Self- versus other-generated interpretations of ambiguous social stimuli are asymmetrically remembered

Clara Sava-Segal (csava.gr@dartmouth.edu)
Department of Psychological & Brain Sciences, Dartmouth College, 3 Maynard St., Hanover, NH 03755 USA

Emily S. Finn (emily.s.finn@dartmouth.edu)
Department of Psychological & Brain Sciences, Dartmouth College, 3 Maynard St., Hanover, NH 03755 USA

Abstract

Ambiguous social situations have a variety of possible interpretations, making them a rich testbed for studying biases in memory processing: people must integrate differing viewpoints of the same sensory information (e.g., their own versus someone else's), ultimately affecting what they remember. Here we probed how individuals encode and remember their own subjective interpretations versus those sourced from others of the same ambiguous sensory input. We find that although the fidelity of both interpretations in memory is relatively high, people, even when confident in their reports, show a marked tendency to merge their memories of other's interpretations to be more like their own original interpretation. This asymmetry suggests a cognitive preference for aligning external interpretations with one's own, extending the understanding of self-referential effects in memory and showing a memory bias towards self-generated narratives in ambiguous situations.

Keywords: memory; social contexts; naturalistic stimuli; natural language processing

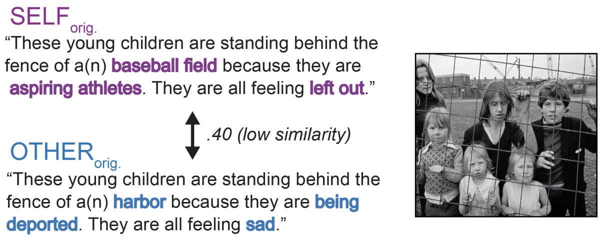
Introduction

Ambiguity arises when the same information can be interpreted in different ways. Consequently, the same events are often remembered in distinct ways within or across people, making ambiguous stimuli a valuable testbed for studying subjective memory recall.

While we typically form initial interpretations of ambiguous information on our own, social contexts (e.g., considering a friend's opinion) can prompt us to reassess our subjective perceptions of an experience. Although previous studies have explored both individual differences in interpretation (Finn et al., 2018; Nguyen et al., 2019) and subjective recalls of conceptually ambiguous stimuli (Lee and Chen, 2022; Sava-Segal et al., 2023) in isolation, our study aims to integrate these questions within a simulated social context. We developed a novel "naturalistic" encoding and recall paradigm that employs complex, conceptually ambiguous stimuli, requiring participants to both generate their own interpretations and engage with those of others. This approach enabled us to not only investigate how we recall multiple subjective interpretations that are attributed to the same experience, but also explicitly compare recalls for interpretations from different sources (i.e., self versus someone else).

Materials and methods

Task paradigm Participants (N=75) did a two-session experiment. During session 1 (encoding), participants were presented with black-and-white photographs (n=45 images/trials) depicting a person or multiple people in ambiguous scenarios, where the ambiguity generally centered around where the people are, what they were doing, what/how they were thinking/feeling, and what their relationships are to one another, extensively piloted (N>500) to ensure that they evoke a wide range of interpretations. On each trial, participants completed a "MadLibs"-style interpretation task where they filled in three blanks to describe what they believed was happening in the image ($\text{SELF}_{\text{original}}$). See Fig. 1 for an example image[1] and interpretations. They were then presented with another participant's interpretation ($\text{OTHER}_{\text{orig.}}$) that varied across trials in semantic similarity (i.e., ranged from being very similar to being very dissimilar) from their own interpretation. They also provided metacognitive reports of their agreement with both interpretations. In session 2, participants recalled both their own ($\text{SELF}_{\text{mem.}}$) and the alternative interpretations they were presented with ($\text{OTHER}_{\text{mem.}}$) in a counter-balanced order. They also reported their subjective confidence in their recall for each interpretation per image.



$\text{SELF}_{\text{orig.}}$
"These young children are standing behind the fence of a(n) **baseball field** because they are **aspiring athletes**. They are all feeling **left out**."

$\text{OTHER}_{\text{orig.}}$
"These young children are standing behind the fence of a(n) **harbor** because they are **being deported**. They are all feeling **sad**."

Figure 1: Sample image presented with two sample interpretations; participants generated their own and were presented with one from another participant.

Recognition Participants completed an old/new task. Performance was at ceiling; analyses focus on recall for interpretations rather than images themselves.

Computing semantic similarity and recall fidelity Interpretations were operationalized as high-dimensional vectors in semantic space defined by miniLM, a compact natural language processing model with transformer-based architecture (Wang et al., 2020). Crucially, the "MadLibs"-style three-blank-structure made the blanks more comparable since they had a shared context and order; embeddings were formulated such that the shared context was used as a baseline and then the deviations within

[1] Photographer: Philip Jones Griffith

each blank were directly compared. To quantify the semantic similarity between two interpretations per trial/participant pair, the cosine similarity between the two interpretations was computed. To quantify memory fidelity, the cosine similarity between the original and remembered interpretation vectors was computed.

Memory merge To quantify the degree to which interpretations remained "distinct" versus "merged"—that is, became more similar to the other in memory— we projected the remembered interpretations ($\text{SELF}_{\text{mem.}}$ and $\text{OTHER}_{\text{mem.}}$) onto an axis defined by the original interpretations in the embedding space. We then calculated the normalized cosine distances from the original points to these projections, yielding a measure of how far each memory deviated from its origin. This measure reflects the percentage of the original distance between the two interpretations that each remembered interpretation *traveled* in memory. We then compared these distances to assess the degree of "merge" for each interpretation (SELF vs. OTHER).

Statistics All analyses used linear mixed effects models with images and participants as random effect. To compute a 'null memory score' to confirm that interpretations were remembered above what would be expected if participants were simply filling in the blanks anew, each remembered interpretation was compared to the bank of all other unseen original interpretations for that image; the median value across participants and images is plotted in Fig. 2A.

Results and Discussion

While recall fidelity was generally high for both interpretations, self-generated interpretations ($\text{SELF}_{\text{mem.}}$) were remembered better than alternative interpretations ($\text{OTHER}_{\text{mem.}}$; Fig. 2A; β=.08; $p<.001$). Subjective confidence paralleled this objective pattern of memory accuracy, with individuals exhibiting greater confidence in recalling $\text{SELF}_{\text{mem.}}$ than $\text{OTHER}_{\text{mem.}}$ ($\chi^2 = 544.14$, df = 3, $p<.001$).

We also found that interpretations *merge* in memory (β=.12; $p<.001$), i.e., the two interpretations are remembered as *less* distinct than they were when originally encoded (Fig. 2B). These effects persisted across trials, regardless of self-reported recall confidence or interpretation preference.

To understand this merge, we next sought to determine if it stemmed from a bias towards one's own interpretation, driving alternative interpretations to align more closely with self-generated ones. We observed notable asymmetry in how different interpretations merged in memory, with $\text{OTHER}_{\text{mem.}}$ consistently *traveling* more in semantic space towards $\text{SELF}_{\text{orig.}}$ than the reverse ($\text{SELF}_{\text{mem.}}$ to $\text{OTHER}_{\text{orig.}}$; Fig. 2C; β=.15; $p<.001$). This asymmetry was robust across various levels of subjective confidence (i.e., even when participants reported being confident in $\text{OTHER}_{\text{mem}}$) and persisted regardless of the order in which recalls were probed. While still present, the asymmetry of the merge was reduced when participants were presented with an alternative interpretation that was more semantically distinct from their own, and when they reported higher agreement with the other interpretation in metacognitive reports from session 1 (β=.38; p=.025).

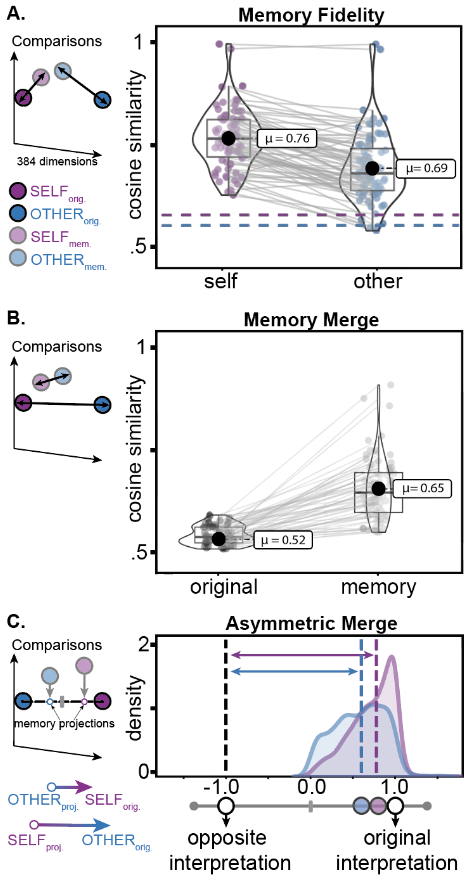


Figure 2: (A) Both interpretations are remembered above baseline, but $\text{SELF}_{\text{mem.}}$ is remembered better. Colored-coded dashed lines indicate the median of the null distribution for each interpretation. (B) Interpretations become more similar in memory. (C) Even though the fidelity is high for both interpretations (both projections remain close to the original interpretation), $\text{OTHER}_{\text{mem.}}$ moves closer to $\text{SELF}_{\text{orig.}}$. than vice versa.

Conclusions

These findings extend the well-established self-reference effect (Rogers et al., 1977), where self-relevant information is better recalled. Self-generated interpretations serve as an "anchor" for how ambiguous situations are remembered. We show that when faced with multiple, equally plausible interpretations, individuals tend to prioritize their own interpretations in memory. This demonstrates a memory bias towards self-generated narratives in ambiguous situations.

Acknowledgements

This work was supported by a National Science Foundation Graduate Research Fellowship (to C.S.S.) and by the National Institutes of Health (grants K99MH120257 and R00MH120257 to E.S.F.).

References

Finn, E. S., Corlett, P. R., Chen, G., Bandettini, P. A., and Constable, R. T. (2018). Trait paranoia shapes inter-subject synchrony in brain activity during an ambiguous social narrative. *Nature Communications*, 9(1):2043. Number: 1 Publisher: Nature Publishing Group.

Lee, H. and Chen, J. (2022). Predicting memory from the network structure of naturalistic events. *Nature Communications*, 13(1):1–14. Number: 1 Publisher: Nature Publishing Group.

Nguyen, M., Vanderwal, T., and Hasson, U. (2019). Shared understanding of narratives is correlated with shared neural responses. *NeuroImage*, 184:161–170.

Rogers, T. B., Kuiper, N. A., and Kirker, W. S. (1977). Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology*, 35(9):677–688. Place: US Publisher: American Psychological Association.

Sava-Segal, C., Richards, C., Leung, M., and Finn, E. S. (2023). Individual differences in neural event segmentation of continuous experiences. *Cerebral Cortex*, page bhad106.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. (2020). MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. arXiv:2002.10957 [cs].

-