After all, despite the remarkably vivid and convincing detail with which humans can recount their past, memory is not a stable representation of what was encountered. Just as Ebbinghaus's nonsense words slipped from his memory over time, aspects of our lived experiences can similarly fade (Misra et al., 2018). Moreover, not only do we forget elements of the past, but we are also prone to embellishing, contradicting, and schematizing the details of our experiences (Schacter, 2022; Schacter et al., 2011). The changeability of autobiographical memory, the memories we have for our personal past, remains a crucial though relatively understudied realm of research. From discrepancies in memory for a shared experience between friends to memory lapses made by political leaders and eyewitnesses, the way autobiographical memory changes is highly relevant to and consequential for humans. Here, we present innovative methodologies for examining changes in autobiographical memory narratives. We describe a novel hand-scoring procedure that quantifies changes in memory over time as well as natural language processing models designed to capture similarity between two texts. We compare and contrast the two approaches and conclude with best practice recommendations for measuring consistency in autobiographical memory research.

Decades of research has highlighted autobiographical memory's dynamic nature. Seminal work in the late 1900s brought the pliability of autobiographical memory to the forefront, with research revealing stark changes in memory for even the most consequential experiences of the time, including the Challenger explosion (Bohannon, 1988; Neisser & Harsch, 1992), the assassination of JFK (Brown & Kulik, 1977; Yarmey & Bull, 1978), and the Watergate

✉ Daniela J. Palombo
daniela.palombo@ubc.ca

[1] Department of Psychology, University of British Columbia, 2136 West Mall, Vancouver, BC V6 T 1Z4, Canada

[2] Department of Psychology, University of Alberta, Edmonton, AB, Canada

[3] School of Psychology, University of Nottingham, Nottingham, UK

Published online: 06 May 2025

scandal (Neisser, 1981). This and related work laid the foundation for myriad influential findings, including the lack of relationship between the consistency of a memory and its phenomenology, such as the vividness with which the memory is recalled (Talarico & Rubin, 2003, 2007) and one's confidence in the accuracy of their recollection (Nielsen et al., 2023; Talarico & Rubin, 2003, 2007), as well as the tendency for positive events to be more prone to distortions than negative events (Kensinger & Schacter, 2006), the role of post-event information in shifting recalls (Loftus, 2003; Thomas & Loftus, 2002), and the tendency for memory changes to level off over time (Hirst et al., 2015; Talarico & Rubin, 2007)—just to name a few. Yet memory is not irredeemably flawed. Memory seems to be less prone to change when the event is rehearsed more frequently (Campbell et al., 2011; Nadel et al., 2007).[1] Similar, repeated events (i.e., "repisodes"; Neisser, 1981) tend to adopt a gist-like quality, whereby fine details of the event(s) may fade or bleed across one another, but the main gestalt remains—relatively accurately—intact (Koriat et al., 2000; Barclay & Wellman, 1986; Goldsmith et al., 2005; also see Fivush & Grysman, 2022). Recently, Diamond and colleagues (2020b) found that memory for details of unique events can also be highly accurate: of the details remembered for a naturalistic staged event, 90–95% were true to what was encoded (though there was the expected decay of memory for details). These findings have begun to elucidate the nuanced boundaries of memory fidelity, indicating the range of ways a memory can be both accurate and distorted.

Observations of how autobiographical and other types of memory change over time have contributed to discussion of the highly adaptive nature of humans' flexible memory system. Many have argued that the malleability of memory indicates that memory is a constructive process through which we extract elements of a past experience, flexibly drawing upon, and even combining across, events in order to access relevant information to be recalled (Devitt et al., 2015; Schacter & Addis, 2007; Schacter et al., 2011; Wardell & Palombo, 2024). This constructive process allows us to retrieve pertinent information for the moment at hand while allowing superfluous or irrelevant details to decay. At the same time, these processes may introduce error and noise into our recollections that can result in changes to the original memory (Wardell & Palombo, 2024). Theories of reconsolidation posit that retrieving a memory makes it labile, and what is recalled is what will be reconsolidated in long-term storage—at least until the next retrieval (Gilboa & Moscovitch, 2021). Destabilization of the memory creates a window in which memory can be updated, in light of our current knowledge, schemas, and expectations (Bartlett, 1932; Fivush & Grysman, 2022; Gilboa & Moscovitch, 2021). Though theories of the utility of memory malleability are well accepted, the way we think about memory outside of research laboratories often seems at odds with current empirical understanding. Eyewitnesses are expected to recount only accurate details. Therapists are expected to shape healthier ways of remembering the past—though without implanting fabricated details. Indeed, connecting with loved ones by discussing shared experiences is only possible if at least some of the details of the experience are shared in the two (or more) separate memory systems over time. The way memory is used in a variety of settings requires that we understand the ways in which memory may reflect an encoded experience, and the ways in which it may not.

The study of autobiographical memory has called upon the creativity of researchers to implement methods that might maintain scientific control over a process that typically begins outside of the laboratory as autobiographical memories are encoded in the messiness of day-to-day life. Researchers have capitalized on diary studies, in which participants document their experiences the day they occurred (e.g., Rubin et al., 2011; Thomsen et al., 2015), and interviews, in which researchers probe participants for mnemonic details of past events, often with standardized and quantifiable protocols such as the Autobiographical Memory Interview (Kopelman et al., 1989), the TEMPau task (Piolino et al., 2009), or the Autobiographical Interview (AI; Levine et al., 2002). More recently, researchers have staged events with naturalistic qualities in a setting over which they can control encoding conditions (Marcotti & St. Jacques, 2018; Diamond et al., 2020a, 2020b). Still, the deep personal meaning, extremes of emotion, and social consequences often of interest to researchers are not easily reproducible in a laboratory. Laboratory-based and real-world studies of memory afford different strengths, and the integration of findings from both has been fundamental in building theories of memory (e.g., Gilboa & Moscovitch, 2021; Palombo et al., 2018; Wardell & Palombo, 2024).

Autobiographical memory is multifaceted and can thus be characterized in multiple ways, including in terms of the phenomenological experience of recalling a past experience (e.g., the imagery, emotions, and sensations the memory elicits; Luchetti & Sutin, 2015; Rubin et al., 2004), the effort required to bring a memory to mind (Barzykowski & Staugaard, 2016), the broader role of an experience in shaping one's life story (Conway et al., 2019; McAdams, 2008), or the mnemonic details that come to mind (e.g., what happened or was experienced; Levine et al., 2002; St. Jacques & Levine, 2007; Wardell et al., 2021). The fidelity of autobiographical memory is typically considered at the level of

[1] Interestingly, rehearsal via media exposure or for highly public events may have opposing effects on memory maintenance, whereby more rehearsal in such contexts has, at times, been associated with less consistent memories (see Talarico & Rubin, 2017, for a review).

mnemonic detail, that is to say, the correspondence of the details we recall with the details of the event that occurred.

Memory accuracy can be thought of as how similar a recall is to the event that took place. A lack of accuracy inherently reflects errors in memory, defined as memory for content that does not correspond with what actually happened (Fisher et al., 2009; Koriat et al., 2000). Objective verification of autobiographical details is often unavailable to researchers, however, as they seldom have access to the details of autobiographical events. Therefore, consistency across multiple recalls for the same experience can be used as a helpful proxy. Consistency can be defined as content that is repeatedly retrieved from memory over time (Baugerud et al., 2014; Fisher et al., 2009; Stanley & Benjamin, 2016). Unlike accuracy, a lack of consistency does not necessarily reflect errors in memory. Inconsistency may reflect new accurate information provided at follow-up, or omitted accurate information from an initial session. In the consistency framework, it is not possible to determine whether new content is erroneous; it may simply have been omitted from the initial recall. Similarly, omitted details may or may not have been erroneous at initial recall, and even inaccurate details might be consistently recalled (Hirst et al., 2015), though data do suggest that the more consistent a memory is, the more accurate it seems to be (Fisher et al., 2009). Moreover, experiences themselves are fleeting. It is what lingers in memory and what we make of that memory that lasts as time passes. Through this lens, consistency is a relevant and fascinating phenomenon in its own right.

At perhaps the broadest level, inconsistencies in memory are often characterized based on Schacter's (1999, 2022) division of omissions, that is, the lack of a memory or access to it (e.g., forgetting), and commissions, that is, the presence of a memory that may not be faithful to encoding (e.g., embellishment). Still, further distinctions within the category of omissions and commissions can be made when considering memory consistency. The eyewitness testimony literature, in which subsequent testimonies from the same individual or testimonies across eyewitnesses might be compared to narrow in on the facts of a crime, has documented how details can be omitted from prior recollections, new details could be provided at subsequent recollections, details between accounts could be consistent, or details can directly contradict each other (Fisher et al., 2009; Orbach et al., 2012). Still, often, details provided during recall do not directly contradict or confirm details from prior recollections, falling into a gray area of similar content.

Researchers have developed scoring schemes to examine how memory for specific details might shift over time (e.g., changes in answers to questions such as "Where were you?", "Who were you with?"; Kensinger & Schacter, 2006; also see Hirst et al., 2015; Talarico & Rubin, 2003). Yet naturalistic recall of autobiographical memories often produces an unfolding narrative or story of what occurred, particularly when discussing past experiences with others (Palombo, 2024; also see Bluck, 2003; Boyd, 2018). Researchers have exploited the richness of autobiographical memory narratives to understand how autobiographical memory differs as a function of age (Levine et al., 2002), emotion (St. Jacques & Levine, 2007; Wardell et al., 2021), mental health (McKinnon et al., 2015; Söderlund et al., 2014), and much more. However, no standardized approach to examining the *consistency* of memory narratives over time has been established in the field.

The goal of the present paper is to outline standardized approaches to measuring the consistency of autobiographical memory narratives. As part of a larger ongoing study on emotional memory consistency, we had participants recall two events, one negative and one neutral, at two time points placed approximately 8 weeks apart (see https://aspredicted.org/rnp6-r5t9.pdf for details about that study). We employ the data from this study to present a novel hand-scoring procedure that builds on Levine and colleagues' widely used Autobiographical Interview (AI; 2002) to quantify memory consistency, namely the Autobiographical Interview Consistency Supplement (AI-CONS). We further compute semantic similarity scores on this dataset using natural language processing models, specifically, latent similarity analysis (LSA; Landauer et al., 1998), DistilBERT (Sanh et al., 2020), and MPNet (Song et al., 2020), demonstrating the potential of large language models to identify similarities between narrative recalls. We consider the strengths and differences between these two approaches by comparing them to each other and to other memory characteristics, including the age of the memory and the detailedness of the memory.

Methods

Open practices statement

Materials and analysis code are available at https://osf.io/msc9n. The reported study was not preregistered.

Participants

A total of 564 participants completed both the initial and follow-up sessions for our ongoing study on emotional autobiographical memory. Data were collected online using the Qualtrics survey platform. Participants were recruited from the Prolific online data collection platform ($n = 530$) and the University of British Columbia's student subject pool ($n = 34$) and were compensated US $8/h or received course credit for their participation, respectively. Of note, 33 of these participants indicated that they could not remember one or both of the events under investigation at the follow-up

session, despite being provided with a title for their event. An additional 18 participants[2] recalled the wrong event at follow-up, despite indicating that they remembered the event in question. The remaining 513 participants ($N_{Prolific}$ = 480; age range 18–46 years; M (SD)$_{Age}$ = 32.4 (7.94) years; 51.1% women, 44.6% men, and 3.3% gender-diverse[3]) resulted in a final dataset of 1,026 events and 2,052 memories. At the initial recall, these events ranged from 1 to 30 days old (M (SD) = 6.36 (4.50), Mdn = 5) and 54 to 106 days old at follow-up (M (SD) = 65.20 (6.20), Mdn = 64), with retention intervals ranging from 53 to 81 days (M (SD) = 58.80 (4.40), Mdn = 58).

Procedure

At the initial session, participants were asked to select one negative and one neutral event from their life that had occurred within the last 2 weeks, not including today. It was requested that events be ones that participants were personally involved in and specific to a time and place but not to include overly mundane events (e.g., brushing teeth) or those that involved substances. Participants were informed that if they could not think of an event(s) from the last 2 weeks, they could select an event from within the last month. A clear and identifiable title for both events was collected to cue participants to recall their memories at the follow-up session. Participants indicated what, when, and where their events took place (e.g., "Walk with Marta on March 15, 2023, at Stanley Park").[4] Participants were then instructed to type out all the details they could remember about their selected events. Specifically, participants were instructed to "Please type out everything you can remember about this specific event. The 'next' arrow will appear after 3 min, but please take as long as you need to recall every detail that comes to mind. Remember that I want you to tell me every detail that comes to mind so I can really picture what you are remembering." Approximately 8 weeks later, participants were presented with their event titles and asked to indicate whether or not they recognized the event from the event title provided. Participants proceeded to re-recall their events, following the same instructions as provided in session 1. Although we excluded recalls for events that participants indicated not remembering, we collected recalls of these events to ensure participants did not indicate not remembering simply to move through the study more quickly.

Following data collection, memory narratives were qualitatively examined in order to confirm that participants had recalled the same event at both sessions. Identifying whether the same event has been recalled at follow-up is crucial for studies of autobiographical memory consistency, as inconsistencies may be overestimated if narratives for different events are compared.

Data processing

Human consistency scoring: The AI-CONS

Autobiographical interview scoring Details provided in memory narratives were identified using the Autobiographical Interview (Levine et al., 2002). Briefly, the Autobiographical Interview scoring procedure is a reliable tool used to measure *types* of mnemonic details in narrative recall (Lockrow et al., 2024). It is primarily used in episodic memory studies but has been expanded to narrative accounts of episodic future simulations (Race et al., 2011), counterfactual events (DeBrigard et al., 2016), and self-referential processing (Verfaellie et al., 2019). The Autobiographical Interview defines a detail as any piece of information. For example, *"I was studying in the library*" would be scored as two details, one for *"I was studying*" and one for *"in the library*." The Autobiographical Interview further categorizes details as internal, that is, directly referring to the event being recalled (i.e., episodic content), and external, that is, a detail that does not refer directly to the event being recalled. Internal details are parsed into five categories: event (i.e., what happened, who was there), perceptual (i.e., sensations and percepts), emotion/thoughts (i.e., emotions and thoughts), place (i.e., location), and time (i.e., temporal setting). External details are parsed into semantics, external events (i.e., episodic details regarding events not specific to the event being recalled), other details (i.e., metacognitive statements, editorializations), and repetitions.[5]

AI-CONS scoring Autobiographical Interview details identified in memory narratives were compared across sessions for their consistency using the AI-CONS (see Table 1). The AI-CONS identifies whether details are consistent,

[2] One participant indicated that they could not remember one of their events and recalled the wrong event for the other. This participant was included in the count of 33 participants who indicated not remembering the event and not in the count of participants who recalled the wrong event.

[3] One participant chose not to report data on age and five participants (1.0%) chose not to report data on gender.

[4] For some research questions concerning memory consistency, events under investigation may be salient enough that general descriptions provide an adequate cue (e.g., "wedding day" or "9/11").

[5] Further delineation of external details has been proposed to better capture nuances in recalled content not directly concerning the event under investigation, such as personal versus general semantic details and repeated versus extended external events (see Strikwerda-Brown et al., 2019; Renoult et al., 2020; Melega et al., 2023).

Table 1 AI-CONS detail types

AI-CONS detail type	Description	Example	
		Initial recall	Follow-up recall
Consistent	Detail was in both the initial recall and follow-up recall	*I met up with my friend Marta*	*Marta and I got together*
Contradictory	Detail in the follow-up recall contradicted detail in the initial recall	*We went to the park*	*We went to the beach*
Similar	Detail in the follow-up recall was similar to a detail in the initial recall	*I was furious*	*I was annoyed*
New	Detail in the follow-up recall that was not in the initial recall	-	*It was very early*
Omitted	Detail in the initial recall that was not in the follow-up recall	*It was so dark out*	-
Other	Detail in either recall is an editorialization or a repetition	*…maybe it was closer to 2:00 PM*	-
		-	*Like I just mentioned, Marta and I met up*

contradictory, similar,[6] omitted, new, or other, allowing researchers to distinguish how memories change over time. *Consistent* details are reserved for details that nearly precisely match one another. *Contradictory* details identify when what was said guarantees that either the follow-up or initial detail must be false. *Similar* details identify when what was said does not guarantee a detail is false but is not directly reflective of another detail. *Omitted* details capture those provided at the initial session that are not mentioned at follow-up, while *new* details capture those provided at follow-up that have no corresponding detail in the initial recall. Finally, editorializations, repeated details, or intra-memory corrections identified as other details in the Autobiographical Interview (e.g., "at 1:30 PM, or maybe it was closer to 2:00 PM, we went to the store") in either session are scored as *other* details.

Layering the AI-CONS onto narratives scored with the Autobiographical Interview not only allows consistency categories to be tagged onto internal (episodic) and external Autobiographical Interview details, but also allows for more fine-grained tagging onto detail subtypes (e.g., events, perceptions, emotions/thoughts, place, time), should the researcher be interested in this level of nuance. We note, however, that Autobiographical Interview sub-details are, in part, determined by how the information is described. For example, "the music stopped" would be scored as a perceptual detail, while "I turned off the music" would be scored as an event detail. The AI-CONS scoring procedure does not penalize participants for recalling consistent information that changes in Autobiographical Interview detail categorization across recalls. For example, consider the following:

> *Initial recall:* "We went to my friend's house. On the way, we drove through leaves."
>
> *Follow-up recall:* "We drove to her house and there were leaves everywhere."

In accordance with the Autobiographical Interview, the details in the initial recall would be scored as an internal place and event detail, respectively, while the details in the follow-up recall would be scored as an internal place and perceptual detail, respectively. All four details would be scored as consistent details in the AI-CONS.

Events were randomly assigned to nine scorers who scored both the initial recall and follow-up recall of their assigned events in accordance with the Autobiographical Interview and AI-CONS protocol. Scorers were undergraduate research assistants who were trained on both scoring procedures. Once scorers reached reliability on the Autobiographical Interview, in accordance with Levine and colleague's training procedure (see https://levinelab.weebly.com/resources.html), scorers began training on the AI-CONS. Training consisted of scoring memories from past datasets used in developing the AI-CONS (see Dev et al., 2022; Wardell et al., 2023), as well as memories provided by the research team. Weekly meetings were held over the course of 6 weeks to discuss scoring disagreements and refine the training protocol with clearer examples and definitions of AI-CONS detail categories. The protocol and training regime can be found at https://osf.io/msc9n.

A random subset of 90 events (i.e., 180 memories) were scored by all nine scorers for reliability analysis. We used intraclass correlation analyses modeled with

[6] We note that similar details were labeled as reminiscent details in Wardell et al., 2023. We altered the term used for these details to avoid confusion with the term reminiscence, which is commonly used in the eyewitness testimony literature (e.g., Gilbert & Fischer, 2006; Orbach et al., 2012; Odinot et al., 2013) and describes details that are present at follow-up but missing from initial recall (akin to new details in the AI-CONS).

Table 2 Cronbach's alpha scores for curators of the AI-CONS

Detail type	Single rater ICC (95% CI)	Average rater ICC (95% CI)	Two-way random effects
Consistent	0.90 (0.88, 0.92)	0.99 (0.98, 0.99)	$F(179, 1{,}440) = 79.12$**
Omitted	0.97 (0.96, 0.98)	0.99 (0.99, 0.99)	$F(179, 1{,}440) = 294.24$**
New	0.95 (0.94, 0.96)	0.99 (0.99, 0.99)	$F(179, 1{,}440) = 186.42$**
Contradictory	0.49 (0.43, 0.55)	0.89 (0.87, 0.92)	$F(179, 1{,}440) = 9.51$**
Similar	0.26 (0.21, 0.33)	0.76 (0.71, 0.81)	$F(179, 1{,}440) = 4.24$**
Other	0.70 (0.66, 0.75)	0.96 (0.95, 0.96)	$F(179, 1{,}440) = 22.47$**

Inter-rater reliability for each AI-CONS detail type using absolute agreement. We note that poor reliability for contradictory and similar details in AI-CONS scoring were likely a result of encountering a floor effect for these detail types, with 97.6% and 88.9% of memories containing two or fewer contradictory and similar details, respectively

ICC intraclass coefficient, ** $p < 0.0001$.

two-way random effects (i.e., we would expect a similar pattern of results if we randomly selected new scorers with similar characteristics) with absolute agreement, as we are most interested in the extent to which scorers achieve the same score. We judge these options to be the most conservative, and thus a stringent test of inter-rater reliability (see Koo & Li, 2016). We report the results for both "single" rater and "mean of *k* raters," as some readers may be interested in using a single rater as the basis of their measurement, whilst others may use an average of raters.

Scorers showed excellent agreement on total Autobiographical Interview details (single rater alpha = 0.96; average raters alpha = 0.99) and AI-CONS consistent, new, and omitted details (all alphas ≥ 0.90). The contradictory category showed moderate agreement, and the similar category showed poor agreement (see Table 2). In exploring the lower agreement observed, we identified that both contradictory and similar details were at floor in memory narratives, with 97.6% of memories containing two or fewer contradictory details (range 0–7; *Mdn* = 0) and 88.9% of memories containing two or fewer similar details (range 0–10; *Mdn* = 1). It is likely that the restricted range of these detail categories contributed to the poor reliability observed.

AI-CONS consistency categories were calculated as the proportion of each consistency tag for a given detail type out of the total number of those details provided. For example, overall consistency was calculated as the total number of consistent details at the initial and follow-up recall divided by the total number of details provided in the initial and follow-up recall (see Fig. 1).[7] This calculation was inspired by the Dice similarity coefficient (Dice, 1945), which is used to measure the similarity between two sets (e.g., memories) out of the total probability space of both sets, while giving equal importance to the overlap provided by both sets by multiplying the value by 2. Notably, nuances in narrative recall can lead to a different raw number of consistent (or similar or contradictory) details between sessions, and so instead of multiplying the "overlap" by 2, we adjust the calculation by adding the number of details in session 1 and the number of details in session 2. For example, consider the following:

> *Initial recall*: "I turned my test over and saw the grade."
>
> *Follow-up recall*: "I flipped over the paper because the grade was on the other side, and I saw my mark."

While all the details recalled are consistent, there are only two details provided in the initial recall ("I turned my test over" and "and saw the grade") while there are three provided in the follow-up recall ("I flipped over the paper," "because the grade was on the other side," and "and I saw my mark"). Notably, Dice similarity coefficients are appropriate for calculations with sets of differing sizes, making it appropriate for memories containing different numbers of details across sessions. In the case of new and omitted details, for which there is no overlap between sessions, proportions are calculated as the total number of new or omitted details out of the total number of details across sessions.

[7] *Other* details are included in the denominator of AI-CONS proportions, as some populations of interest to memory researchers may have systematic differences in *other* details, particularly repeated or corrected details. We note that when the AI is conducted in person, participants may make requests that are unrelated to their recall (e.g., "Can I go to the washroom?"). We do not include these editorializations in calculation of the AI-CONS.

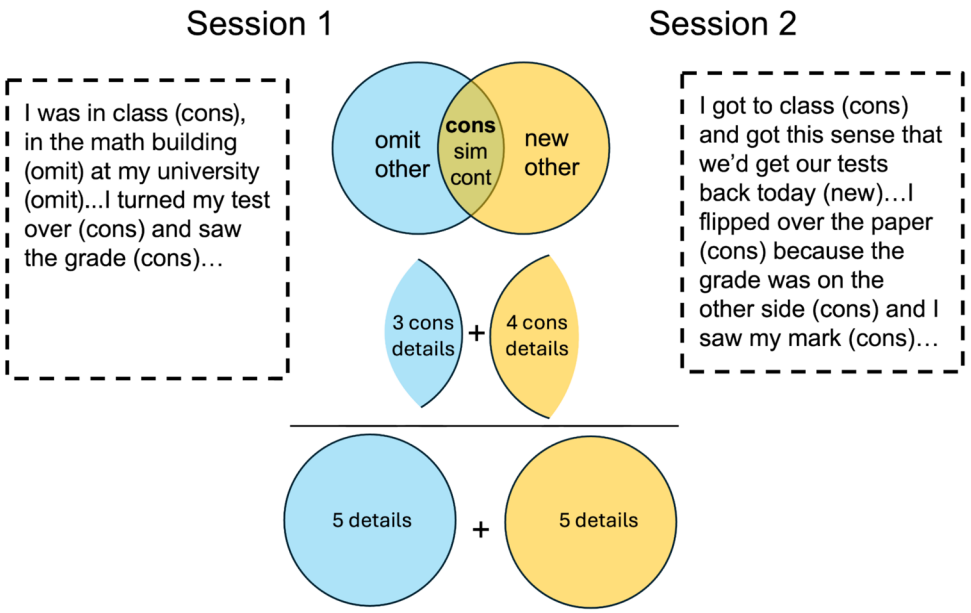

Fig. 1 Calculating AI-CONS proportions. *Note*. Schematic of proportion calculations for AI-CONS consistency categories. This memory excerpt would receive a consistency proportion of 0.70 (7 consistency details divided by 10 total details). omit = omitted details; cons = consistent details; sim = similar details; cont = contradictory details; new = new details; other = other details

Machine consistency scoring: Semantic similarity analysis

We then computed the semantic similarity of event recalls across sessions using natural language processing models. Semantic similarity analyses quantify the similarities between two texts and may be promising for use in assessing the consistency of memories over time. Notably, the term semantic here refers to conceptual similarity (as opposed to semantic memory). These models convert input text, such as a memory narrative, into a vector representation, that is, a string of numbers that capture meaningful semantic information about a text based on the training data the model was exposed to (see Harispe et al., 2015). Here, we explore three models for vector transformations; LSA (Landauer et al., 1998), DistilBERT (Sanh et al., 2020), and MPNet (Song et al., 2020). We chose these models both for their previous use in memory research (e.g., Fowler et al., 2024; Mistica et al., 2024; Ren & Coutanche, 2021; Steyvers et al., 2005; van Genugten & Schacter, 2024) and because they each adopt a unique approach to quantifying similarity.

These models are based on related, but distinct, computational approaches for mapping words to high-dimensional semantic spaces. LSA is a language model built from the TASA corpus[8] that is based on word co-occurrences; words that frequently occur together are considered semantically similar. Vectors representing texts are extracted via singular value decomposition: A matrix containing rows of words and columns of input text is constructed, and then the number of rows are compressed into a single value that preserves the relationships between columns. LSA does not account for the distinct word meanings (e.g., "bank" of river bank and financial bank would be intermixed). In contrast, DistilBERT—an advancement on BERT (or bidirectional encoder representation of transformers) with higher processing capabilities—is a masked language model. It encodes meaning for words bidirectionally to take into account the words that come before and the words that come after a target word the model is attempting to predict, generating a vector representing each word that has the context of that word embedded into the numeric representation. Once each word has been transformed, DistilBERT mean-pools word vectors into a sentence vector that represents the entire text. DistilBERT mimics output expected from BERT, which was trained on BookCorpus[9] and English Wikipedia. Finally, MPNet further builds upon BERT models by using the bidirectional encoding approach in tandem with permuted language modeling to incorporate dependency among the words being predicted when computing word vectors. In addition to BookCorpus and Wikipedia, MPNet is trained on news articles and web content. Though MPNet tends to be more accurate than DistilBERT, DistilBERT has lower computational cost than MPNet, which can make it an appealing model for researchers working with large datasets.

Once a vector representation of the text is computed by any of these models, the vector can be mapped onto a high dimensional space. The similarity between two texts can then be calculated

[8] The TASA corpus consists of 37,600 text samples from ungraded grade school and college level English documents.

[9] BookCorpus is a collection of over 7,000 free, self-published books.

by computing the cosine similarity, which entails taking the cosine of the angle between the two vectors. A cosine can range from −1 to +1, and the smaller the angle (i.e., the closer a cosine gets to 1), the more similar the text. The contextualized nature of DistilBERT and MPNet vectors may better lend themselves to identifying consistent content across memory recalls. Alternatively, LSA may be more sensitive to deviations between memory recalls due to its reliance on syntactic structure.

Data analysis

We present the results of the AI-CONS and semantic similarity analyses in quantifying memory consistency. As there is no standardized approach to measuring consistency of autobiographical memory narratives in the field to date, to assess the validity of these tools, we examined the relationship between the age of the memory and each consistency score, as memories have been found to stabilize after an initial consolidation window (see Hirst et al., 2015; Gilboa & Moscovitch, 2021).[10] Specifically, using the R package *rmcorr* to compute repeated-measure correlations (Bakdash & Marusich, 2017), we tested the hypothesis that memories that were older (e.g., more remote events) at *time 1* would be more consistent. We further used repeated-measures correlations to assess the ability of these tools to discriminate memory consistency per se, by examining whether the total number of Autobiographical Interview details provided across sessions were related to consistency measures. Finally, we compared whether differences observed between the different consistency measures significantly differed between each other using the Williams method, which we calculated in R using *r.test* from the Psych package (Revelle, 2023). The Williams method compares a correlation between *X* and *Y* and *X* and *Z* (using Fisher *Z*-transformed data), while accounting for both the dependency in the data (i.e., paired data) and the dependency from the overlapping variable (i.e., *X*).

Results

Memory consistency

Results from the AI-CONS procedure suggested that, on average, details recalled were 33.6% consistent across sessions (95% CI [32.7%, 34.5%]). Inconsistencies were largely observed due to the omission of details provided at the initial recall ($M = 38.4\%$, 95% CI [37.5%, 39.3%]) and the addition of new details at follow-up ($M = 19.5\%$, 95% CI [18.8%, 20.2%]). In contrast, on average, only 1.4% (95% CI [1.2%, 1.6%]) of details directly contradicted time 1 details at time 2, and 3.9% (95% CI [3.6%, 4.2%)] were similar. This general pattern was observed across Autobiographical Interview episodic detail subtypes, though nuances emerged (see Fig. 2). On average, there were proportionally fewer consistent, and more new emotion/thought and perceptual details. Further, place details show some evidence of high consistency and time details show some signs of high omission, though due to floor effects of these detail types—place details ranged from 0 to 11 ($Mdn = 1$) and time details ranged from 0 to 7 ($Mdn = 0$)—we interpret these with caution.

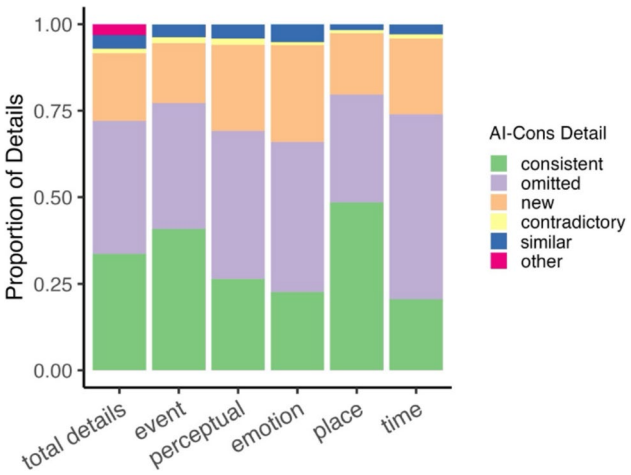

Fig. 2 Consistency of Autobiographical Interview sub-detail types. *Note.* Distribution of average AI-CONS details within each Autobiographical Interview detail category. *Total details* reflect the summation of internal and external details in the Autobiographical Interview, while *event*, *perceptual, emotion, place,* and *time* details reflect Autobiographical Interview sub-details of Autobiographical Interview internal details specifically. We note that place (0–11; $Mdn = 1$) and time (0–7; $Mdn = 0$) details were restricted in number and should be interpreted cautiously

On average, LSA computed intrapersonal memory consistency as having an average cosine similarity of 0.87 (95% CI [0.87, 0.88]). DistilBERT computed memories as having an average cosine similarity of 0.70 (95% CI [0.70, 0.71]). Finally, MPNet computed memories as having an average cosine similarity of 0.72 (95% CI [0.71, 0.73]). Mathematically, a proportion and a cosine similarity score are very different constructs, and so direct comparison of average ratings across the AI-CONS and natural language models is limited. Still, numerically, natural language models seem to be judging memories to be highly consistent, whereas this was not so for human scorers. Average cosine

[10] From a validity point of view, we might not expect memory consistency measures to converge with performance on laboratory-based memory tasks, as performance on autobiographical memory tasks is consistently found to differ from laboratory-based memory performance (Diamond et al., 2020a, 2020b; Gilboa, 2004; LePort et al., 2017; Palombo et al., 2015). Further, different characteristics of autobiographical memory do not always predict another: for example, the phenomenology of a memory can be starkly dissociated from the consistency of mnemonic details (Talarico & Rubin, 2003, 2007). Therefore, here we opted to explore the age of memories and memory detailedness in validity analyses.

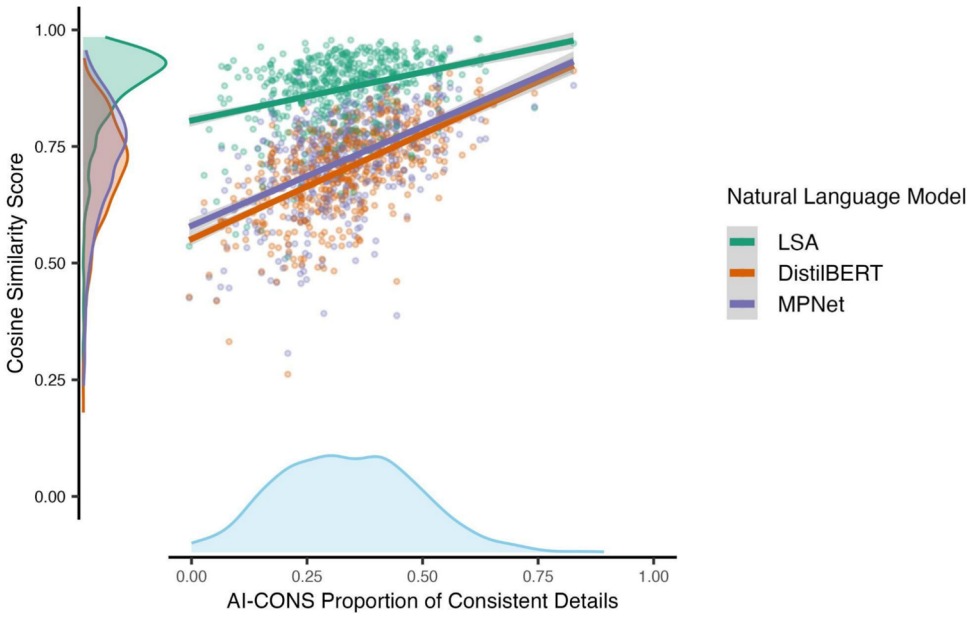

Fig. 3 Comparing the AI-CONS and automated scoring approaches. *Note*. Figure 3 presents the linear relationship between the proportion of consistent details hand-scored using the AI-CONS and similarity scores computed by LSA, DistilBERT, and MPNet, averaging events within participants for readability. Of note, for readability, we have truncated the *y*-axis to range from 0 to 1; however, cosine similarity scores can range from − 1 to 1. Plotted values are on the full, not rescaled, scales

similarity scores were high on their respective scale of − 1 to + 1, at or above 0.70, while the average proportion of AI-CONS consistent details, at 0.34, was modest to low on its respective scale of 0 to + 1 (see Fig. 3). Rescaling cosine similarity scores from a scale of − 1 to + 1 to a scale of 0 to + 1 increased the discrepancies between the proportion of AI-CONS consistent details and the semantic similarity scores,[11] with average LSA scores increasing to 0.94, DistilBERT to 0.85, and MPNet to 0.86.

Memory age and consistency

In line with our hypothesis, we found that the older a memory was at time 1, the greater the proportion of AI-CONS consistent details ($r(512) = 0.15$, $p < 0.001$, 95% CI [0.07, 0.24]),[12] though the effect was small. We observed a similar relationship between memory age and LSA ($r(512) = 0.18$, $p < 0.001$, 95% CI [0.94, 0.26]) and MPNet ($r(512) = 0.12$, $p = 0.008$, 95% CI [0.03, 0.20]) similarity scores. However, we did not find the expected relationship between the age of the initial recall and similarity scores computed with DistilBERT ($r(512) = 0.08$, $p = 0.076$, 95% CI [− 0.01, 0.16]). DistilBERT may not be a sensitive enough measure of memory consistency to identify this relationship: however, all other approaches to measuring memory consistency adequately captured the expected association.

Memory details and consistency

AI-CONS consistent details were related to the detailedness of the memories, defined as the total number of Autobiographical Interview details provided across sessions, though the effect size was small ($r(512) = 0.17$, $p < 0.001$, 95% CI [0.09, 0.25]). Cosine similarity scores were also related to the detailedness of the memories, with small to moderate effects observed with LSA ($r(512) = 0.22$, $p < 0.001$, 95% CI [0.13, 0.30]), DistilBERT ($r(512) = 0.26$, $p < 0.001$, 95% CI [0.18, 0.34]), and MPNet ($r(512) = 0.29$, $p < 0.001$, 95% CI [0.21, 0.37]). The relationship between AI-CONS consistent details and memory detailedness was significantly smaller than the relationship observed with DistilBERT ($t(511) = 2.05$, $p = 0.041$) and MPNet ($t(511) = 2.52$, $p = 0.012$), but not LSA ($t(511) = 0.83$, $p = 0.410$). While we might expect that the more mnemonic content provided, the greater the likelihood of providing

[11] Cosine similarity scores were rescaled using the formula $(X - X_{min}/X_{range}) \times n$, where X is the original cosine similarity score, X_{min} the minimum observed value on the original scale (i.e., − 1), X_{range} the range on the original scale (i.e., 2), and n the upper limit of the rescaled variable (i.e., 1).

[12] Degrees of freedom in repeated-measure correlations are calculated as $N(k-1)-1$ (Bakdash & Marusich, 2017), where N is the number of participants and k the number of repeated measures. Here, we had 513 participants recalling two events each; thus, degrees of freedom were calculated as $513(2-1)-1$.

consistent content, this analysis suggests that, generally, the AI-CONS may be better suited in discriminating consistency from memory detailedness than natural language models, particularly those sensitive to linguistic context like DistilBERT and MPNet.

Comparing humans and machines

Interestingly, despite humans seeming to score consistency more conservatively than natural language models, AI-CONS scoring was moderately correlated with DistilBERT, ($r(512) = 0.46$, $p < 0.001$, 95% CI [0.39, 0.53]) and MPNet ($r(512) = 0.39$, $p < 0.001$, 95% CI [0.32, 0.46]), though the correlation between LSA and AI-CONS scoring was smaller ($r(512) = 0.21$, $p < 0.001$, 95% CI [0.13, 0.29]). Further, the correlation between AI-CONS scoring and LSA was significantly smaller than the correlations observed between AI-CONS and DistilBERT ($t(511) = 5.35$, $p < 0.001$) and MPNet ($t(511) = 3.74$, $p < 0.001$). Human scoring showed similar rank order with DistilBERT ($r(512) = 0.46$, $p < 0.001$, 95% CI [0.39, 0.53]) and MPNet ($r(512) = 0.39$, $p < 0.001$, 95% CI [0.32, 0.47]), with a smaller relationship observed between LSA and AI-CONS scoring ($r(512) = 0.27$, $p < 0.001$, 95% CI [0.18, 0.34]).[13] These findings suggest some agreement between the AI-CONS and bidirectional encoding models.

Based on these results, we hypothesized that natural language models may be liberal in assessing consistency because the model may be judging consistency based on thematic similarity between the narratives. For example, take the following statements generated by the researchers:

Initial recall: "We won the finals for our soccer league."
Follow-up recall: "My soccer team lost in the finals."

AI-CONS scorers would tag the follow-up detail as contradictory. The soccer team could not have both won the final game and lost the final game. However, LSA rates these statements as having a cosine of 0.35, DistilBERT as 0.66, and MPNet as 0.70. After all, both statements are about a final soccer match, and holistically considering all the statements that could have possibly been provided, they are quite similar despite detailing very different accounts of what happened. While contradictory statements are rare in our narratives, this illustrates the thematic similarity that natural language models may be capturing in similarity scores that human scorers are not scoring as consistent in the AI-CONS.

We wondered whether natural language models, particularly models using bidirectional encoding such as DistilBERT and MPNet, were being conflated by external Autobiographical Interview details, that is, details that might offer context to a narrative but provide generic, thematically similar verbiage with the potential to inflate similarity scores. To assess this possibility, we removed external details from memory narratives so that only internal details remained. We then compared cosine similarity scores of narratives that contained only internal details with the proportion of consistent internal details hand-scored. A total of two participants had no internal details in one of their memories, leaving 511 participants and 1,022 events for this analysis. Average cosine similarity scores remained numerically high for LSA ($M = 0.86$, 95% CI [0.86, 0.87]), DistilBERT ($M = 0.69$, 95% CI [0.68, 0.70]), and MPNet ($M = 0.70$, 95% CI [0.69, 0.71]), and the proportion of consistent details scored with the AI-CONS remained moderate ($M = 0.37$, 95% CI [0.36, 0.38]). Numerically, we observed a slight weakening of the positive relationship between the two approaches, such that the size of the correlation was diminished for LSA ($r(510) = 0.20$, $p < 0.001$, 95% CI [0.12, 0.28]), DistilBERT ($r(510) = 0.42$, $p < 0.001$, 95% CI [0.34, 0.49]), and MPNet ($r(510) = 0.38$, $p < 0.001$, 95% CI [0.31, 0.45]). It seems that external details are not misleading natural language models. Instead, similar themes across internal and external details may be driving comparatively high similarity scores in this automated approach.

We then wondered whether we would see similarly high cosine similarity scores when comparing omitted details from session 1 narratives and new details from session 2 narratives. We theorized that, if natural language models were being inflated by thematic similarity, comparing omitted and new details would result in high similarity scores despite capturing details identified as not being present across recalls by human scores (i.e., omitted details were not observed in follow-up recalls, and new details were not observed in initial recalls). A total of 17 participants had no omitted details in their initial memory(s) or no new details in their follow-up memory(s), leaving 496 participants and 992 events for this analysis. As expected, we observed a numeric decrease in similarity scores across automated methods, yet we found the cosine similarity scores remained relatively high when computed with LSA ($M = 0.74$, 95% CI [0.73, 0.75]), DistilBERT($M = 0.48$, 95% CI [0.47, 0.49]), and MPNet ($M = 0.50$, 95% CI [0.49, 0.51]).

[13] To calculate rank order while taking into account our repeated-measure design, we assigned a rank order for cosine similarity scores and the proportion of AI-CONS consistency details and proceeded to calculate a repeated-measure correlation comparing the two rank orders with the rmcorr package. For completeness, we also calculated a traditional Spearman correlation by first averaging consistency and similarity scores within participants, and the pattern of results did not change.

Discussion

Here, we have described standardized approaches to measuring the consistency of autobiographical memory narratives over time. We demonstrated how a novel hand-scoring procedure, the AI-CONS, is implemented and presented data to suggest that recent events are recalled with approximately 33% consistency a few months later. We further demonstrated how the AI-CONS compares to natural language models with the potential to automate scoring, namely LSA, DistilBERT, and MPNet. Though AI-CONS and natural language models were moderately correlated with one another, we found that, on average, humans provided a conservative rating whereas natural language models numerically provided a liberal consistency rating. In the following, we explore our findings and outline best practices for measuring autobiographical memory consistency using these two approaches.

Summary of the AI-CONS

The AI-CONS allows for the nuanced examination of the types of details that change in memory over time, including specific episodic content as defined by Levine and colleagues' (2002) Autobiographical Interview. The AI-CONS further identifies how inconsistencies in memory recall develop over time, by classifying not only consistent details but also omitted, new, contradictory, and similar details.

Omitted details were very common in our dataset: 64.0% of details in session 1 recalls were omitted. Omitted details may reflect a type of forgetting, in that content previously accessible during free recall is not accessed at a later time. Still, that details were not freely recalled at a follow-up session does not mean they are unavailable to the rememberer. Similarly, much of the content provided at follow-up (45.1% of details in session 2 recalls) was new information not mentioned in the initial recall. New content may reflect fabrications or elaborations that have been embedded into the memory over time. However, new details could also reflect content not mentioned previously or task demands (see Gurguryan et al., 2024). Directed probing of certain details when collecting recalls would likely help to more thoroughly extract mnemonic content available to the rememberer at each session, offering researchers stronger evidence that omitted details reflect forgotten content and new details reflect embellished content. Importantly, although participants were instructed to recall the same event in the same way at both sessions, differences in an individual's goal or motivational state as well as the retrieval context shape what details are retrieved (Kensinger & Ford, 2020). Future research examining consistency across more than two time points will be interesting to explore, as details may weave in and out of recalls depending on factors such as retrieval demands and cues (see Hirst et al., 2009, 2015).

Direct probing can also reveal more inconsistencies in memory (Fisher et al., 2009). Here, we found direct contradictions to be uncommon in free recall (also see Wardell et al., 2023), which suggests that our flexible memory system, though far from perfect, has safeguards in place against introducing completely fictitious content. This may be because we fill in episodic gaps with our best-guess inference for what likely occurred based on our largely valid schemas of how the world works (Brewin et al., 2020; Fivush & Grysman, 2022; Reyna et al., 2016; Wardell & Palombo, 2024). Though research has shown that misinformation can be implanted in memory (Murphy et al., 2019; Gabbert et al., 2004; Jalbert et al., 2021; Shaw & Porter, 2015; see Loftus, 2005, for a review), details tangential to the main storyline of an event seem to be more prone to misinformation than the central components of what occurred (Kaplan et al., 2016; Putnam et al., 2017; Sharma et al., 2023). The core content of memory may be a largely reliable reference of our experiences. Still, the large number of new and omitted details in our data indicate that memory recalls change considerably, at least within the first few months following encoding.

Consistent, new, and omitted details can be scored reliably across a number of researchers, while contradictory and similar details are not only less common in narrative recalls of past events, but also relatively nuanced detail categories that may pose more challenges in achieving acceptable inter-rater reliability at the level of single raters (also see Wardell et al., 2023). Still, these detail types seem to capture qualitatively distinct concepts unaccounted for by other categories. For example, in a prior dataset used in developing the AI-CONS procedure, one memory we collected of early days of the COVID-19 lockdown involved the following:

> *Initial recall:* "I remember how quiet the city was, no one out to grab dinner with friends or shop on the mile—just empty streets and dark windows and me."
> *Follow-up recall:* "The streets were deserted, and all the windows were dark—I felt very alone in this gigantic city."

While deserted streets and dark windows were consistently recalled over time (earning two consistent details), the elaboration of feeling very alone was not explicitly discussed in the initial recall, leaving it inappropriate to categorize as a consistent detail. Still, themes of loneliness are not completely missing from the initial recall, as such a new detail would also be inappropriate. The similar category offers a detail that appropriately represents the consistency of this mnemonic content over time (see https://osf.io/msc9n for further examples). Nevertheless, depending on research questions,

goals, and resources (e.g., number of scorers, time available to train), inclusion of similar and contradictory details may not be practical. In cases when the precision of similar and contradictory details are not of interest, we recommend collapsing similar and consistent detail categories, and contradictory and new or omitted detail categories, depending on the session. Alternatively, there may be some studies that lend themselves to inclusion of this level of precision, such as projects with more directed questioning than free recall procedures (e.g., use of specific probing in the Autobiographical Interview). Some research questions, such as the study of eyewitness testimonies or work with patient samples, may warrant high precision in the measure and require the use of similar details. In such cases, employing protocols that lend themselves to average rater intraclass coefficients (ICCs) may be needed to obtain acceptable reliability.

Human versus machine scoring

The AI-CONS procedure was developed by synthesizing theory and past scoring schemes present in autobiographical memory, flashbulb memory, and eyewitness testimony research to provide the field with a standardized approach to measuring memory consistency. Semantic similarity analyses have been developed to aid humans in identifying the relatedness of units of language (e.g., words, sentences; see Harispe et al., 2015). Investigation of the relationship between consistency, memory age, and memory detailedness in the present study provides initial indication of the validity of both the AI-CONS and natural language processing approaches in capturing memory consistency. Specifically, the consistency of memories as measured by the AI-CONS, MPNet, and LSA (though not DistilBERT) seems to stabilize over time, a pattern observed in past examinations of memory consistency (e.g., Hirst et al., 2015). The detailedness of a memory was weakly correlated with our consistency measures, suggesting that, though memory consistency and memory detailedness are related, they are distinct constructs. Scores computed by natural language processing models showed greater overlap with memory detailedness relative to the AI-CONS, suggesting that the AI-CONS may be a more sensitive measure. Still, overall, the AI-CONS and semantic similarity analyses were moderately correlated, which suggests that both tools are converging towards similar constructs, namely the shared overlap between two narrative recalls.

The time and labor demands of human scoring procedures such as the AI-CONS are often not feasible, and here we have shown that natural language models offer a promising alternative. We found that semantic similarity scores, particularly those computed from bidirectional encoding models, had a positive relationship with the proportion of AI-CONS consistency details, with a moderate effect size. Still, when considering the numeric value on each of their respective scales, on average, natural language models showed high consistency of memories whereas human scorers showed moderate to low consistency. As cosine similarity coefficients and proportions are distinct mathematical constructs, direct comparison of the two approaches is limited, though it is still useful to consider where each measure falls on their respective scales. Automated approaches seem to be apt at capturing thematic similarity across texts. When participants recall the same event, related words will contribute to higher semantic similarity scores, even when exact meaning may differ between texts. Human scorers seem to be better at identifying a lack of consistency between specific details, perhaps offering more nuance and specificity. For researchers interested in capturing nuanced mnemonic changes, such as shifts in episodic content, differences among clinical populations where sample sizes may be more modest, or what drives inconsistency in a memory (e.g., new or omitted details), employing the AI-CONS is recommended. However, in instances when assessing gestalt similarity in large datasets is a priority, natural language models offer an accessible proxy for measures of consistency. The AI-CONS seems to have more face validity in identifying how consistent a memory is, yet our data also suggest that similarity analyses are sensitive to reductions in consistency. For example, while average similarity scores were still moderately high when comparing only new and omitted details across sessions, similarity scores were still lower than they were when all details were included. Thus, although detail-by-detail consistency may be too nuanced for current natural language models, these models are still able to capture a gestalt sense of memory consistency. As the field of natural language processing continues to rapidly develop, it is possible that new models or hybrid approaches may be better able to reduce labor demands of human scoring. For example, models like ChatGPT may be capable of being trained to implement protocols such as the Autobiographical Interview and the AI-CONS. Ethical considerations of working with local versus cloud-based programs and feeding these models participants' personal memories will be important for researchers to keep in mind.

In considering how human and machine scoring compares, it is also important to note that human scoring is not perfect. Assessing reliability mitigates concerns of human error in scoring, and here we show that consistent, new, and omitted details can be reliably scored across a number of researchers. Nevertheless, it is still expected that scoring disagreements and errors will be encountered. We have found that intermittently conducting reliability analyses while scorers are scoring a dataset encourages accurate scoring, particularly when scorers are blind to which memories are used in reliability analyses. Notably, to honor the results of our ICCs, our laboratory opts to not correct narratives when we identify details that we may disagree with.

Future directions in memory consistency research

Autobiographical research has often conceptualized "good" memory as memories that are highly detailed, particularly in episodic content (see Addis & Szpunar, 2024, for discussion), and detailed memories are rated as more accurate and trustworthy (Bastin et al., 2022). Yet our data suggest that a very detailed memory is not necessarily a consistent one. While more detailed memories tended to be more consistent in our data, the effect size with AI-CONS consistent details was small ($r = 0.17$), suggesting that the two concepts have dissociable features and warrant independent investigation. When assessing consistency, researchers should be aware that timing matters: The more remote a memory is at time 1, the less change is likely to be observed at follow-up, at least with retention intervals of a few months. This is likely, at least in part, because memories tend to become more general, gist-like stories over time (Boccia et al., 2019; Grilli et al., 2019; Robin & Moscovitch, 2017; Rudoy et al., 2009; Wardell & Palombo, 2024). With this in mind, it will be important for researchers to control for or match conditions on the age of the event at initial recall to ensure effects observed are associated with experimental manipulations and not the passage of time.

By proposing a standardized approach to assessing consistency in memories, we hope to aid the field in pursuing a range of important research questions. There may be a variety of factors that shape the consistency of autobiographical memories, including event characteristics, time, and individual differences. Our standardized approach to measuring the consistency of autobiographical memory narratives opens exciting avenues for future research to explore such boundary conditions. It will also be interesting for future work to explore changes in memory across multiple time points (e.g., Hirst et al., 2015), either by expanding the AI-CONS to categorize types of (in)consistencies across three or more recalls or by employing semantic similarity analyses. Beyond *intra*personal consistency, we can also consider *inter*personal consistency in narrative recall, examining how shared experiences are remembered similarly or divergently between individuals, be it romantic partners, parent–child dyads, or strangers (see Maswood et al., 2019; Lee et al., 2020; also see Fowler et al., 2024). Finally, assessing memory consistency as a function of the rememberer's audience offers an important way to consider how retrieval demands inform the mnemonic content we bring to mind (e.g., Momennejad et al., 2019). The AI-CONS offers a standardized, systematic approach to exploring these and related research questions by identifying the types of details that change in a memory over time and how that change occurs.

Conclusion

Our autobiographical memories are our reference point for understanding how our past has led to where we are in the present. Yet our memories change, shifting with the passage of time, to meet retrieval demands, or to best fit with our current knowledge. Examining the consistency with which we remember our past offers us an important window into memory malleability. While capturing change in naturalistic contexts is complicated, the AI-CONS and large language models offer exciting ways for researchers to examine changes in naturalistic narrative recall of autobiographical experiences. The AI-CONS offers a standardized approach to assessing nuanced mnemonic differences, including changes in episodic content. This approach may be particularly important for work examining differences among clinical populations (e.g., amnesia, trauma) or in eyewitness testimony work where nuanced changes in content are highly relevant. In other instances when gestalt or thematic similarity is of interest, machine scoring may offer an accessible method for large datasets that can help guide more specific questions that warrant the time and labor of human scoring.

Acknowledgements The authors thank Deniz Basakci, Yuxan Chen, Sabrina C. Co, Joaquin Gomez de la Torre, Emma Haight, Sarah Lacusta, Kim Marty, Khushi Sharma, and Wendi Zhong for scoring memory narratives and Oliver J. Bontkes for double-checking all the statistics.

Funding D.J.P. is supported by an NSERC Discovery [grant number RGPIN-2019–04596] and the John R. Evans Leaders Fund from the Canada Foundation for Innovation [grant number 38817], V.W. is supported by a SSHRC Doctoral Award, a Killam Memorial Doctoral Scholarship, and a Killam Donald N. Byers Memorial Prize.

Data availability Data and the AI-CONS training protocol can be found at https://osf.io/uat5w/.

Code availability Analysis code can be found at https://osf.io/uat5w/.

Declarations

Ethics approval Study procedures were approved by the University of British Columbia's Research Ethics Board (H18-02583).

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication Participants signed informed consent regarding publishing their data, including memory narratives.

Conflicts of interest/competing interests The authors declare no conflicts of interest, including no financial conflicts of interest with regard to any software products cited in this paper.

References

Addis, D. R., & Szpunar, K. K. (2024). Beyond the episodic–semantic continuum: The multidimensional model of mental representations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1913). https://doi.org/10.1098/rstb.2023.0408

Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00456

Barclay, C. R., & Wellman, H. M. (1986). Accuracies and inaccuracies in autobiographical memories. *Journal of Memory and Language, 25*(1), 93–103. https://doi.org/10.1016/0749-596x(86)90023-9

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

Barzykowski, K., & Staugaard, S. R. (2016). Does retrieval intentionality really matter? Similarities and differences between involuntary memories and directly and generatively retrieved voluntary memories. *British Journal of Psychology, 107*(3), 519–536. https://doi.org/10.1111/bjop.12160

Bastin, C., Folville, A., & Geurten, M. (2022). "I trust you if your memory is detailed": Interpersonal memory fidelity judgments and social bonding. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4303218

Baugerud, G. A., Magnussen, S., & Melinder, A. (2014). High accuracy but low consistency in children's long-term recall of a real-life stressful event. *Journal of Experimental Child Psychology, 126*, 357–368. https://doi.org/10.1016/j.jecp.2014.05.009

Bluck, S. (2003). Autobiographical memory: Exploring its functions in everyday life. *Memory, 11*(2), 113–123. https://doi.org/10.1080/741938206

Boccia, M., Teghil, A., & Guariglia, C. (2019). Looking into recent and remote past: Meta-analytic evidence for cortical re-organization of episodic autobiographical memories. *Neuroscience & Biobehavioral Reviews, 107*, 84–95. https://doi.org/10.1016/j.neubiorev.2019.09.003

Bohannon, J. N., III. (1988). Flashbulb memories for the space shuttle disaster: A tale of two theories. *Cognition, 29*(2), 179–196. https://doi.org/10.1016/0010-0277(88)90036-4

Boyd, B. (2018). The evolution of stories: From mimesis to language, from fact to fiction. *WIREs Cognitive Science, 9*(1), e1444. https://doi.org/10.1002/wcs.1444

Brewin, C. R., Andrews, B., & Mickes, L. (2020). Regaining consensus on the reliability of memory. *Current Directions in Psychological Science, 29*(2), 121–125. https://doi.org/10.1177/0963721419898122

Brown, R., & Kulik, J. (1977). Flashbulb memories. *Cognition, 5*(1), 73–99. https://doi.org/10.1016/0010-0277(77)90018-x

Campbell, J., Nadel, L., Duke, D., & Ryan, L. (2011). Remembering all that and then some: Recollection of autobiographical memories after a 1-year delay. *Memory, 19*(4), 406–415. https://doi.org/10.1080/09658211.2011.578073

Conway, M. A., Justice, L. V., & D'Argembeau, A. (2019). The self-memory system revisited. In J. Mace (Ed.), *The organization and structure of autobiographical memory* (pp. 28–51). Oxford University Press. https://doi.org/10.1093/oso/9780198784845.003.0003

De Brigard, F., Giovanello, K. S., Stewart, G. W., Lockrow, A. W., O'Brien, M. M., & Spreng, R. N. (2016). Characterizing the subjective experience of episodic past, future, and counterfactual thinking in healthy younger and older adults. *Quarterly Journal of Experimental Psychology, 69*(12), 2358–2375. https://doi.org/10.1080/17470218.2015.1115529

Dev, D. K., Wardell, V., Checknita, K. J., Te, A. A., Petrucci, A. S., Le, M. L., Madan, C. R., & Palombo, D. J. (2022). Negative emotion enhances memory for the sequential unfolding of a naturalistic experience. *Journal of Applied Research in Memory and Cognition, 11*(4), 510–521. https://doi.org/10.1037/mac0000015

Devitt, A. L., Monk-Fromont, E., Schacter, D. L., & Addis, D. R. (2015). Factors that influence the generation of autobiographical memory conjunction errors. *Memory, 24*(2), 204–222. https://doi.org/10.1080/09658211.2014.998680

Diamond, N. B., Abdi, H., & Levine, B. (2020a). Different patterns of recollection for matched real-world and laboratory-based episodes in younger and older adults. *Cognition, 202*, 104309. https://doi.org/10.1016/j.cognition.2020.104309

Diamond, N. B., Armson, M. J., & Levine, B. (2020b). The truth is out there: Accuracy in recall of verifiable real-world events. *Psychological Science, 31*(12), 1544–1556. https://doi.org/10.1177/0956797620954812

Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology, 26*(3), 297–302. https://doi.org/10.2307/1932409

Ebbinghaus, H. (1885). *Memory: A contribution to experimental psychology*. Dover Publications.

Fisher, R. P., Brewer, N., & Mitchell, G. (2009). The relation between consistency and accuracy of eyewitness testimony: Legal versus cognitive explanations. *Handbook of Psychology of Investigative Interviewing*, 121–136. https://doi.org/10.1002/9780470747599.ch8

Fivush, R., & Grysman, A. (2022). Accuracy and reconstruction in autobiographical memory: (Re)consolidating neuroscience and sociocultural developmental approaches. *WIREs Cognitive Science*, *14*(3). https://doi.org/10.1002/wcs.1620

Fowler, Z., Palombo, D. J., Madan, C. R., & O'Connor, B. B. (2024). Collaborative imagination synchronizes representations of the future and fosters social connection in the present. *Proceedings of the National Academy of Sciences, 121*(25), e2318292121. https://doi.org/10.1073/pnas.2318292121

Gabbert, F., Memon, A., Allan, K., & Wright, D. B. (2004). Say it to my face: Examining the effects of socially encountered misinformation. *Legal and Criminological Psychology, 9*(2), 215–227. https://doi.org/10.1348/1355325041719428

Gilbert, J. A. E., & Fisher, R. P. (2006). The effects of varied retrieval cues on reminiscence in eyewitness memory. *Applied Cognitive Psychology, 20*(6), 723–739. https://doi.org/10.1002/acp.1232

Gilboa, A. (2004). Autobiographical and episodic memory—one and the same?: Evidence from prefrontal activation in neuroimaging studies. *Neuropsychologia, 42*(10), 1336–1349. https://doi.org/10.1016/j.neuropsychologia.2004.02.014

Gilboa, A., & Moscovitch, M. (2021). No consolidation without representation: Correspondence between neural and psychological representations in recent and remote memory. *Neuron, 109*(14), 2239–2255. https://doi.org/10.1016/j.neuron.2021.04.025

Goldsmith, M., Koriat, A., & Pansky, A. (2005). Strategic regulation of grain size in memory reporting over time. *Journal of Memory and Language, 52*(4), 505–525. https://doi.org/10.1016/j.jml.2005.01.010

Grilli, M. D., Coste, S., Landry, J. E., & Mangen, K. (2019). Evidence that an episodic mode of thinking facilitates encoding of perceptually rich memories for naturalistic events relative to a gist-based mode of thinking. *Memory, 27*(10), 1468–1474. https://doi.org/10.1080/09658211.2019.1657461

Gurguryan, L., Yang, H., Köhler, S., & Sheldon, S. (2024). Lifetime familiarity cue effects for autobiographical memory. *Psychological Research Psychologische Forschung, 88*(5), 1456–1470. https://doi.org/10.1007/s00426-024-01968-3

Harispe, S., Ranwez, S., Janaqi, S., & Montmain, J. (2015). *Semantic similarity from natural language and ontology analysis* (1st ed.). Springer Chan. https://doi.org/10.1007/978-3-031-02156-5

Hirst, W., Phelps, E. A., Buckner, R. L., Budson, A. E., Cuc, A., Gabrieli, J. D., Johnson, M. K., Lustig, C., Lyle, K. B., Mather, M., Meksin, R., Mitchell, K. J., Ochsner, K. N., Schacter, D. L., Simons, J. S., & Vaidya, C. J. (2009). Long-term memory for the terrorist attack of September 11: Flashbulb Memories, event memories, and the factors that influence their retention. *Journal of Experimental Psychology: General, 138*(2), 161–176. https://doi.org/10.1037/a0015527

Hirst, W., Phelps, E. A., Meksin, R., Vaidya, C. J., Johnson, M. K., Mitchell, K. J., Buckner, R. L., Budson, A. E., Gabrieli, J. D., Lustig, C., Mather, M., Ochsner, K. N., Schacter, D., Simons, J. S., Lyle, K. B., Cuc, A. F., & Olsson, A. (2015). A ten-year follow-up of a study of memory for the attack of September 11, 2001: Flashbulb Memories and memories for flashbulb events. *Journal of Experimental Psychology: General, 144*(3), 604–623. https://doi.org/10.1037/xge0000055

Jalbert, M. C., Wulff, A. N., & Hyman, I. E. (2021). Stealing and sharing memories: Source monitoring biases following collaborative remembering. *Cognition, 211*, 104656. https://doi.org/10.1016/j.cognition.2021.104656

Kaplan, R. L., Van Damme, I., Levine, L. J., & Loftus, E. F. (2016). Emotion and false memory. *Emotion Review, 8*(1), 8–13. https://doi.org/10.1177/1754073915601228

Kensinger, E. A., & Ford, J. H. (2020). Retrieval of emotional events from memory. *Annual Review of Psychology, 71*(1), 251–272. https://doi.org/10.1146/annurev-psych-010419-051123

Kensinger, E. A., & Schacter, D. L. (2006). When the Red Sox shocked the Yankees: Comparing negative and positive memories. *Psychonomic Bulletin & Review, 13*(5), 757–763. https://doi.org/10.3758/BF03193993

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Kopelman, M. D., Wilson, B. A., & Baddeley, A. D. (1989). The Autobiographical Memory Interview: A new assessment of autobiographical and personal semantic memory in amnesic patients. *Journal of Clinical and Experimental Neuropsychology, 11*(5), 724–744. https://doi.org/10.1080/01688638908400928

Koriat, A., Goldsmith, M., & Pansky, A. (2000). Toward a psychology of memory accuracy. *Annual Review of Psychology, 51*, 481–537. https://doi.org/10.1146/annurev.psych.51.1.481

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2–3), 259–284. https://doi.org/10.1080/01638539809545028

Lee, H., Bellana, B., & Chen, J. (2020). What can narratives tell us about the neural bases of human memory? *Current Opinion in Behavioral Sciences, 32*, 111–119. https://doi.org/10.1016/j.cobeha.2020.02.007

Leport, A. K., Stark, S. M., McGaugh, J. L., & Stark, C. E. (2017). A cognitive assessment of highly superior autobiographical memory. *Memory, 25*(2), 276–288. https://doi.org/10.1080/09658211.2016.1160126

Levine, B., Svoboda, E., Hay, J. F., Winocur, G., & Moscovitch, M. (2002). Aging and autobiographical memory: Dissociating episodic from semantic retrieval. *Psychology and Aging, 17*(4), 677–689. https://doi.org/10.1037/0882-7974.17.4.677

Linton, M. (1975). Memory for real-world events. In D. A. Norman & D. E. Rumelhart (Eds.), *Explorations in cognition* (pp. 376–404). Freeman.

Lockrow, A. W., Setton, R., Spreng, K. A., Sheldon, S., Turner, G. R., & Spreng, R. N. (2024). Taking stock of the past: A psychometric evaluation of the autobiographical interview. *Behavior Research Methods, 56*(2), 1002–1038. https://doi.org/10.3758/s13428-023-02080-x

Loftus, E. F. (2003). Make-believe memories. *American Psychologist, 58*(11), 867–873. https://doi.org/10.1037/0003-066x.58.11.867

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*(4), 361–366. https://doi.org/10.1101/lm.94705

Luchetti, M., & Sutin, A. R. (2015). Measuring the phenomenology of autobiographical memory: A short form of the memory experiences questionnaire. *Memory, 24*(5), 592–602. https://doi.org/10.1080/09658211.2015.1031679

Marcotti, P., & St. Jacques, P. L. (2018). Shifting visual perspective during memory retrieval reduces the accuracy of subsequent memories. *Memory*, *26*(3), 330–341https://doi.org/10.1080/09658211.2017.1329441

Maswood, R., Rasmussen, A. S., & Rajaram, S. (2019). Collaborative remembering of emotional autobiographical memories: Implications for emotion regulation and collective memory. *Journal of Experimental Psychology: General, 148*(1), 65–79. https://doi.org/10.1037/xge0000468

McAdams, D. P. (2008). Personal narratives and the life story. In O. P. John, R. W. Robins, & L. A. Pervin (Eds.), *Handbook of personality: Theory and research* (3rd ed., pp. 242–262). The Guilford Press.

McKinnon, M. C., Palombo, D. J., Nazarov, A., Kumar, N., Khuu, W., & Levine, B. (2015). Threat of death and autobiographical memory: A study of passengers from Flight AT236. *Clinical Psychological Science, 3*(4), 487–502. https://doi.org/10.1177/2167702614542280

Melega, G., Lancelotte, F., Ann-Kathrin, N., Hornberger, M., Levine, B., & Renoult, L. (2023). Evoking episodic and semantic details with instructional manipulation: The Semantic Autobiographical Interview. *PsyArXiv Preprints*. https://doi.org/10.31234/osf.io/6h23t

Misra, P., Marconi, A., Peterson, M., & Kreiman, G. (2018). Minimal memory for details in real life events. *Scientific Reports, 8*, 16701. https://doi.org/10.1038/s41598-018-33792-2

Mistica, M., Haylock, P., Michalewicz, A., Raad, S., Fitzgerald, E., & Hitchcock, C. (2024). A natural language model to automate scoring of autobiographical memories. *Behavior Research Methods, 56*(7), 6707–6720. https://doi.org/10.3758/s13428-024-02385-5

Momennejad, I., Duker, A., & Coman, A. (2019). Bridge ties bind collective memories. *Nature Communications*, *10*(1). https://doi.org/10.1038/s41467-019-09452-y

Murphy, G., Loftus, E. F., Grady, R. H., Levine, L. J., & Greene, C. M. (2019). False memories for fake news during Ireland's abortion referendum. *Psychological Science, 30*(10), 1449–1459. https://doi.org/10.1177/0956797619864887

Nadel, L., Campbell, J., & Ryan, L. (2007). Autobiographical memory retrieval and hippocampal activation as a function of repetition and the passage of time. *Neural Plasticity, 2007*, 1–14. https://doi.org/10.1155/2007/90472

Neisser, U. (1981). John Dean's memory: A case study. *Cognition, 9*(1), 1–22. https://doi.org/10.1016/0010-0277(81)90011-1

Neisser, U., & Harsch, N. (1992). Phantom flashbulbs: False recollections of hearing the news about challenger. *Affect and Accuracy in Recall*, 9–31. https://doi.org/10.1017/cbo9780511664069.003

Nielsen, N. P., Gehrt, T. B., & Berntsen, D. (2023). Individual differences in autobiographical memory predict memory confidence but not memory accuracy. *Journal of Applied Research in Memory and Cognition, 12*(4), 542–551. https://doi.org/10.1037/mac0000082

Odinot, G., Memon, A., La Rooy, D., & Millen, A. (2013). Are two interviews better than one? Eyewitness memory across repeated

cognitive interviews. *PLoS ONE*, *8*(10). https://doi.org/10.1371/journal.pone.0076305

Orbach, Y., Lamb, M. E., La Rooy, D., & Pipe, M. (2012). A case study of witness consistency and memory recovery across multiple investigative interviews. *Applied Cognitive Psychology, 26*(1), 118–129. https://doi.org/10.1002/acp.1803

Palombo, D. J. (2024). Beyond memory: The transcendence of episodic narratives. *Canadian Journal of Experimental Psychology / Revue Canadienne De Psychologie Expérimentale, 78*(3), 155–162. https://doi.org/10.1037/cep0000345

Palombo, D. J., Alain, C., Söderlund, H., Khuu, W., & Levine, B. (2015). Severely Deficient Autobiographical Memory (SDAM) in healthy adults: A new mnemonic syndrome. *Neuropsychologia, 72*, 105–118. https://doi.org/10.1016/j.neuropsychologia.2015.04.012

Palombo, D. J., Sheldon, S., & Levine, B. (2018). Individual differences in autobiographical memory. *Trends in Cognitive Sciences, 22*(7), 583–597. https://doi.org/10.1016/j.tics.2018.04.007

Piolino, P., Desgranges, B., & Eustache, F. (2009). Episodic autobiographical memories over the course of time: Cognitive, neuropsychological and neuroimaging findings. *Neuropsychologia, 47*(11), 2314–2329. https://doi.org/10.1016/j.neuropsychologia.2009.01.020

Putnam, A. L., Sungkhasettee, V. W., & Roediger, H. L. (2017). When misinformation improves memory: The effects of recollecting change. *Psychological Science, 28*(1), 36–46. https://doi.org/10.1177/0956797616672268

Race, E., Keane, M. M., & Verfaellie, M. (2011). Medial temporal lobe damage causes deficits in episodic memory and episodic future thinking not attributable to deficits in narrative construction. *Journal of Neuroscience, 31*(28), 10262–10269. https://doi.org/10.1523/jneurosci.1145-11.2011

Ren, X., & Coutanche, M. N. (2021). Sleep reduces the semantic coherence of memory recall: An application of latent semantic analysis to investigate memory reconstruction. *Psychonomic Bulletin & Review, 28*, 1336–1343. https://doi.org/10.3758/s13423-021-01919-8

Renoult, L., Armson, M. J., Diamond, N. B., Fan, C. L., Jeyakumar, N., Levesque, L., Oliva, L., McKinnon, M., Papadopoulos, A., Selarka, D., St. Jacques, P. L., & Levine, B. (2020). Classification of general and personal semantic details in the Autobiographical Interview. *Neuropsychologia, 144*, 107501. https://doi.org/10.1016/j.neuropsychologia.2020.107501

Revelle, W. (2023). psych: Procedures for personality and psychological research (Version 2.4.3) [R package]. https://cran.r-project.org/web/packages/psych/index.html

Reyna, V. F., Corbin, J. C., Weldon, R. B., & Brainerd, C. J. (2016). How fuzzy-trace theory predicts true and false memories for words, sentences, and narratives. *Journal of Applied Research in Memory and Cognition, 5*(1), 1–9. https://doi.org/10.1016/j.jarmac.2015.12.003

Robin, J., & Moscovitch, M. (2017). Details, gist and schema: Hippocampal–neocortical interactions underlying recent and remote episodic and spatial memory. *Current Opinion in Behavioral Sciences, 17*, 114–123. https://doi.org/10.1016/j.cobeha.2017.07.016

Rubin, D. C., Dennis, M. F., & Beckham, J. C. (2011). Autobiographical memory for stressful events: The role of autobiographical memory in posttraumatic stress disorder. *Consciousness and Cognition, 20*(3), 840–856. https://doi.org/10.1016/j.concog.2011.03.015

Rubin, D., Schrauf, R., & Greenberg, D. (2004). Stability in autobiographical memories. *Memory, 12*(6), 715–721. https://doi.org/10.1080/09658210344000512

Rudoy, J. D., Weintraub, S., & Paller, K. A. (2009). Recall of remote episodic memories can appear deficient because of a gist-based retrieval orientation. *Neuropsychologia, 47*(3), 938–941. https://doi.org/10.1016/j.neuropsychologia.2008.12.006

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv*. https://doi.org/10.48550/arXiv.1910.01108

Schacter, D. L. (1999). The seven sins of memory: Insights from psychology and cognitive neuroscience. *American Psychologist, 54*(3), 182–203. https://doi.org/10.1037/0003-066X.54.3.182

Schacter, D. L. (2022). The seven sins of memory: An update. *Memory, 30*(1), 37–42. https://doi.org/10.1080/09658211.2021.1873391

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society b: Biological Sciences, 362*(1481), 773–786. https://doi.org/10.1098/rstb.2007.2087

Schacter, D. L., Guerin, S. A., & St. Jacques, P. L. (2011). Memory distortion: An adaptive perspective. *Trends in Cognitive Sciences, 15*(10), 467–474. https://doi.org/10.1016/j.tics.2011.08.004

Sharma, P. R., Wade, K. A., & Jobson, L. (2023). A systematic review of the relationship between emotion and susceptibility to misinformation. *Memory, 31*(1), 1–21. https://doi.org/10.1080/09658211.2022.2120623

Shaw, J., & Porter, S. (2015). Constructing rich false memories of committing crime. *Psychological Science, 26*(3), 291–301. https://doi.org/10.1177/0956797614562862

Söderlund, H., Moscovitch, M., Kumar, N., Daskalakis, Z. J., Flint, A., Herrmann, N., & Levine, B. (2014). Autobiographical episodic memory in major depressive disorder. *Journal of Abnormal Psychology, 123*(1), 51–60. https://doi.org/10.1037/a0035610

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T. (2020). MPNet: Masked and permuted pre-training for language understanding. ArXiv. https://doi.org/10.48550/arXiv.2004.09297

Stanley, S. E., & Benjamin, A. S. (2016). That's not what you said the first time: A theoretical account of the relationship between consistency and accuracy of recall. *Cognitive Research: Principles and Implications*, *1*(14). https://doi.org/10.1186/s41235-016-0012-9

Steyvers, M., Shiffrin, R. M., & Nelson, D. L. (2005). Word association spaces for predicting semantic similarity effects in episodic memory. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 237–249). American Psychological Association. https://doi.org/10.1037/10895-018

St. Jacques, P. L., & Levine, B. (2007). Ageing and autobiographical memory for emotional and neutral events. *Memory*, *15*(2), 129–144https://doi.org/10.1080/09658210601119762

Strikwerda-Brown, C., Mothakunnel, A., Hodges, J. R., Piguet, O., & Irish, M. (2019). External details revisited–A new taxonomy for coding 'non-episodic' content during autobiographical memory retrieval. *Journal of Neuropsychology, 13*(3), 371–397. https://doi.org/10.1111/jnp.12160

Talarico, J. M., & Rubin, D. C. (2003). Confidence, not consistency, characterizes flashbulb memories. *Psychological Science, 14*(5), 455–461. https://doi.org/10.1111/1467-9280.02453

Talarico, J. M., & Rubin, D. C. (2007). Flashbulb memories are special after all; in phenomenology, not accuracy. *Applied Cognitive Psychology, 21*(5), 557–578. https://doi.org/10.1002/acp.1293

Talarico, J. M., & Rubin, D. C. (2017). Ordinary memory processes shape flashbulb memories of extraordinary events: A review of 40 years of research. In O. Luminet & A. Curci (Eds.), *Flashbulb memories: New challenges and future perspectives* (2nd ed., pp. 73–95). Psychology Press. https://doi.org/10.4324/9781315623481

Thomas, A. K., & Loftus, E. F. (2002). Creating bizarre false memories through imagination. *Memory & Cognition, 30*, 423–431. https://doi.org/10.3758/BF03194942

Thomsen, D. K., Jensen, T., Holm, T., Olesen, M. H., Schnieber, A., & Tønnesvang, J. (2015). A 3.5 year diary study: Remembering and life story importance are predicted by different event characteristics. *Consciousness and Cognition, 36*, 180–195. https://doi.org/10.1016/j.concog.2015.06.011

van Genugten, R. D., & Schacter, D. L. (2024). Automated scoring of the autobiographical interview with natural language processing. *Behavior Research Methods, 56*(3), 2243–2259. https://doi.org/10.3758/s13428-023-02145-x

Verfaellie, M., Wank, A. A., Reid, A. G., Race, E., & Keane, M. M. (2019). Self-related processing and future thinking: Distinct contributions of ventromedial prefrontal cortex and the medial temporal lobes. *Cortex, 115*, 159–171. https://doi.org/10.1016/j.cortex.2019.01.028

Wardell, V., Madan., C. R., Jameson, T. J., Cocquyt, C. M., Checknita, K. J., Liu, H., & Palombo, D. J. (2021). How emotion influences the details recalled in autobiographical memory. *Applied Cognitive Psychology, 35*(6), 1454–1465. https://doi.org/10.1002/acp.3877

Wardell, V., Jameson, T., Bontkes, O. J., Le, M. L., Duan, T., St. Jacques, P. L., Madan, C. R., & Palombo, D. J. (2023). Fade in, fade out: Do shifts in visual perspective predict the consistency of real-world memories? *Psychological Science*, *34*(8), 932–946. https://doi.org/10.1177/09567976231180588

Wardell, V., & Palombo, D. J. (2024). Stability and malleability of emotional autobiographical memories. *Nature Reviews Psychology, 3*(6), 393–406. https://doi.org/10.1038/s44159-024-00312-1

Yarmey, A. D., & Bull, M. P. (1978). Where were you when President Kennedy was assassinated? *Bulletin of the Psychonomic Society, 11*(2), 133–135. https://doi.org/10.3758/bf03336788