Cognitive and Neural Mechanisms of False Memory Formation: An Information Processing Perspective

Authors: Guo Ying, Gong Xianmin, Wang Dahua, Wang Dahua

Date: 2020-08-16T00:00:00+00:00

Abstract

From an information processing perspective, and based on the categorization of different information sources, we analyze how the series of processing stages –including encoding, storage (consolidation), reactivation/reconsolidation, and retrieval–lead to the formation of false memories, thereby summarizing three possible causes of false memory: (1) Due to the lack of memory representations for item-specific details of target items, there is an emphasis on encoding and retrieving abstract memory representations shared by target and non-target items, causing subjects to rely more heavily on abstract representations to reconstruct missing target details, thereby triggering false memories; (2) Target items activate corresponding schemas, leading to enhanced memory representations of schema-related non-target items, which triggers false memories; (3) Misleading information interferes with the memory representations of target items in the reactivated state, preventing accurate memory reconsolidation and thus triggering false memories. Future research could further investigate issues such as the representational regions of item-specific details of target items, the specific mechanisms by which different types of schema representations enhance memory representations of non-target items, and the influence of schema reinstatement during the retrieval phase on false memory formation.

Full Text

Preamble

The Cognitive and Neural Mechanisms of False Memory Formation: An Information Processing Perspective

Guo Ying[1], Gong Xianmin[2], Wang Dahua[1] (Corresponding Author)
([1] Institute of Developmental Psychology, Beijing Normal University, Beijing,

100875, China)
([2] Department of Psychology, University of Zurich, Zurich, 8050, Switzerland)

Abstract: From an information processing perspective, this paper analyzes how a series of processing stages—encoding, storage (consolidation), reactivation/reconsolidation, and retrieval—contribute to false memory formation when different information sources are distinguished. We identify three potential causes of false memory: (1) When memory representations lack item-specific details of target events, individuals tend to encode and retrieve abstract representations shared by targets and non-targets, leading them to reconstruct missing target details based on these abstract representations and thereby producing false memories; (2) Target events activate corresponding schemas, which enhances memory representations of schema-related non-target events and triggers false memories; and (3) Misleading information interferes with reactivated memory representations of target events during reconsolidation, preventing accurate memory reconsolidation and thus causing false memories. Future research should further investigate the brain regions representing item-specific details of target events, the specific mechanisms by which different types of schema representations promote memory representations of non-target events, and how schema reactivation during retrieval influences false memory formation.

Keywords: false memory; cognitive schema; neural mechanisms; information processing

1 Introduction

When individuals recall experienced events, the memories retrieved by the brain are often not faithful reproductions of reality but rather reconstructions of original events that incorporate prior experiences and external interference. This can lead people to erroneously recall events that never occurred or recall details that do not match reality, resulting in false memory. How exactly does false memory arise? Some scholars have already analyzed and reviewed the cognitive and neural mechanisms underlying false memory from different perspectives. For example, Jiang and Li (2015) summarized the causes of associative false memory from the perspectives of fuzzy-trace theory and associative activation theory. Wang and Geng (2010) further analyzed how gist traces and verbatim traces in fuzzy-trace theory influence associative false memory. Liu et al. (2015) reviewed possible causes of implanted false memory. Although these reviews undoubtedly benefit our understanding of false memory mechanisms, they have limitations: First, they mostly focus on one specific type of false memory (e.g., associative false memory, implanted false memory) and lack an overarching framework that encompasses multiple types of false memory (e.g., Liu et al., 2015). Second, they tend to emphasize the static influence of certain factors (e.g., gist traces, verbatim traces) on false memory, analyzing the relationship between the presence or absence of a factor and false memory formation, while lacking a summary of

its dynamic mechanisms (e.g., Wang & Geng, 2010).

To address these two issues, this paper adopts an information processing perspective. On the one hand, we distinguish different information sources and discuss how processing of target events, internal schemas, and external interference leads to false memory formation. We place false memories caused by internal processing (e.g., associative false memory) and those caused by external processing (e.g., implanted false memory) within the same information processing framework. This allows us to move beyond previous reviews that were limited to specific types of false memory. Moreover, all types of false memory involve information processing of original target events. Integrating information processing related to correct memory formation with that related to internal schemas and external interference helps us develop a more comprehensive and systematic mechanism of false memory formation. On the other hand, we distinguish different information processing stages and discuss possible causes of false memory during encoding, storage, reactivation/reconsolidation, and retrieval. Memory is a dynamic process, and errors in any stage—encoding, storage, or retrieval—may trigger false memories. Therefore, when discussing factors influencing false memory, we must consider not only "presence or absence" but also "which processing stage," thereby dynamically analyzing their role in false memory formation.

Thus, this paper combines the distinction of information sources (see gray boxes in Figure 1 [Figure 1: see original paper]) with the distinction of information processing stages (see white boxes in Figure 1) to summarize three possible mechanisms of false memory formation within an information processing framework (see dashed boxes in Figure 1). Each mechanism focuses on the processing stages of information from a particular source. Specifically: Mechanism One corresponds to processing of target events, where lacking item-specific detail representations during encoding, storage, and retrieval leads to false memory. Mechanism Two corresponds to processing of internal schemas, where rapid schema activation by target events changes memory representations of other related events during encoding and increases false memory through schema reactivation during retrieval. Mechanism Three corresponds to processing of external interference, where encoding misleading information inconsistent with target events during reactivation of target memory representations interferes with original memories and triggers false memory. Below, we elaborate on these three mechanisms in detail.

Figure 1. Schematic diagram of false memory mechanisms from an information processing perspective. Note: Gray boxes represent different information sources (from target events, internal schemas, and external interference); white boxes represent different information processing stages (encoding, storage, reactivation/reconsolidation, and retrieval); dashed boxes represent possible mechanisms of false memory formation, each focusing on the processing stages of information from a particular source, labeled and corresponding to the white boxes below. The three mechanisms are independent yet interconnected[2] and

jointly cause false memory. It should be noted that although Mechanism Three focuses on processing external interference, it also involves processing of target events; therefore, the dashed box representing Mechanism Three appears over both types of gray boxes.

[2] The three mechanisms are both independent and interconnected. Their independence is reflected in: each mechanism corresponds to different information sources and relates to different types of false memory formation. For example, Mechanism Two can explain associative false memory formation, while Mechanism Three can explain implanted false memory formation. Their interconnection is reflected in: Mechanism One, related to correct memory formation, typically co-occurs with Mechanisms Two and Three, which are related to false memory formation. Both associative and implanted false memories lack distinctive item-specific representations of target events. Additionally, there may be mutual influences between Mechanism One and Mechanism Two, and between Mechanism One and Mechanism Three. For instance, schema' s inhibitory effect on distinctive details in Mechanism Two (e.g., van der Linden et al., 2017) and competition between misleading information and distinctive details in Mechanism Three (e.g., Okado & Stark, 2005).

---

2 Mechanism One: Lack of Distinctive Detail Representations for Target Events

Different memory contents form different memory representations in abstract psychological space, corresponding to different neural representations at the neural level. The distinctive representation discussed in this section refers to unique memory/neural representations for specific memory content that can distinguish it from other memory contents. Neural representations vary in their degree of distinctiveness. Less distinctive neural representations can only distinguish memory contents with obvious perceptual/semantic differences (e.g., distinguishing pictures of the "Sydney Opera House" and "Great Wall of China"), whereas more distinctive neural representations can further distinguish memory contents with subtle detail differences (e.g., distinguishing pictures of similar-looking but non-identical "cats").

False memory research typically examines participants' false alarm/rejection responses to related lures (i.e., probes that are highly similar to target items in perceptual/semantic features but differ in certain details, hereafter referred to as lures). Rejecting lures—i.e., distinguishing them from targets—depends on highly distinctive neural representations of target details. When such distinctive representations exist, the match between target and lure memory representations decreases, and participants' memory strength for lure probes declines (Norman, 2010), leading to greater tendency to reject lures. Conversely, when such distinctive representations are lacking and only abstract memory representations reflecting common features of targets and lures exist, participants tend to rely

on abstract representations to reconstruct missing target details, increasing false alarm rates. Below, we discuss the internal mechanisms and related neural evidence regarding distinctive detail representations for targets (hereafter referred to as distinctive representations) during encoding and retrieval stages in false memory formation.

2.1 Encoding Stage: Lack of Distinctive Representations

Lacking distinctive representations during encoding may lead to false memory formation. Research on the hippocampus provides direct evidence for this mechanism. The constructive memory framework emphasizes the important role of the hippocampus' s pattern separation function in forming accurate memories (Schacter et al., 1998). Pattern separation refers to the hippocampus' s ability to transform similar memory contents into orthogonal neural representations. Specifically, the hippocampus minimizes overlap between neural representations of similar memory contents, enabling individuals to form distinctive, diagnostic representations for each memory content that facilitate accurate discrimination between different memories (LaRocque et al., 2013; Stevenson et al., 2020; Yassa et al., 2011). Wing et al. (2020) directly confirmed this mechanism. In their study, participants first learned different exemplar pictures under several semantic categories (e.g., "tent"). One day later, participants performed a recognition test on studied target pictures, unstudied lure pictures from the same semantic categories as target exemplars, and unstudied unrelated pictures from categories different from any target exemplars. Using representational similarity analysis (RSA) in multivoxel pattern analysis (MVPA)[3], the study calculated category representations reflecting common perceptual/semantic features among target exemplars and between targets and lures during encoding. They found that when primary visual cortex and superior parietal lobule showed high-level category representations, high-level category representations in the hippocampus led participants to false alarm to lures, whereas low-level category representations in the hippocampus (i.e., hippocampal pattern separation) led participants to reject lures. Thus, hippocampal pattern separation can effectively modulate whether highly overlapping cortical representations impair accurate memory formation. However, hippocampal pattern separation is not unlimited. When memory contents are too similar beyond the separation limit, the hippocampus must engage pattern completion, reactivating similar representations and comparing them with current representations to obtain specific detail information reflecting subtle differences between them, thereby guiding correct memory discrimination (van den Honert et al., 2016).

[3] This method treats neural signals from multiple voxels in the current state as a multidimensional variable (i.e., spatial pattern) and constructs a representational dissimilarity matrix (RDM) for spatial patterns between all item pairs. This matrix can also be derived from spatial patterns in other states, item attribute models, computational models, behavioral models, etc. By comparing relationships between representational dissimilarity matrices, this method en-

ables interpretation of neural representations in the current state (Kriegeskorte et al., 2008).

In addition to the hippocampus, numerous studies have found that some cortical regions, such as primary visual cortex (or secondary visual cortex under specific tasks), can also encode detail information for similar memory contents in a distinctive manner (Baym & Gonsalves, 2010; Garoff-Eaton et al., 2005; Pidgeon & Morcom, 2016; St-Laurent et al., 2014; Xiao et al., 2017). For example, Baym and Gonsalves (2010) found that secondary visual cortex could represent specific detail information distinguishing targets from lures (e.g., if the target event was buying "bananas" in a story context and the lure event was buying "oranges" in the same context, "bananas" would be the specific detail distinguishing the target from the lure). If secondary visual cortex activation decreased during encoding, participants might lack distinctive encoding of target details, thereby increasing false alarms to lures.

However, many studies have not found that distinctive representations in primary/secondary visual cortex during encoding predict lure rejection (e.g., Pidgeon & Morcom, 2016). This may be because distinctive representations in primary visual cortex during encoding often cannot be stably reactivated during retrieval (Kuhl & Chun, 2014; St-Laurent et al., 2014; Staresina et al., 2012; Xiao et al., 2017). Another reason may be that distinctive representations in primary visual cortex during retrieval often do not directly influence memory performance (Lee et al., 2019; Wing et al., 2020). Nevertheless, we cannot ignore the important role of primary visual cortex in forming distinctive memory representations. As a processing region that retains the most original details, it may also influence memory performance by affecting hippocampal pattern separation (Gordon et al., 2014).

2.2 Retrieval Stage: Lack of Distinctive Representations

Lacking distinctive representations during retrieval may also trigger false memory. Three possible reasons may account for this situation: (1) Encoding failure of specific details. Numerous studies have shown that successful memory retrieval actually reflects reactivation of neural representations from encoding (Nyberg et al., 2000; Staresina et al., 2012; Wheeler et al., 2000). Therefore, lacking distinctive representations during encoding may cause reactivated neural representations during retrieval to be similarly too abstract and generalized, and memory reconstruction based on such representations increases false alarms to lures. Wing et al. (2020) used representational similarity analysis to calculate the degree of reactivation during retrieval of category representations from encoding (see Section 2.1) and found that stronger neural reactivation of category representations in the hippocampus led participants to false alarm to lures. This suggests that false memory formation is accompanied by neural reactivation of category-level abstract representations, whereas forming highly diagnostic distinctive representations during encoding is a necessary but insufficient condition for inhibiting false memory. Whether participants can successfully retrieve

item-specific details of targets based on successful encoding is also influenced by storage quality and retrieval mode.

(2) Storage failure of specific details. Research shows that specific detail information decays faster, more extensively, and is more vulnerable to interference than other types of information (Brainerd & Reyna, 1993, 2002; Sekeres et al., 2016). Therefore, specific detail information may be lost during storage and become unretrievable, triggering false memory.

(3) Inappropriate retrieval mode of specific details. Other research suggests that specific detail information is not completely lost but still exists in our memory system, merely inaccessible due to inappropriate retrieval mode (Gonsalves & Paller, 2000; Kensinger & Schacter, 2006; Slotnick & Schacter, 2004). When appropriate retrieval cues are provided (e.g., representing targets), the specific details of target items can still be successfully retrieved and inhibit false memory formation (Guerin et al., 2012a, 2012b; Sekeres et al., 2016; Weinstein et al., 2010; Chen et al., 2015).

Although these three reasons may all lead to lacking distinctive representations during retrieval, they differ in that neural reactivation reflects direct connection between retrieval and encoding stages, whereas storage loss and inappropriate retrieval are only reflected in the retrieval stage itself. Below, we specifically discuss how neural representations in the angular gyrus during retrieval influence false memory formation after excluding direct connection with encoding stage (i.e., hippocampal neural reactivation). Unlike primary visual cortex (see Section 2.1), angular gyrus function is crucial for memory retrieval. On the one hand, Xiao et al. (2017) used representational connectivity analysis in multivoxel pattern analysis and found that the internal representational structure of target items collected by visual processing cortex during encoding was reproduced by angular gyrus during retrieval. On the other hand, angular gyrus is located at the junction between visual processing cortex and prefrontal cortex responsible for cognitive control, making it more likely to influence behavioral memory performance (Lee et al., 2019).

The important role of angular gyrus in memory retrieval urges researchers to focus on its neural representations, as this directly affects memory retrieval quality. Numerous studies have found that neural representations in angular gyrus during retrieval are indeed distinctive (Kuhl & Chun, 2014; Lee et al., 2019; Xiao et al., 2017). However, these studies focused on distinctive representations that distinguish obvious perceptual/semantic differences between items (e.g., scene pictures of "Sydney Opera House" and "Great Wall of China"), rather than the distinctive representations that distinguish detail differences between items, which is the focus of false memory research. The two categories differ in scope.

Current research has not reached a consensus on the specific relationship between distinctive representations in angular gyrus and false memory formation. On the one hand, Ye et al. (2016) used the Deese-Roediger-McDermott (DRM)

paradigm and found in angular gyrus that neural indices representing lure words' memory traces obtained from all target words partially mediated the predictive relationship between semantic similarity among all words (including targets and lures) in the same semantic category and memory strength for lure words. This may suggest that distinctive representations in angular gyrus during memory retrieval mainly reflect obvious semantic differences between target items (Kuhl & Chun, 2014), and this lower-level distinctive representation cannot guide memory discrimination for lure probes in false memory research. This is consistent with findings from Kurkela and Dennis (2016) that angular gyrus shows consistent activation during false memory retrieval in meta-analysis, and with McDermott et al. (2017) that angular gyrus activation during DRM paradigm reflects perceived oldness of memory probes. On the other hand, some research suggests that distinctive representations in angular gyrus reflect detail differences between memory contents (Richter et al., 2016), and participants can use such specific details to inhibit false alarms to lure probes (Guerin et al., 2012a; Lee et al., 2019). Current research has not reached consistent conclusions about how angular gyrus function during retrieval influences false memory. Future research could examine the relationship between distinctive representations in angular gyrus and false memory formation using false memory paradigms.

---

3 Mechanism Two: Cognitive Schemas Enhance Related Memory Representations

Individuals tend to use prior knowledge—i.e., cognitive schemas—to process information in a top-down elaborative manner. This schema-driven elaborative processing helps individuals quickly and effectively understand, absorb, organize, and structure incoming information, thereby improving memory efficiency. It also facilitates extraction of event themes and meanings, making abstraction and generalization possible. However, this elaborative processing may also create byproducts, namely false memory formation (Schacter et al., 2011).

Schemas refer to high-level knowledge networks formed through countless experiences that embody common features among things. They encompass various types of prior knowledge, including semantic networks and script networks that receive considerable attention in false memory research. The former reflects associative relationships between concepts and hierarchical relationships between concepts and categories, while the latter reflects sequences of episodes and causal associations in common events (e.g., buying groceries or ordering takeout) and typical scenes within these episodes (Gilboa & Marlatte, 2017).

At the neurobiological level, schemas also refer to interconnected neocortical representations. Different types of knowledge networks involve different neocortical regions. Key hubs for semantic networks include the inferior frontal gyrus, middle/inferior temporal gyrus, and temporal pole, whereas key hubs for script networks include the medial prefrontal cortex, posterior cingulate cortex,

retrosplenial cortex, precuneus, and inferior parietal lobule (Gilboa & Marlatte, 2017). Below, we discuss the internal mechanisms and related neural evidence of how schemas cause false memory formation.

3.1.1 Encoding Stage: Schema Instantiation

Schema instantiation refers to the process of activating cognitive schemas and maintaining them as generic information processing "templates" (Ghosh et al., 2014). Schema instantiation occurs spontaneously approximately 170 milliseconds after target presentation, preceding conscious awareness (Gilboa & Moscovitch, 2017). The ventromedial prefrontal cortex (vmPFC) is the key region for schema instantiation. It is responsible for identifying corresponding schemas for targets (e.g., the word "appointment" belongs to the "seeing a doctor" schema rather than the "sleeping" schema; Ghosh et al., 2014), recognizing consistency/inconsistency between targets and schemas (e.g., "cactus in a sink" is inconsistent with the cactus schema; Spalding et al., 2015), and strengthening schema-related information processing through enhanced functional connectivity with posterior cortices such as middle/inferior temporal gyrus, inferior parietal lobule, precuneus, and posterior cingulate (Bonasia et al., 2018; Gilboa & Moscovitch, 2017; Sommer, 2017), while weakening information processing unrelated to or inconsistent with schemas (Spalding et al., 2015; Sweegers et al., 2015; van der Linden et al., 2017). When vmPFC is damaged and cannot instantiate schemas, cognitive operations that depend on schema-driven false memory formation (such as associative activation and/or gist extraction[4]) cannot proceed smoothly. Research has found that after learning schema-consistent target pictures (e.g., "pizza in an oven"), vmPFC patients show reduced false alarms to similar lure pictures (Spalding et al., 2015) and reduced false alarms and false recall to lure words in DRM paradigm (Warren et al., 2014). Similar results were obtained when applying repetitive transcranial magnetic stimulation (rTMS) to inhibit vmPFC function in healthy participants (Berkers et al., 2017).

[4] Associative activation refers to the process where individuals spontaneously generate spreading activation among associated items under the guidance of cognitive schemas. Gist extraction refers to the process where individuals actively extract common features among items driven by cognitive schemas. Both are schema-driven cognitive operations that may lead to false alarms to lures (Gallo, 2006).

3.1.2 Encoding Stage: Schema Modulation of Related Memory Representations/Strength

After schemas are instantiated as generic information processing "templates," they top-down guide online information processing in two ways: strengthening schema-related information processing and weakening schema-unrelated information processing. Schema-guided online processing paves the way for subsequent modulation of memory representation/strength formation in terms of

processing time and processing regions.

In terms of processing time, schema-guided online processing occurs at schema instantiation (approximately 170 ms after target presentation), whereas schema modulation of memory representation/strength formation occurs approximately 400 ms after target presentation. At this time, semantic elaboration of schema-consistent target items in relevant knowledge contexts enhances memory strength for target probes (Packard et al., 2017). In terms of processing regions, the transition from schema-guided online processing to schema modulation of memory representation/strength formation involves a shift from vmPFC to posterior cortex (Gilboa & Marlatte, 2017), with vmPFC playing a leading and binding role in neural activity in posterior cortex (Sommer, 2017; Baldassano et al., 2018).

Similar to schema instantiation, vmPFC function during schema-guided online processing is a prerequisite for schema modulation of memory representation/strength. Subsequently, neural activity in posterior cortex truly determines memory representation/strength. Moreover, neural activity in posterior cortex mainly depends on specific content of targets and cognitive operations performed. Specifically, the left inferior frontal gyrus (Addis & McAndrews, 2006; Cooper & Ritchey, 2020; Sommer, 2017) and middle/inferior temporal gyrus (Binder & Desai, 2011), especially the temporal pole (Patterson et al., 2007), are mainly responsible for forming semantic associations among target words using semantic networks, whereas medial prefrontal cortex, posterior cingulate cortex, retrosplenial cortex, precuneus, and inferior parietal lobule are mainly responsible for reconstructing occurrence patterns and typical scenes of target events using script networks (Baldassano et al., 2017; Baldassano et al., 2018; Chen et al., 2017). The two schema networks respectively determine formation of memory representation/strength under each schema. Below, we discuss how semantic networks and script networks in each region modulate memory representation/strength of lure probes and cause false memory formation.

(1) Modulation of lure probe memory representation/strength by semantic networks. Kim and Cabeza (2007) required participants to learn several category word lists (e.g., "farm animals") and perform recognition. They found that both hitting targets and false alarming to lures activated the left inferior frontal gyrus responsible for semantic processing to the same degree during encoding (Cabeza et al., 2001; Garoff-Eaton et al., 2007; Kubota et al., 2006). This suggests that semantic networks activated by category word lists[5] enhance memory strength not only for target probes but also for lure probes under the same semantic network.

[5] Whether activation of key regions such as inferior frontal gyrus, middle/inferior temporal gyrus, medial prefrontal cortex, retrosplenial cortex, and inferior parietal lobule can represent schema function: Whether from psychological or neural representation perspectives, an essential element of being a schema is having an associative network organization (Ghosh & Gilboa, 2014). Therefore, even key

regions cannot simply represent schemas themselves, but this does not mean they cannot represent schema function, for two reasons: (1) Activation of key regions is driven by schemas. For example, research found that the greater the load of forming associative processing (e.g., forming semantic associations among toy, lily, and wool for associative memory), the stronger the activation of left inferior frontal gyrus (Addis & McAndrews, 2006). Although only left inferior frontal gyrus activation is observed in neural results, its activation actually reflects the driving effect of semantic networks. (2) Activation of key regions is relative, reflecting relative schema strength. For example, research found that associating two items within typical scenes (e.g., "bulldozer" and "warning post") activated retrosplenial cortex and related regions more than associating two items within atypical scenes (e.g., "camera" and "scissors" ; Aminoff et al., 2008). Although only retrosplenial cortex activation is observed in neural results, its activation actually reflects the relative strength of script networks formed through experience accumulation. Therefore, although key regions are not equivalent to schemas themselves, they can represent schema function.

In addition to left frontal cortex, research has also found the important role of middle/inferior temporal gyrus (Dennis et al., 2008), especially temporal pole, in semantic network modulation of lure memory representation (Chadwick et al., 2016; Zhu et al., 2019). Chadwick et al. (2016) used representational similarity analysis in left temporal pole and found that neural representational similarity between processing semantically related target word lists and processing lure words could significantly positively predict lure false alarm probability in another independent group of participants. This provides convincing evidence that temporal pole uses cross-subject schema representations activated by target word lists to promote formation of lure probe memory representations under the same schema.

(2) Modulation of lure probe memory representation/strength by script networks. Current research mostly demonstrates from behavioral level that script networks promote lure probe memory strength. Specifically, participants associative activate and/or gist extract lure events that are script-consistent but never actually occurred, producing false memories. For example, Bower et al. (1979) found that after participants read a text about going to a restaurant, those familiar with this script would erroneously recall the "ordering" episode that was not mentioned in the text (false memory for causally linked episodes). Hannigan and Reinitz (2001) found that after participants saw a scene depicting a woman pulling an orange from the bottom of a tall pile of oranges, they would false alarm to having seen a scene of oranges scattered on the floor (false memory for causally linked episodes). Friedman (1979) found that after participants saw a picture depicting a specific scene (e.g., children' s playroom), they would falsely recall typical items not presented in that scene (e.g., teddy bear; false memory triggered by typical scenes in scripts).

To our knowledge, only one study has examined script-based false memory for-

mation at the neural level. Aminoff et al. (2008) first presented participants with pairs of object pictures and required them to construct a common scene for each pair, classifying them into strong scene frames (e.g., "bulldozer" and "warning post") and weak scene frames (e.g., "camera" and "scissors") based on scene typicality. One day later, participants performed recognition on target pictures from strong/weak scene frames (e.g., bulldozer or camera), lure pictures related to strong scene frames but not presented (e.g., warning sign), and unrelated pictures not presented and unrelated to strong/weak scene frames (e.g., chandelier). Results showed that compared with unrelated pictures, participants false alarmed more to lure pictures, and brain regions for false memory formation (false alarm vs. correct rejection) overlapped with regions for typical scene processing (strong vs. weak scene frames), both activating medial prefrontal cortex, retrosplenial cortex, and inferior parietal lobule. Therefore, this study indirectly demonstrated that script networks activated by typical scenes promote memory strength for lure pictures under the same network. However, this study did not directly examine the relationship between schema representations and lure probe memory representations under the same schema, as Chadwick et al. (2016) did. Future research could use multivoxel pattern analysis to explore how cross-modal (Baldassano et al., 2017; Baldassano et al., 2018), cross-stage (Baldassano et al., 2017; Chen et al., 2017), and cross-subject (Baldassano et al., 2018; Chen et al., 2017) script representations promote formation of lure probe memory representations.

3.1.3 Retrieval Stage: Schema Reactivation

Research has found that when participants recognize schema-consistent target pictures (van der Linden et al., 2017) or recall schema-consistent target events (Chen et al., 2017), schemas are reactivated during retrieval and promote hits to target probes through enhanced functional connectivity with posterior cortex (Bonasia et al., 2018; Kesteren et al., 2010). This phenomenon also occurs when recognizing lure probes. For example, reactivation of key regions for semantic networks—left frontal cortex and middle/inferior temporal gyrus—during retrieval promotes false alarms to lure probes (Garoff-Eaton et al., 2007; Moritz et al., 2006; Webb et al., 2016). However, to our knowledge, no study has used multivoxel pattern analysis to directly examine the relationship between reactivated schema representations during retrieval and lure probe memory representations under the same schema.

Different from the above research approach, Buuren et al. (2014) found that the existence of spatial schemas made participants retrieve schema-related probes more slowly but more accurately, and this highly strategic retrieval process activated the superior parietal lobule responsible for top-down attentional control and the dorsolateral prefrontal cortex responsible for cognitive monitoring and evaluation. In false memory research, neural activity in these regions is commonly observed during recollection rejection of lure probes (Bowman & Dennis, 2016). Does this suggest that schema reactivation during retrieval may also

inhibit false memory formation through enhanced monitoring? Future research could further investigate the influence of schema reactivation during retrieval on false memory.

3.2 Relationship Between Schema and Distinctive Details

In Section 3.1.2, we discussed the strengthening process of schema-related memory representation/strength. In this section, we turn to the weakening process of schema-unrelated memory representation/strength and how this process may cause false memory formation (Spalding et al., 2015). In false memory research, distinctive details that distinguish targets from lures are usually not beneficial for schema identification and are therefore suppressed as schema-unrelated information. Research has found that during schema instantiation, angular gyrus related to schema processing inhibits hippocampal activity, causing the hippocampus to gradually withdraw from encoding target pictures, and stronger angular gyrus inhibition of hippocampus leads participants to false alarm more to lure pictures under the same schema (van der Linden et al., 2017). Doss et al. (2018) believe this is because enhanced schema processing fluency interferes with hippocampal pattern separation, preventing the hippocampus from forming distinctive representations for details and thereby increasing false memory (see Section 2.1). However, this does not mean that lacking distinctive representations only influences false memory formation as a byproduct of schema modulation of memory representation. When target presentation methods are insufficient to instantiate corresponding schemas (e.g., presenting two similar pictures with intervals), distinctive representations can also independently influence false memory formation (van den Honert et al., 2016).

---

4 Mechanism Three: Memory Updating from Misleading Information

As individuals constantly interact with the external environment, memories they form are influenced not only by internal schemas but also by environmental changes. Facing environmental changes, the dynamic memory system adaptively updates memory by flexibly incorporating new useful information from the environment into existing memory representations, thereby better guiding behavioral practice. However, sometimes new information in the environment is misleading information inconsistent with original events (i.e., target events). In such cases, the originally adaptive memory updating process may create a byproduct—false memory formation—due to incorporation of misleading information (Schacter et al., 2011).

Investigation of this type of false memory typically uses the misinformation interference paradigm (original event—misinformation—memory test). Specifically, after learning the original event, participants either relearn the modified original event containing misleading information or recall content of the original

event under misleading questions, thereby internally inducing reactivation of the original event context (see Section 4.2.1) and prompting participants to newly process and learn misleading information inconsistent with the original event (see Section 4.2.2). This causes participants to incorporate misleading information into memory for the original event during memory tests for the original event, showing intrusion of misleading information into original memory and false alarms to misleading information. Before elaborating on this process, we must clarify several boundary conditions that must be satisfied for memory updating to occur under memory reconsolidation theory.

4.1 Boundary Conditions for Memory Updating

Memory reconsolidation theory posits that reactivated consolidated memories undergo protein decomposition within minutes, during which memories temporarily return to an unstable state. Only after several hours of new protein synthesis do memories restore their original consolidated state (Hardt et al., 2010). Within the time window when original memories are reactivated and become vulnerable, original memories may undergo dynamic changes in memory strength, such as memory enhancement (Jonker et al., 2018; Koen & Rugg, 2016; Kuhl et al., 2010) and memory weakening (Kim et al., 2014), as well as dynamic changes in memory content, such as memory updating (e.g., Sinclair & Barense, 2018).

Memory updating from misleading information must satisfy three boundary conditions: (1) Reactivation of original memory. According to memory reconsolidation theory, only within the time window when reactivated original memory enters an unstable state can original memory undergo dynamic changes including memory updating. (2) Existence of misleading information beyond original memory content. Research shows that when original memory is interfered with by other information beyond its content and new learning of interfering information occurs, reactivated original memory is more likely to undergo changes in memory content rather than memory strength (Hardt et al., 2010). In the misinformation interference paradigm, participants' new learning of misleading information inconsistent with the original event provides necessary "material" for content change of original memory. (3) Moderate reactivation strength of original memory. The nonmonotonic plasticity hypothesis (NMPH) posits a U-shaped nonlinear relationship between memory reactivation strength and dynamic changes: weak reactivation does not trigger memory change, strong reactivation instead weakens memory, and only very strong reactivation enhances memory (Ritvo et al., 2019). In the misinformation interference paradigm, participants only learn the original event once, so original memory reactivation strength typically falls within the "weak–strong" interval (Kim et al., 2014). In this case, original memory is more likely to be weakened and, on this basis, interfered with by misleading information and undergo memory updating.

This paper adopts an information processing perspective, focusing on dynamic processes of encoding, storage, reactivation/reconsolidation, and retrieval of in-

formation from different sources. To avoid scattered content, this paper does not elaborate on reactivation strength, but this remains an important topic for future research. Below, we discuss how reactivation of original context and new learning of misleading information lead to false memory formation, where the former provides the time window for false memory formation and the latter provides content material. Both are indispensable.

4.2.1 Time Window: Reactivation of Original Context

The misinformation interference paradigm requires participants to relearn modified original events or recall content of original events under questions, thereby internally inducing reactivation of original context[6]. As an inherent cognitive process involved in the paradigm, reactivation of original context is difficult to disentangle from the paradigm itself, which somewhat hinders further exploration of the specific relationship between this process and false memory formation (misleading information updating/intrusion into original memory). The paradigm introduced below can better address this issue and further demonstrate that reactivation of original context is a critical window for memory intrusion, which is reflected in two aspects: whether this window opens is crucial, and when this window opens is also crucial.

[6] Reactivation of original context discussed in this section is a refinement of original memory reactivation in memory reconsolidation theory, particularly applicable to the misinformation interference paradigm. This is because original events involved in this paradigm are typically extremely rich in content (e.g., presenting a series of pictures depicting a daily activity), and any form of reactivation cannot perfectly reproduce all content of original events (Sinclair & Barense, 2019), but rather reflects reappearance of original event context (Jacques et al., 2013). In contrast, other classic paradigms in memory reconsolidation framework, such as the "AB-BC" paradigm, typically involve relatively simple original events (e.g., paired pictures), making them more suitable for examining how reactivation of original event content itself influences dynamic memory changes.

First, whether this window opens—i.e., whether original context is reactivated—is crucial. In Hupbach et al. (2009) study[7], all participants first learned a set of items (Set 1). Two days later, the reminder group was instructed to recall how they learned Set 1 two days earlier in the same experimental environment before learning another set of items (Set 2), whereas the no-reminder group directly began learning Set 2 in a different experimental environment. Two days later, all participants performed a recognition test and further recalled whether old items came from Set 1 or Set 2. The study found that compared with the no-reminder group, the reminder group falsely alarmed more to Set 2 items as coming from Set 1, but not more to Set 1 items as coming from Set 2. The reminder group' s unidirectional intrusion of Set 2 into Set 1 memory occurred because reactivation of Set 1 learning context when learning Set 2 opened a time window for Set 2 to interfere with Set 1 memory, making memory intrusion possible.

[7] In addition to memory reconsolidation theory, the temporal context model (TCM) can also provide a reasonable explanation for Hupbach et al. (2009) results. Specifically, learning two sets of items forms not only memory for items themselves but also associative memory between items and their temporal context. The reminder group reinstated the temporal context of Set 1 items before learning Set 2 items and performed new learning of Set 2 items within this time window. Set 1 context became bound with Set 2 items, causing Set 1 items and Set 2 items to become connected through sharing the same Set 1 context, thereby increasing the possibility of Set 2 items intruding into Set 1 memory. Moreover, the learning order relationship between Set 1 and Set 2 items allows Set 2 items to be bound with Set 1 context but not Set 1 items with Set 2 context, so the reminder group only shows unidirectional intrusion of Set 2 into Set 1 memory (Sederberg et al., 2011). TCM' s focus on context reinstatement and new learning of new items is largely consistent with memory reconsolidation theory. However, TCM does not assume that new information modifies and updates original memory but rather believes that they only strengthen connections through Hebbian synaptic effects while original memory content remains relatively intact. Moreover, TCM does not emphasize time-dependency of memory updating as much as protein synthesis/decomposition-based memory reconsolidation theory. Sinclair and Barense (2018) found that both theoretical models can jointly explain memory intrusion.

Second, when this window opens—i.e., whether original context is reactivated before or after misleading information presentation—is also crucial. Research shows that only when original context is reactivated before misleading information presentation (Sinclair & Barense, 2018) and continues into the misleading information presentation stage does it increase the probability of misleading information intruding into original memory (Jacques et al., 2013). Gershman et al. (2013) used multivoxel pattern classification[8] to accurately lock the timing of original context neural reactivation to 2 seconds before misleading information presentation. Using Hupbach et al. (2009) paradigm, this study found that the higher the degree of Set 1 learning context neural reactivation 2 seconds before Set 2 item presentation, the more likely that Set 2 item would intrude into Set 1 memory, and this predictive effect of neural reactivation on memory intrusion only occurred in the 2 seconds before Set 2 item presentation. Neural reactivation during and after Set 2 item presentation could not cause Set 2 items to intrude into Set 1 memory.

[8] This method borrows pattern classification technology from machine learning research, using spatial patterns in different cognitive states to train a classifier, then testing classifier performance with independent experimental data. This method not only has stronger detection power for cognitive representation differences but also allows researchers to infer participants' cognitive states from neural signals based on evidence provided by the classifier, facilitating deeper understanding of how cognitive states are represented in the brain (Lei et al., 2010).

4.2.2 Content Material: New Learning of Misleading Information

Lee (2009) pointed out that memory reconsolidation does not simply refer to the process from reactivation to automatic restoration of stable state but rather provides conditions for continuous memory modification and updating. Therefore, memory updating only occurs when there is interfering information beyond original memory content. New learning of this information provides necessary material for content change of original memory. Misleading information in the misinformation interference paradigm is inconsistent with original events and thus also belongs to the new information required for memory updating. Investing more cognitive resources and performing more cognitive/neural processing when encoding this information promotes false alarms to misleading information and produces false memory.

Okado and Stark (2005) used the misinformation interference paradigm and found that if participants activated the left posterior hippocampus and perirhinal cortex more when encoding original events, they were more likely to reject lure events. Conversely, if they activated these regions more when encoding lure events containing misleading information, they were more likely to false alarm to lure events and produce false memory. This interaction suggests that the left posterior hippocampus and perirhinal cortex are responsible for processing specific detail information distinguishing original events from lure events (e.g., whether a thief stole a girl' s wallet and hid behind a "door" or a "tree"). Rejecting lure events results from distinctive detail representation when encoding original events (see Section 2.1), whereas false alarming to lure events results from excessive cognitive/neural processing of misleading details when encoding lure events. If encoding lure events could activate the default mode network more, such as anterior/posterior cingulate and left precuneus, thereby disengaging from current cognitive/neural processing of misleading information, participants would be more likely to reject misleading information in lure events and reduce false memory (Baym & Gonsalves, 2010).

---

5 Conclusion and Reflection

From an information processing perspective, this paper elaborates on possible causes of false memory during encoding, storage, reactivation/reconsolidation, and retrieval of information from different sources (target events, internal schemas, and external interference). On the one hand, this makes our discussion of cognitive mechanisms of false memory emphasize the dynamic nature of information processing more than previous reviews and be less limited to specific types of false memory (e.g., Liu et al., 2015). On the other hand, when discussing neural mechanisms of false memory, we did not follow the traditional approach of examining similarities and differences in neural activity between false and true memories to 描绘 false memory' s "neural portrait" (e.g., Dennis et al., 2015). Instead, we aimed to reveal causes of false memory formation.

Therefore, we approached from an information processing perspective of information and consistently discussed within the information processing framework, enabling us to more comprehensively and systematically present causes of false memory formation. Moreover, to our knowledge, previous reviews of false memory neural mechanisms almost without exception only covered studies using univariate activation analysis (Dennis et al., 2015; Johnson et al., 2012; Straube, 2012). Constrained by methodology, although such studies can better reveal causes of lure probe memory strength formation, they are quite limited in revealing specific content of memory (Davis & Poldrack, 2013). To investigate memory content, it is necessary to use multivoxel pattern analysis to extract information contained in corresponding neural representations. This paper cites numerous cutting-edge studies using this analytical method, enabling us to further reveal causes of lure probe memory representation formation from the content level. This is also a good supplement and extension to previous reviews.

Summarizing the full text, we believe false memory may arise from three causes: First, the hippocampus fails to form and reactivate distinctive detail representations for targets during encoding and retrieval, leaving only abstract memory representations reflecting common features of targets and lures. This makes participants more likely to rely on the latter to reconstruct missing detail information, increasing false memory. Additionally, the specific relationship between neural representations in angular gyrus during retrieval and false memory formation is currently a hot topic in the field. Future research could examine the relationship between distinctive representations in angular gyrus and false memory using false memory paradigms. Moreover, since neural representations in angular gyrus are modulated by task demands (Kuhl et al., 2013), future research could also investigate the influence of functional interaction between angular gyrus and prefrontal cortex on false memory formation. Furthermore, researchers could consider this question: Does the existence of distinctive detail representations necessarily lead to reduced false memory? For example, can distinctive details of visually presented target words guide memory discrimination between auditorily presented targets and lures?

Second, after target items instantiate schemas, semantic networks (inferior frontal gyrus, middle/inferior temporal gyrus, temporal pole, etc.) and script networks (medial prefrontal cortex, posterior cingulate cortex, retrosplenial cortex, precuneus, inferior parietal lobule, etc.) enhance lure probe memory representations related to schemas and weaken target item detail representations unrelated to schemas, leading to false memory formation. Current research trends in this direction include: (1) Compared with script networks, more focus on false memory formation driven by semantic networks; (2) Compared with the influence of schema reactivation during retrieval on false memory, more focus on how schemas modulate lure probe memory representation during encoding. Future research could use multivoxel pattern analysis to explore how script representations promote formation of lure probe memory representations, and whether reactivated schema representations during retrieval increase false mem-

ory by promoting lure probe memory representations under the same schema, or inhibit false memory by enhancing monitoring of lure probes. Additionally, researchers could further consider this question: Do schemas necessarily increase schema-consistent false memory? Hannigan and Reinitz (2001) found that participants with scripts for causal event sequences would falsely recall event outcomes after being presented with causes. However, the existence of scripts also made participants form expectations for causes after being presented with outcomes (though causes were not actually presented), thereby reducing false alarms to causes. Does prediction error formation caused by schemas inhibit false memory formation? What role does the hippocampal CA1 subregion, responsible for automatically monitoring match/mismatch between expectations and actual events, play in this process (Chen et al., 2011)?

Third, new learning and processing of misleading information within the time window of original context reactivation interferes with memory reconsolidation, causing participants to update misleading information into original memory and produce false memory. Current research in this direction pays less attention to retrieval processes of misleading information. Some studies have found competition between memory representations during retrieval, and relative representation strength can directly predict participants' behavioral reports (Kuhl et al., 2012; Kuhl et al., 2011). Following this line of thinking, future research could examine whether competition exists between misleading and original information during retrieval and how this competition influences false alarms to misleading information. Additionally, although not the focus of this paper, investigating reactivation strength of original memory may suggest feasible methods for reducing false memory. For example, according to the nonmonotonic plasticity hypothesis (see Section 4.1), overlearning original events to make original memory reactivation reach the range for memory enhancement may help inhibit memory updating from misleading information. Moreover, Bridge and Voss (2014) found that only memories dominating active retrieval would combine with new information. In misinformation interference paradigm, when participants actively recall original event content, they typically prioritize extracting event context rather than specific details (Putnam et al., 2017). At this time, is the combination between dominant context memory and misleading information, rather than detail memory, the deep cause of false memory formation? Investigating this question could deepen our understanding of misleading information updating and false memory formation.

Acknowledgments: We thank Anastasia Besika for carefully proofreading the English abstract.

References (Chinese): Chen, H., Guo, C., & Yang, H. (2015). Effects of delay interval and retrieval conditions on short-term false memory. *Studies of Psychology and Behavior*, 13(1),

Jiang, R., & Li, X. (2015). Developmental reversal of false memory: Why does it increase with age? *Advances in Psychological Science*, 23(8),

Lei, W., Yang, Z., Zhan, M., Li, H., & Weng, X. (2010). Decoding cognitive neural representations using multivoxel pattern analysis of brain imaging: Principles and applications. *Advances in Psychological Science*, 000(012), 1934-1941.

Liu, Z., Liu, T., Han, J., & Mu, S. (2015). Implantability of false memory. *Advances in Psychological Science*, 23(5), 806-814.

Wang, M., & Geng, H. (2010). The adaptive nature of memory from a lifespan developmental perspective of associative memory illusions. *Chinese Science Bulletin*, 55(4), 307-315.

[The English references that follow are already in English and should be preserved exactly as given]

Note: Figure translations are in progress. See original paper for figures.