Jeremy R. Manning
Dartmouth College
Department of Psychological & Brain Sciences
HB 6207 Moore Hall
Hanover, NH 03755

September 30 2020

To the editors of *Psychological Review*:

I have enclosed my revised manuscript entitled *Episodic memory: mental time travel or a quantum 'memory wave' function?* (submission REV-2019-1167). I appreciate the reviewers' insightful comments, and I believe the manuscript has been substantially improved as a result.

To summarize my changes, the first reviewer requested additional background and clarifying details and background, which I have added to the manuscript. The second reviewer raised some important points about whether the descriptions in the original manuscript were appropriate, how they might be clarified, and how they might be integrated with related work. They also suggested several relevant papers, and I have added to the manuscript references and discussions of that work, along with the other requested information, modifications, and clarifications.

I have included point-by-point responses to each comment on the following pages. The reviewers' comments are shown in *italics* and my responses are shown in **bold**. Thank you for considering my revised manuscript.

Sincerely,

Jeremy R. Manning
Jeremy.R.Manning@Dartmouth.edu

Reviewer #1: This is an interesting and timely theoretical piece making the argument that a prominent (possibly preeminent) verbal theory of how memory recall operates is out of step with modern computational theories of how memory recall operates. The verbal theory is mental time travel as developed by Tulving and others over many years, in which it is suggested that one can revisit a particular moment of one's past during recall or reminiscence. The computational theory, present in many models of recall, but here specifically developed in terms of retrieved-context models, suggests that recall involves reactivation of a blend of memories extending over a range of times. Given the ubiquity of MTT theory, this piece will likely be of wide interest to many psychologists, regardless of whether they specialize in memory.

The piece focuses on recent research on recall of naturalistic events, television narratives, and contrasts research in this domain with research from the list learning literature. I thought this was done quite effectively, stepping through theoretical elements of interest and contrasting how they have been applied in both domains (e.g., regarding the effects of event boundaries). Finally, the piece reviews advanced analytic techniques allowing one to examine the semantic structure of narrative recall sequences (topics models and associated techniques), in order to support the contention that recall of naturalistic experiences are best thought of as reflecting a blend of event memories distributed over time.

Thank you for the positive feedback!

I think the heart of this article is theoretically sound, and it is certainly well written. I have some comments which may help Manning clarify the theoretical position being proposed here, followed by a handful of minor comments.

(1) I think the article could use some more detail regarding how mental time travel is conceptualized in the modern literature. This would allow the author to draw a sharper distinction between MTT theory and the quantum wave function (QWF) proposal. I think it would be useful to have some clear examples from the literature where MTT is conceptualized as revisiting a specific past event. Elaborating on MTT theory as currently used could involve discussing the substantial body of literature that focuses on MTT and future planning (often referred to as pre-living an experience). Does the QWF idea inform our understanding of future planning? It strikes me that retrieving a blend of past experiences could serve as a useful template for the structure of future experiences.

I have added additional references to the Introduction and Overview (p. 2) that specifically describe the conceptualization of episodic recall as mental time travel, or mentally "jumping back in time" to the moment we experienced the event we are remembering. I have also added a new section (*Defining the* quantum memory wave function *metaphor and comparing it to the* mental time travel *metaphor*, p. 6--8) and Figure (Fig. 1) to elaborate on the MTT vs. QWF distinctions. That section also includes a paragraph elaborating on how the QWF idea may be used to conceptualize how we

predict future experiences, as well as examples of specific references to how mental time travel is conceptualized in the prior literature. For example (p. 6):

> **"The *mental time travel* metaphor is defined by Tulving as "the type of awareness experienced when one thinks back to a *specific moment* in one's personal past and consciously recollects some prior episode or state as it was previously experienced..."**

(2) The paper contrasts two sets of annotations, one made scene-by-scene as the movie narrative unfolds, and another made during recall of the movie narrative. The scene-by-scene annotations show markedly less long-range structure than the recall annotations. I was wondering if part of this difference could arise from a difference in demand characteristics between the two annotation tasks. Like if in the scene-by-scene task participants are more focused on lower level characteristics of the movie. Perhaps some more detail could be provided regarding either the instructions given in the scene-by-scene annotation task, or a characterization of the qualitative differences in the kinds of descriptions given?

There are some important points here. First, the annotations themselves reflect the following information (now described on p. 11): a brief narrative description of what was happening, the location where the scene took place, whether that location was indoors or outdoors, the names of all on-screen characters, the names of the character(s) who were in focus in the camera shot, the name(s) of any character(s) who spoke during the scene, the camera angle of the shot, a transcription of any on-screen text, and whether or not there was any music present in the auditory background. In other words, the annotations comprise a mix of low-level and high-level properties. Participant's recalls also comprise a mix of low-level and high-level properties. The annotations and recalls may be downloaded here: https://github.com/ContextLab/mental-time-travel-paper.

The primary qualitative difference between the annotations versus participants' recalls is that the annotations comprise descriptions of information that is immediately available to the annotators, whereas recalls are generated internally by the participants (i.e., from memory). For this reason, the annotations tend to focus on short-timescale information, whereas recalls reflect information at a range of timescales. Critically, the recalls also tie in information *across* events in ways that the annotations do not. For example (p. 14):

> **"...early in the episode, Sherlock and John appear to be investigating a series of suicides. Later in the episode, it becomes clear that the apparent suicides are actually murders. The annotations refer to early deaths (including that shown in Fig. 3B) as suicides, whereas participants often refer to those early deaths as "murders" even though they could not have known that they were murders early on in the episode. In this way, participants' experiences *subsequent to* a target remembered event may distort its memory by incorporating information learned during other experiences."**

More generally, the key phenomenon of interest here is in these sorts of "distortions" between annotator's descriptions of what is directly happening in each scene versus

how participants recall those same events later (p. 11--14). The example in Figure 3 shows that even though the target scene's annotations themselves (gray curve) are temporally localized, the way people *remember* that scene later, after viewing the rest of the episode, tend to incorporate relevant information about that scene that the participant gleaned from other relevant parts of the episode beyond that one scene (black curve).

(3) A substantial portion of the paper deals with using topic models to understand the semantic structure of narrative annotations. I would love to see more detail on how these models work, I think this would help readers that maybe are familiar with classic vector space models of semantics like LSA but not necessarily topics models. This would also lay the groundwork for a bit more detail on the idea of a 'topic trajectory' developed later in the paper. Perhaps there's a way to more clearly establish that the shape of these trajectories by no means had to end up looking similar across people, to set a bar for how impressed we should be? (I think it is quite impressive but it doesn't come across clearly in the current write up.)

I have added two new sections to address these insightful comments. First, to address the point that the shapes of these trajectories "by no means had to end up looking similar across people," I have added a new section entitled "The *matching problem*" (p. 18--19). The section provides additional context and an explanation of why the matching trajectory shapes are "impressive."

Second, to address the reviewer's request for additional background information on word embedding models, I have added a new section entitled "Defining the geometries of thoughts and memories" (p. 19--25, Fig. 5). The section provides additional history of word embedding spaces and explains several key models and advances (including LSA and topic models, as well as more recent deep learning-based approaches).

Minor comments.

(1) The quantum wave function (QWF) analogy is described in detail, but the motivation for specifically calling this QWF is not well developed. That is, the paper doesn't describe QWFs explicitly at any point. There is some potential for confusion with the recent popularity of 'quantum probability' decision models. Those models explicitly use the version of probability theory developed to explain quantum mechanics in physics. Here, I believe the point is simply that the representation that is retrieved during recall is best thought of as a superposition of representations corresponding to multiple non-contiguous timepoints.

To address this comment (along with the reviewers Major Comment 1, above), I have added a new section entitled *Defining the* quantum memory wave function *metaphor and comparing it to the* mental time travel *metaphor* (p. 6--8, Fig. 1). The QWF metaphor derives directly from quantum wave functions (Schrödinger, 1926) for describing the spatial positions of quantum particles. The wave function describes, for each spatial

location *x*, the probability that the particle is located at that position. In the original formulation of this theoretical construct, Schrödinger conceptualized particles as occupying *every* position in space simultaneously, where more "probability mass" is placed at the peaks of the particle's wave function.

The analogy to memory is that when we remember our past, we spread our thoughts over *every* moment from our past. The quantum memory wave function describes how much our thoughts (at time *t* when we are engaging in remembering of our past) incorporate information about each moment along our autobiographical timeline. Importantly, as the reviewer notes, the quantum memory wave function may have multiple peaks (e.g., at non-contiguous timepoints or time ranges).

I appreciate the suggestion to describe how this work relates to quantum probability decision models. The framework I propose is analogous to how decision models that incorporate quantum probability formalisms describe information states (e.g., Khrennikov et al., 2018); the present manuscript extends these ideas into the domain of episodic memory. I have added a note to this effect on page 6.

(2) On page 10 it is proposed that the QWF logic may provide a more accurate reflection of what neural patterns represent. Many of the neural examples early in the paper focus on neural representations at multiple timescales but I don't think these explicitly address this idea of expecting a neural representation that reflects discontinuous experiences. Perhaps the literature on transitive inference would be useful to mention in terms of neural effects where temporally discontinuous points are blended into one representation. It could also be of value to include some neural predictions to guide future work.

I have clarified that statement to read as follows (p. 26):

> **"Traditional approaches then "clean up" this distribution over possible mental states by taking its average (expected), maximum, or modal value. However, following the "quantum wave function" logic described in this manuscript, these multi-peaked response patterns may provide a mechanism for associating non-contiguous representations."**

In other words, the point I was trying to make was about the underlying mental representations that neural activity patterns reflect. Specifically, the representations themselves may reflect weighted combinations of multiple (potentially non-contiguous) mental states or constructs. If so, this should be reflected in our approaches to decoding neural activity patterns as mental states.

I have also added some references to work reporting non-contiguous spatial and temporal receptive fields (primarily in the rodent hippocampus and entorhinal cortex); p. 26-27:

> **"Further, a number of studies have suggested that some neurons respond to distinct (non-contiguous) places (Fenton et al., 2008; Rich et al., 2014; Lee et al., 2019; Derdikman et al., 2009; Grieves et al., 2020) and times (Pastalkova et al., 2008). Taken together, these studies suggest that variability in the set of locations and/or times to which a neuron responds is not merely a reflection of noise or imprecision. Rather, this variability might have some functional relevance in terms of facilitating the learning of paths around obstacles and towards goals."**

(3) On page 7 a distinction is made between semantic and conceptual "semantically (but not conceptually)". These terms mean so many things to so many people that it is worth explicitly defining each, explicating how the author thinks they are different.

The sentence the reviewer is referencing that appeared to distinguish semantic versus conceptual overlap was poorly worded. I did not intend to draw the distinction the reviewer is suggesting. Rather, I was trying to distinguish events that happened to share semantic attributes (e.g., two events that took place in a park, or two events involving an overlapping set of characters) versus events that were directly linked in the episode's narrative. I have reworded that paragraph (now on p. 14) as follows:

> **"The way they described the example target event did *not* mirror the way they described other events that were not directly narratively linked with the target event, even when those other events overlapped semantically. For example, thematic elements of events involving unrelated violence, or Sherlock and John interacting with other characters about other subjects, tended not to feature in participants' recountings of the target event. Taken together, this suggests that these specific events were associated in participants' memories, perhaps contributing to (or reflecting) their understanding that those events were linked in the episode's narrative."**

(4) The author makes the point that the shape of a topics trajectory might have importance. I think this point could be elaborated / clarified, with regards to what it means for the path to go in a certain direction. Presumably a common topics model is estimated for all the participants, which puts their trajectories into a common representational space. Is it really shape that is important? An alternative that could be contrasted with the shape idea would be something like: If you had two people visit the same sequence of landmarks in a city, their path shapes would look quite similar, but this would be because they are visiting the same series of places. Can an analysis establish that it is really shape of trajectory that's important and not a common sequence of places being visited in this abstract space?

The reviewer raises some interesting points. Is it possible to describe an experience *without* reproducing that experience's trajectory? For example, is navigating in semantic space analogous to navigating in a spatial environment, where visiting the same sequence of landmarks necessarily results in the same path through physical space?

In brief, I see the specific shape of an experience's (or memory's) topic trajectory as important and informative, beyond simply visiting a sequence of fixed landmarks in a physical environment. To examine this question in detail, I found it helpful to consider some key differences between mental and physical navigation. I have added a paragraph (p. 25--26) comparing "mental" navigation (i.e., topic trajectories describing how we think about experiences and remember them) and physical navigation (e.g., through real-world spaces):

> **The way we navigate through *thought space* (i.e., word embedding spaces; Figs. 4, 5) differs from how we navigate in physical spatial environments. First, we cannot teleport in real space, whereas our thoughts can exhibit rapid jumps between different non-contiguous regions of word embedding space. Second, when two people visit the same fixed physical landmark, they necessarily visit the same physical location. When two people describe a shared experience, their idiosyncratic thoughts, goals, prior experiences, etc. can lead them to process and remember the experience differently. Third, when we revisit a fixed physical landmark, we (by definition) revisit the same spatial location. In other words, it is possible to visit the same spatial location twice. However, we cannot revisit our prior experiences in this way. We only get to experience each moment once; subsequent "visits" to a prior experience must occur through a different medium than the original experience (e.g., our memory, another person's memory, a physical recording, a written account, etc.). Fourth, we can revisit different aspects of the same prior experience on different "visits"-- e.g., when we remember a specific event we can focus on the physical occurrence in one remembering, the emotional occurence in another remembering, the social implications in another remembering, etc. We can also revisit the same experience in different levels of detail or depth. There are no analogs of these phenomena in spatial navigation. Finally, *physically* (at a macro scale), we are always located at a single moment and in a single location. The main argument of the quantum memory wave function framework I propose here is that we may be *mentally* spread over many times or locations simultaneously. For example, this provides a way for us to integrate information from multiple prior experiences into a single decision or conceptualization, even if those experiences were not temporally contiguous.**

Many of these differences also hold for the distinction between physically navigating through space versus mentally navigating through space (e.g., tracing out a potential route in one's mind; I have added a note to this effect on p. 26).

A second distinction worth highlighting is between memory for high-level versus low-level details of an experience. Reproducing (by recounting) the low spatial frequency properties of the trajectory shape of an experience requires capturing its high-level (essential) details. In other words, the trajectory shape's gross scaffolding describes how its conceptual content changes over long timescales and across different discrete events. By contrast, recounting low-level details is reflected in the high spatial frequency properties of the trajectory shape (i.e., its within-event details). Concretely, high-level details include major narrative plot points that are necessary to understand the overall "story." Low-level details include non-essential elements of the story (e.g., the

color of a character's shirt, an interaction between minor characters, etc.). The main point that my group makes in one of our prior studies (whose findings are featured prominently in this paper) is that it is usually more important to remember the high-level details than the low-level details of an experience, even if the amount of information between these two sets of features is roughly comparable. In other words, it is better to match (in your memory of an experienced event) the word-embedding coordinates related to the event's high-level details than its low-level details. Therefore reproducing the trajectories overall "shape" is more important than correctly reproducing the precise coordinates of its control points (as long as the low spatial frequency properties of the shape are roughly preserved). In that prior work we also make the point that there are no analogs of this high-level versus low-level distinction in most classic memory tasks (e.g., word list learning tasks), where the primary objective is to exactly reproduce a given word or sequence from memory. I have added a paragraph with some more detail from that prior study (p. 17--18):

> **"In their paper, Heusser et al. (2018a) use the geometric shapes of experience trajectories (through word embedding space) to functionally distinguish high-level essential details of the experience from low-level (non-essential) details. When the trajectory exhibits a sharp change in direction, this indicates a rapid shift in the thematic content of the episode. Heusser et al. (2018a) used hidden Markov models to detect these rapid shifts and segment the trajectories into discrete *events*. The high-level (essential) details of the episode are reflected in the shape defined by the sequence of coordinates of its events. This may be characterized by examining the low spatial frequency properties of the trajectory's shape. Low-level details are characterized by (typically smaller-scale) within-event geometric patterns, which are captured by the high spatial frequency properties of the trajectory's shape. A key finding of that work was that people who watched the episode recalled high-level details reliably, and in a similar way across people, whereas people varied substantially in how they remembered low-level details. This suggests that our memory systems may prioritize certain types of information (e.g., high-level details) about our experiences over others (e.g., low-level details) as we encode our experiences into memories or when we recount our prior experiences."**

Reviewer #2: Comments on "Episodic memory: mental time travel or quantum "memory wave" function?" by Manning

In this interesting opinion paper, the author warns on the difficulty and hand waviness of the contextual approach to episodic memory. The author reviews a selected set of findings and suggests a modified conceptual view in which memory retrieval would rely on "mentally casting ourselves back simultaneously to many time points from our past".

Coming from time research, I find this piece interesting and perhaps unorthodox for the memory field of research. I would perhaps stress that the title, while certainly catchy, is very misleading and perhaps an unnecessary metaphorical stretch of imagination in light of the sober content of the article.

I appreciate the reviewer's feedback. I do feel that the "quantum wave function" metaphor is appropriate, although I take the reviewer's point that this metaphor was not sufficiently explained or expanded on in my prior draft. I therefore decided to keep the title unchanged, but I have added a new section entitled *Defining the* quantum memory wave function *metaphor and comparing it to the* mental time travel *metaphor* (p. 6--8, Fig. 1). This section provides additional clarifying detail about the metaphor and links it directly to Schrödinger (1926)'s quantum wave function formulation.

I will highlight three specific points discuss by the author, which could be refined:

(1) The author suggests that multiple "temporal drifting" coexist yielding temporal hierarchies. However, I would contend that the empirical findings do not show "temporal drifting" per se, but rather "temporal tuning" or "time windows", and these do not imply a decay rate as much as temporal integration. For instance, and complementary to the discussed work by Honey's team:

The notion of multiplexing can seminally be found in auditory and speech research in which the coexistence of different temporal resolutions for information coding seems to be the rule rather than the exception:
** Viemeister, N. F., & Wakefield, G. H. (1991). Temporal integration and multiple looks. The Journal of the Acoustical Society of America, 90(2), 858-865.*
It can nevertheless be found in vision:
** Gauthier, B., Eger, E., Hesselmann, G., Giraud, A. L., & Kleinschmidt, A. (2012). Temporal tuning properties along the human ventral visual stream. Journal of Neuroscience, 32(41), 14433-14441.);*

More generally, and as a side comment, the author may be interested in the notion of "temporal window of integration".

I appreciate the pointers to these studies, and I have integrated these references into the Introduction and overview (p. 3):

> **"This notion, that our brain maintains parallel mental representations drifting at different time scales, has been well-characterized in a series of elegant fMRI and ECoG studies by Uri Hasson and colleagues (Hasson et al., 2008; Lerner et al., 2011; Honey et al., 2012; Aly et al., 2018). Subsequent work by the same group has shown that this hierarchy of drifting mental representations plays a central role in how we remember continuous experiences and segment our experiences into discrete events (Baldassano et al., 2017). This body of work suggests that different brain regions reflect *temporal receptive windows* that characterize the timescale(s) to which that region is maximally sensitive. This research complements work on the *temporal window of integration* in the speech and perception literatures (e.g., Viemeister and Wakefield, 1991; Gauthier et al., 2012; van Wassenhove et al., 2007). The temporal window of integration refers to the duration over which physically distinct stimuli are treated (by some brain system or network) as a single percept or an atomic unit. This systems-level work mirrors the discovery of *time cells* in the rodent**

hippocampus that respond preferentially to a specific time relative to a temporal reference (e.g., time elapsed since starting to run on an exercise wheel), whereby each cell appears to integrate incoming information at a different rate (Pastalkova et al., 2008; MacDonald et al., 2011) and the population activity serves to represent information at a range of timescales (Mau et al., 2018).”

I also agree that the empirical findings discussed do not specifically identify drift as a mechanism, but rather they point to sensitivity to different timescales as the reviewer notes. I have modified the text the reviewer referenced (e.g., see example on p. 4) as follows (emphasis added):

> **“Collectively, these experimental and theoretical studies make two important contributions to our understanding of how we remember our past. First, the studies suggest that our ongoing experiences are “blurred out” in time by neural integrators (where different brain structures or sub-structures integrate with different time constants). Second, the studies suggest that representations that *reflect different timescales* become bound together such that when we retrieve memories about our past, we reactivate representations that carry information at a range of time scales. These reactivations explain how our brains reactivate prior contexts when we remember the past.**

(2) While the author highlights the novelty of the view that simultaneous and parallel "nows" coexist in the brain, I would take this as a re-interpretation of "temporal multiplexing", which has been reported in several lines of research including memory, speech or even timing research. For instance, explicit discussions of "temporal multiplexing" can be found here:
** Caruso, V. C., Mohl, J. T., Glynn, C., Lee, J., Willett, S. M., Zaman, A., ... & Groh, J. M. (2018). Single neurons may encode simultaneous stimuli by switching between activity patterns. Nature communications, 9(1), 2715.*
** Knight, R. T., & Eichenbaum, H. (2013). Multiplexed memories: a view from human cortex. Nature neuroscience, 16(3), 257.)*
** Panzeri, S., Brunel, N., Logothetis, N. K., & Kayser, C. (2010). Sensory neural codes using multiplexed temporal scales. Trends in neurosciences, 33(3), 111-120.*
Hence, what is perhaps unclear in the argument put forward by the author is at which representational level the coexistence of "nows" should be considered: is it but a metaphor? Are multiple "nows" to be found at a computational and pre-attentive level (as would currently suggest the literature)? Or is it to be taken literally at a conscious level?

I appreciate the pointer to these temporal multiplexing papers and have added references to them in the introduction (p. 3--4):

> **“In addition to this work suggesting that individual system or neurons maintain representations at their preferred timescales, a number of studies also report neurons and small populations exhibiting *temporal multiplexing*, whereby information at different timescales is reflected in different features of an individual neuron’s (or population’s) activity patterns, such as spike timing versus firing rate or local field potential oscillations at different frequencies (Caruso et al., 2018; Knight and Eichenbaum, 2013; Watrous et al.,**

2013; Panzeri et al., 2010)."

The question of our subjective experience of "multiple nows" is an interesting one, and to my knowledge remains unanswered. I have added the following paragraph to address the reviewer's question (p. 8):

> **"While the studies presented in the *Introduction and overview* on temporal receptive windows, the temporal window of integration, time cells, and temporal multiplexing present strong empirical evidence that our brains maintain parallel representations at multiple timescales, questions remain about how or whether we experience those representations at a conscious level. For example, are we consciously aware of spreading our mental state over many moments from our past? Or is our subjective experience of remembering more like mentally time traveling back to a single "hybrid" moment that blends information from several of our (actual) prior experiences?**

(3) I am very sympathetic to the discussion on how the contiguity in episodic memory may not be necessary, along with the interesting and canonical issue of the limitation of serial order in memory - notably in Human resarch. Recent work dealing with "mental time travel" empirically speaks to the limitation of serial order in memory. For instance:

** Arzy, S., Adi-Japha, E., & Blanke, O. (2009). The mental time line: An analogue of the mental number line in the mapping of life events. Consciousness and cognition, 18(3), 781-785.*
** Gauthier, B., & van Wassenhove, V. (2016). Cognitive mapping in mental time travel and mental space navigation. Cognition, 154, 55-68.*
** Gauthier, B., Pestke, K., & van Wassenhove, V. (2018). Building the Arrow of Time… Over Time: A Sequence of Brain Activity Mapping Imagined Events in Time and Space. Cerebral Cortex.*
** Arzy, S., & Schacter, D. L. (2019). Self-Agency and Self-Ownership in Cognitive Mapping. Trends in cognitive sciences.*

I appreciate the pointers to these papers; I have added references to them in my revised manuscript.