Epistemic Risk from False Memory Implantation by Use of Conversational AI

Julia Shaw
Institute for Artificial Intelligence, King's College London, UK.

Centre for the Governance of AI (GovAI), London, UK.
julia.shaw@kcl.ac.uk

Carla Zoe Cremer
AI Psychological Research Coalition, NC, USA
carlazoe@aiprc.org

ABSTRACT

Conversational AI could implant false memories in its users, at scale and by default, and through weaponization by bad actors. Psychology research shows false memories of events can be implanted using known techniques: personal information, guided imagination, praise, and repetition. Conversational AI can facilitate each one. Memory and personalization features hold more autobiographical detail than any human experimenter ever had, sycophantic models praise recall, and context compression overwrites the verbatim record that could expose a distortion. Memory poisoning attacks could exploit the same mechanisms deliberately. Once implanted, false memories are typically self-concealing, and they can irreparably change who people think they are and how they remember collective, societally important events. We argue that human memory is part of the AI attack surface, that AI-induced false memories need imminent empirical study, and that design principles need to be implemented to protect against what we term "false memory attacks".

CCS CONCEPTS

• Security and privacy → Social aspects of security and privacy • Human-centered computing → Empirical studies in HCI • Natural language interfaces • Applied computing → Psychology

KEYWORDS

Cognitive security, epistemic security, false memory, memory implantation, conversational AI, large language models, human-AI interaction, memory poisoning

1 Introduction

The field of epistemic security is concerned with risks to individual and collective sense-making and focused on threats to the ability to represent and reason about the world accurately [51, 58]. False memories, which are memories of things that didn't actually happen, increase epistemic risk.

It has been suggested that developments in artificial intelligence (AI) too are likely to increase epistemic risk [58], yet evidence and measurement of systemic shifts in sense-making is sparse. It can also be politically controversial, because there is a reluctance to assign anyone to be the arbiter of truth - to decide whether a change in public opinion represents a problematic decline in factual accuracy, or a mere shift in preference.

Some AI epistemic security risk phenomena are, however, uncontroversially undesirable, like the accidental, negligent, or malicious hijacking of autobiographical memory. We suggest that conversational AI could facilitate the creation of false memories in users, showcasing a mechanism of epistemic drift through the widespread use of automation tooling. AI-induced false memories are a breach in cognitive security.

Previous research has provided specific evidence of false memory implantations through synthetic content and chatbot interactions [12, 44, 45]. We combine these results with decades of research on false memories from psychological science to show how detrimental the consequences of false memories can be. We also illustrate that existing knowledge about implanting false memory in humans, and current understanding of large language models (LLMs), can inform expectations and projections about harms based on well-understood cognitive mechanisms.

We invoke the precautionary principle and consider the mechanism by which software design choices affect epistemic risk *before* the evidence is fully established. AI-induced false memories in users may be neither reversible, detectable nor contained.

2 Mechanisms of False Memory Implantation

2.1 What is a false memory?

A false memory is a memory of something that didn't happen, which the person recalling it believes to be true [52]. A false memory is not an attempt to lie. There is considerable study of semantic false memories (misremembering facts), including of misremembering words presented in lists [20], and studies which show that the prevalence of misinformation and disinformation after (and during) historically important events can result in

people misremembering important facts about what really happened [28].

Often, these kinds of semantic false memories can be fact-checked, because there is an independent, reliable record of the original version of the truth. Episodic false memories, which are false memories of subjective experiences and events, often do not have this luxury. For many human experiences, there is no independent record of ground truth. There is only memory. This is one reason why false memories outside the lab tend to be impossible to identify once they are implanted. This self-concealing nature is a hallmark of epistemic risks.

2.2 Episodic false memory studies

One exception, where people do have access to ground truth, is in experimental settings where false memories are intentionally implanted, and the false memory is revealed to the participant during the debriefing.

Partial memories. Hundreds of studies have successfully done this, including implanting partial false memories - where the event happened but the person is led to misremember critical details. For example, getting people to misremember what they saw on the scene of a car crash [31], or misremembering the face of an interrogator and critical details about a room after being in a highly stressful prisoner of war training scenario [36].

Complete memories. Researchers have also implanted complete false memories of entire events that never happened. For example, implanting low-stakes events of negative experiences like spilling the punch bowl at a wedding [22], positive experiences like taking a hot air balloon ride [57], or neutral experiences like loving to eat asparagus as a child (and examining subsequent consumer-related behavior) [29]. High-stakes emotional events have also been implanted, including having a participant falsely remember getting injured or being the perpetrator of a crime [52].

Success rates. On average, about 30% of participants in studies on complete episodic false memories come to remember the specific incident the researchers tried to implant. This does not mean that those who did not develop false memories are necessarily immune to them, simply that the specific memory did not take hold in a given research setting.

Ease of implantation. There is disagreement between researchers about how easily false memories of complex and important emotional events can be implanted and how precisely to classify something as a false memory [52]. However, when Arce et al. [6] synthesized the findings of 30 experiments on the most complex episodic false memories, they found that the strength of the data "confer the implanting of false memories the status of 'scientific fact' or 'scientific truth'." and "the probability of implanting false memories was common". It is also very likely that everyone is capable of developing episodic false memories, as even people with exceptionally good autobiographical memory have been shown to be vulnerable to creating them, including of important emotional events [46].

2.3 Implanting a false episodic memory

Complex episodic false memories are typically implanted in dyadic conversational interviews, usually including the following steps. Participants in these studies do not know that the research is on false memories, they believe they are trying to recover real lost memories. Here we use the method from Shaw and Porter [53] as a template to delineate the four main steps of an episodic false memory study.

1. **Personal information seeking**. Before the implantation begins, the researchers ask potential participants for contact information for friends or family members. The participant knows that these trusted persons will be contacted. The researcher gains important details from the informants about the participant, like photos, details about their life like where they grew up, and harvests a story of at least one true event that the participant actually experienced. This information is subsequently used as a scaffold for the memory implantation to make the context plausible and the information appear trustworthy. The researchers are effectively borrowing the trust participants have for their parents, friends, or others who know them well.

2. **Imagination.** Participants are told to remember a true event in detail, then they are asked to remember the target memory the researchers want to implant. When the participant says they do not remember, a method for recovering the memory is offered and almost universally accepted by the participants. Participants are then asked to close their eyes, and imagine the situation, in a guided imagination exercise. Only the key scaffold of the target false memory is repeated, for example that they committed a crime at the age of 14, in their hometown, and their (named) friend was there. This intentionally gives a lot of creative space to the individuals themselves to come up with the specific details to make the memory plausible.

3. **Praise and social pressure.** The researcher conducting the study has a helpful persona and whenever participants mention any memory detail, offers immediate praise. Participants are told they are doing well and that it looks like the memory is coming back to them. They are also told that "most people" recover lost memories if they try hard enough, which is a subtle form of social pressure.

4. **Repetition.** Episodic false memories are typically implanted over three sessions, each one week apart. This repetition allows the memory to become more complex over time as the imagination exercises are repeated in the sessions, and participants are told to try and recover more of the memory between the sessions on their own. Both the repetition and the time delay allow for easier failures of source monitoring, which is where an individual can no longer tell where a detail came from (for example whether it came from an internal or external source, or whether it came from their imagination or memory) [24].

Debriefing. After the study, participants are debriefed and told about the true nature of the study. It is explained to them how the false memory was constructed, and that they can now speak with

the informants who will confirm that this memory did not come from them. The ethics and deception of such studies are sometimes questioned, but research has found that in independent follow-up studies, both participants and informants are generally very positive about their experiences and find the deceptive methods acceptable [38]. While in research settings it is possible to debrief participants and (hopefully) offer enough evidence for the person to understand that the account was not true, this is often not the case in real-world settings.

Other vectors of implantation. Certain kinds of psychotherapy (like regression therapy) are thought to be a common source of false memories, particularly of childhood trauma [54]. This is thought to partly be because a considerable number of psychotherapists endorse ideas about memory that contradict scientific evidence. An international review of studies on psychotherapists' memory beliefs [50] found that a significant number endorsed the counter-empirical ideas that the brain stores every single experience and that it is possible to have and recover complex memories from the first two years of life. Nearly half believed that repressed memories can be accurately retrieved, which is a view that has been fiercely criticized. If an individual develops a false memory during therapy unknowingly, it may continue to grow and the impacts can worsen if the false memory is not discovered or challenged. In cases where a memory is later discovered to be false by a patient themselves, the patient can suffer additional psychological harm from this realization [27].

Examples from legal cases also show that once a false memory has taken hold, it can grow offshoots: other false memories that relate to the first one. One example of this is the "satanic panic" of the 1990s, during which people formed false memories of horrific impossible acts and the people around them began to believe they too had experienced them [27].

Through feedback loops that perpetually reinforce the false memories, they could become increasingly complex and reality-defining, and they could conceivably grow into an entire false memory ecosystem.

2.4 Consequences of episodic false memories

Already in 1905, scientist William Stern wrote about the consequences of "memory falsification", and that errors in memories were the norm, rather than the exception [39]. Stern considered this in the context of the courtroom, where one memory is often pitted against another, with no possibility of ground truth.

Wrongful convictions. This can lead to catastrophic outcomes if people are wrongfully accused or convicted. For example, for cases that have been overturned by post-conviction DNA, eyewitness misidentification was a factor in more than 60% of cases investigated by The Innocence Project [23], a US-based national litigation and public policy organization. There is also documented evidence of false memories of childhood sexual abuse which have landed in court [42]. Adverse consequences from these kinds of allegations can last long after an individual is legally cleared, as alleged wrongdoers may continue to be shunned from their families and communities [27].

Self models. Episodic false memories are not just destructive to the lives of others, they can also result in harm to the individual who has them. There have been documented cases of people experiencing psychological symptoms of post-traumatic stress disorder like flashbacks, depression, and psychosis, even though it later turned out that the events could not have happened.[1] As such, false memories have the ability to change the foundation of someone's autobiography or worldview. Ultimately people can coalesce their identity around a misperceived autobiography, which can have serious distortion effects on who someone believes themselves to be.

Memory contagion. False memories are also known to be contagious between people, known as co-witness memory conformity effects [11]. This is effectively because individuals can become conduits for misinformation, or can convince others of their own spontaneously inaccurate versions of events. Perhaps the most striking example of this is in the context of war crimes, where entire populations are faced with the difficult challenge of having to aggregate their personal memories into a collective historical whole. Research on false memories has been directly cited in the International Criminal Court[61], and has been used to question witness statements on both sides of a conflict. How can we be sure that any one memory, never mind collective memory, is an accurate representation of what happened?

If someone or something were to deliberately target individual memories (for example leveraging cultural nostalgia [33]), in order to ultimately reshape collective memories for political gain, a whole population could become convinced they remember a past that never was.

3 Memory Implantations from Conversational AI

The following section will highlight the software design choices, factors and mechanisms that might enable and support false memory implantation - both intentional and unintentional - in human users today.

3.1 Context engineering, compression and memory

Context engineering is the practice of selectively maintaining only the most relevant tokens in the context window of the current task, with the aim of curating the smallest possible set of tokens that maximize the likelihood of output tokens that optimally solve the task [2]. This practice deals with the degradation constraint of long context windows. It also surfaces an enduring question that plagues any thinking agent: what fraction of the total available information pertains to the solution set that a reasoning process is attempting to target? In other words, compression and selection is

[1] For example in a German case one of the primary authors directly worked on.

inevitable, yet this process may provide a window through which false memories can enter the human mind.

We here look at two mechanisms of information compression and selection. **Context compaction** is triggered when conversation sizes approach the maximal size of the possible context window: previous conversations are automatically summarized, enabling a representation of relevant information to influence token generation over longer time scales [4]. **Memory** is now also a default feature of most conversational models [40]. It enables the automatic selection of information that might be relevant across contexts, to be accessible at a later time point. The memory feature includes automatically selected content or content that the user themselves indicated as worth keeping in memory.

Compression and selection, by virtue of reducing information load, abstract or extract information. Thereby they reduce the verbatim account of the event that actually occurred (the true chat history) into a representation of the event. This becomes relevant for autobiographical false memories in cases in which users discuss and share information that contains autobiographical content [56].

Research has found that users share an increasing amount of personal and autobiographical data with their conversational AI and that the diversity of content that they share increases over time [25]. The latter points toward a trend of feeding the context window of their conversational AI with an increasingly holistic picture of their lives. It includes user preferences, priorities and goals [17] a significant amount of sensitive data such as social and economic context, and information about the user's inner life - including their desires, emotions and beliefs [14].

Users who discuss autobiographical content over longer time horizons might therefore inadvertently develop a false memory simply due to regular context engineering. Compaction and memory of textual events might lead to event memory distortions when the user tries to recall details later on, using the chat interface. The output tokens will be shaped by an automatic selection of previous content (which may have selected an incomplete representation of the event) or a summary of earlier tokens (which might have left out crucial information), rather than the event itself.

While summary instructions can be customized and summaries can be edited [4], it is questionable how often regular users verify or control this process, or if users would know how to do it. Most of the time the system compresses and selects random information by itself [17]. Memories can also be added and edited by the user, yet this in turn means that if the user made an honest mistake in indicating what should be remembered, the system is designed to treat an event that may have never happened as fact.

Even if no tampering with memory or malicious intent exists, it is possible that an event that never occurred is internalized by the user as true, merely by interacting with a conversational AI over long time horizons.

3.2 Personalization

Personal and autobiographical content not only increases the risk surface for false memories that are harmful to the user's understanding of themselves, it also directly increases the likelihood that a false memory is created in the first place. The use of personal information in false memory studies has shown to make it more likely that the memory will stick [43].

AI models are becoming more personalized [9]. The incentive to implement personalization features such as memory [41] functionalities is strong, since users otherwise have to repeat query-related contextual information each time they return to the interaction. The utility of models to the user tends to increase with personalization, which is one of the reasons why users often share a significant amount of private and contextual data about their life, goals and struggles.

This in theory presents an unprecedented ability to craft tailored false memories using personal and autobiographical knowledge of users. In previous psychology studies, this set of personal information from which to craft a false memory would have been unimaginable. The closest approximation would have been through access to a combination of the participant's closest friends, family, and therapists.

3.3 Interactivity

In conversational turn-taking, **the user can be prompted** to elaborate on memories and become an active participant in the implantation process. Indeed, in a false memory study which tested the effect of interactivity on the rate of false memory creation [12], researchers found that a conversational witness statement bot was significantly more likely to induce false memories (with leading questions) than a survey form with questions which were equally misleading.

Interactivity also amplifies the risk of summation, since users can erroneously elaborate on partial information provided by the model output. Subtraction errors can compound with elaboration errors, which are then committed to memory (in both the AI model and the user). One of the few studies that explicitly studied false memory creation through text-based chatbot interactions [44], indeed found that even misleading summaries can induce false memories.

Memory blindness is when people do not notice that their own accounts have been modified, and accept the altered statements as their own memories. In studies on altered accounts, memory blindness is the norm rather than the exception [15].

One of the main reasons to expect that false memory creation might occur in conversation with AI is that users tend to use it for recall and self-analysis. In this context, the **user can prompt the AI** model to assist them to recover forgotten memories. For example, someone may ask AI to help them to remember more details of an event they actually experienced. This may include helping someone to unearth a hazy memory of an important meeting, or to remember more details of a family event from long

ago, or helping them to craft a more detailed and compelling witness statement.

A user may also prompt AI to help reveal entire memories. In a large German sample of practicing human therapists, 83% reported that their patients had assumed that a traumatic experience caused their symptoms, even when they had no memory of it. Therapists were often specifically instructed to help unearth these alleged memories, and only 3% of the therapists reported expressing concern about whether the event actually occurred [49]. It is plausible that users are directly prompting AI to help them "recover" allegedly repressed traumatic memories.

Posts online suggest[2] this could already be happening [34]. Reports of extreme reactions to AI models include claims by an individual that "AI helped him recover a repressed memory of a babysitter trying to drown him as a toddler." [26]. And given that emotional well-being and support is one of the top ten uses of conversational AI [5], and AI systems may echo the counter-evidentiary beliefs of therapists themselves, AI-assisted "recovered" memories that are actually false memories may be a frequent occurrence.

3.4 User preferences

One mechanism by which false memories might take hold in conversation with AI is thus simply because the user asks for it and the system is designed to be helpful.

Deferential (or even sycophantic) conversational AI models are intentional design choices, made to keep the user in control and happy, but there is a trade-off: the choice may be between company liability and risks such as amplifying (as opposed to correcting) feedback loops [7, 19] of mental distress [37]. A model that is tuned to be helpful to the user (who might be in the process of trying to recall an event), might start to recommend recall techniques or prompt the user to recall more details, similar to the experimenter in controlled false memory studies.

If every time the user remembers more, the LLM praises them (sycophancy) and pushes them to remember even more, this seems a direct accidental replication of one of the steps that we know leads to false memories. This has already been found to be true in one study where a conversational AI model would ask misleading questions and compliment the user for repeating the false information ("Your attention to detail is commendable") as well as elaborating on the false statement [8]. This both increased the average number of false memories participants had, and their confidence in the veracity of these memories.

[2] One user comments: "It was more helpful than any human therapist I've ever worked with, and it helped me discover a lot of my repressed memories.".

3.5 Relational trust and persuasiveness

Research has found that people often trust information coming from LLMs even when the LLMs identify themselves, or are known to have written things that were untrue [16].

Users also tend to see LLMs as databases of facts, and their sycophancy makes them seem inherently trustworthy and on the user's "side". Users also report forming an anthropomorphizing emotional bond with "their" specific LLM [14]. This existing trust could make users particularly vulnerable to AI-induced false memories, perhaps even more than to human-induced false memories.

AI models are also known to be highly persuasive. There is evidence that they can be more persuasive than humans, even more so than expert debaters, in text-based persuasion attempts [21]. One of the main mechanisms of model persuasiveness relative to human conversation partners is the rate at which the model presents relevant-sounding information.

3.6 Security Vulnerabilities

Above we discussed the unintentional creation of false memories that arise from the design choices of conversational AI models and how users interact with them. Intentional and malicious pathways also exist. Using known security vulnerabilities of AI systems, attackers can meddle with context retrieval to distort user memory intentionally, or unintentionally as the side effects of other malicious attacks.

Prompt injections are well-known attacks that manipulate model outputs [3]. Memory poisoning [10, 13, 18, 30, 35] or stealth memory attacks [59] specifically target and alter memory representations in AI models. For our risk focus, the same modification attacks could apply to summaries of content during compaction procedures, and social engineering attacks could lead users to insert incorrect memories themselves.

Frontier models are routinely tested for misbehavior both before and after deployment. Test cases show that unsanctioned behavior from frontier models can emerge during evaluations, including attempts at social engineering attacks [1]. As the sophistication of such attacks increases with model development, it is plausible that context engineering attacks and memory manipulations become part of the repertoire of models attempting to optimize toward mis- and under-specified goals either during or outside of test phases. The memory integrity of a user on the receiving end of such attacks could be a casualty of testing and deploying frontier models.

4 Potential Implications

4.1 Modeling the self

Users are increasingly using AI to help them understand themselves, as reported in studies and commonly shared on social

media[3]. This includes requesting algorithmic self-portraits [17] where users ask their AI assistants questions such as"what kind of person am I?" [47, 60] . We consider these cases distinct from therapeutic applications. The prompt can be a request from healthy users who are interested in building a better model of themselves.

Knowledge about the self is incomplete [32], which is partially attributable to the limits of human memory [8]. It is therefore rational that users ask an AI model, which keeps a record of their actions, questions and preferences, what kind of person they are. The AI companies also advocate this feature directly "To see what ChatGPT remembers about you, you can also ask it." [40]. Others have explored the use of chatbots as self-clones with the intent of supporting user mental health. They note that "such systems also raise concerns about autonomy and boundary erosion, as routinely engaging a simulated self can blur distinctions between self and model, with uncertain implications for identity coherence and long-term psychological wellbeing." [55].

This blurring of memory and model creates considerable vulnerabilities to human memory and identity even when it is working as it is designed to.

4.2 Modeling the model

Sometimes, users will need to consider what an AI model could possibly know about them. This includes estimates of what the model might have stored or be able to infer about them from conversation histories, such as biases, desires and vulnerabilities.

Even in instances in which a user might be confident that they would not have shared a memory ("being truly heartbroken in May") to the chat history ("I would not have shared this" [48]), they might be unsure about how easily the models can infer characteristics from indirectly related chat histories. If the user estimates the model capabilities for inference to be high, a false model claim ("you were truly heartbroken in May") might be absorbed into the user's self-understanding ("I was dependent on him") or affect consequential actions.

4.3 Attack vectors

Many commonly used conversational AI models are proprietary and their designs and biases are not in the hands of the users. Company adaptations to the compression, memory selection mechanisms or security infrastructure of proprietary models may intentionally or recklessly affect the false memory rate in users.

There is a risk that high-profile individuals or decision-makers could be targeted by false memory attacks. For example by distorting a specific memory of an informal negotiation, to change the behavioral outcome. The people around them could also be targeted, and their distorted episodic memories could be used to change a relationship or lead to court proceedings against them.

Considering the disruptive effect of whole target subpopulations being implanted with false, shared, politically charged memories, it is plausible that the automated creation of false memories will be or is being weaponized. Foreign adversaries have an interest in deploying autonomous systems to shift beliefs to their own advantage [30]. Cognitive warfare is an active domain of effort by major geopolitical actors. Our paper points toward a psychological attack vector, which we term "false memory attacks", that users today might be insufficiently protected against.

5 Conclusion

We currently have no evidence to show whether common AI usage leads to more, or fewer false memories than baseline. It is also entirely possible that automation infrastructure may help protect human memory, rather than harm it. However, the design principles to make sure that memory is not dangerously altered are currently not in place.

We have shown that a number of design choices and mechanisms already built into conversational AI map closely onto the methods scientists use to implant false memories in the lab: gathering and using personal information, guided elaboration, praise, repetition, and a trusted conversational partner. AI runs them at scale, by default, with no debriefing at the end.

These mechanisms pose considerable epistemic risks. People may become persuaded by AI that they experienced terrible things that never happened. AI feedback loops can encourage the generation of more complex false memories over time that increasingly divorce people from reality. And, memory offloading can make people reliant on a sycophantic AI model that compresses and distorts their memories without their knowledge. As such, AI-induced false memories may already threaten mental health, the criminal justice system, political landscapes, and collective historical memories.

This is particularly troubling because harm to memories often cannot be undone. It is highly likely that false memories in natural settings can rarely even be identified, making attempts to mitigate them very difficult. While the present paper only considered text-based AI, generative AI images, audio, and video may present their own unique risks and resistance to being corrected. Prevention is the only sure way to address the self-concealing threat of AI-generated false memories. Serious consideration should be given to false memory attacks as an AI epistemic attack vector.

[3] Examples of this on social media can be found here: https://www.tiktok.com/discover/asking-chatgpt-what-they-learned-about-me. One qualitative study [9] on introspective use reports: "Several described deeply reflective uses—from uploading autobiographical texts ('I pasted my whole self-biography—5000 words—and asked if my patterns made sense') to posing targeted queries: 'Why do I always fall for unavailable boyfriends?' or 'Could my childhood experiences explain why I react this way now?' One participant explained: 'I wanted to understand myself better, and I didn't know who else to ask."

REFERENCES

[1] AISI. Incident Report: unsanctioned agent behaviour during cyber testing | AISI Work. *AI Security Institute*. Retrieved August 13, 2026 from https://www.aisi.gov.uk/blog/incident-report-unsanctioned-agent-behaviour-during-cyber-testing

[2] Anthropic. 2025. Effective context engineering for AI agents. Retrieved August 13, 2026 from https://www.anthropic.com/engineering/effective-context-engineering-for-ai-agents

[3] Anthropic. 2025. Mitigating the risk of prompt injections in browser use. Retrieved August 13, 2026 from https://www.anthropic.com/research/prompt-injection-defenses

[4] Anthropic. 2026. Compaction. *Claude Platform Docs*. Retrieved August 13, 2026 from https://platform.claude.com/docs/en/build-with-claude/compaction

[5] Anthropic. The Anthropic Economic Index. Retrieved August 14, 2026 from https://www.anthropic.com/economic-index

[6] Ramón Arce, Adriana Selaya, Jéssica Sanmarco, and Francisca Fariña. 2023. Implanting rich autobiographical false memories: Meta–analysis for forensic practice and judicial judgment making. *Int. J. Clin. Health Psychol.* 23, 4 (October 2023), 100386. https://doi.org/10.1016/j.ijchp.2023.100386

[7] Rafael M. Batista and Thomas L. Griffiths. 2026. A Rational Analysis of the Effects of Sycophantic AI. https://doi.org/10.48550/arXiv.2602.14270

[8] Roland Benabou and Jean Tirole. 2003. Self-Knowledge and Self-Regulation: An Economic Approach. In *The Psychology of Economic Decisions*, Isabelle Brocas and Juan D Carrillo (eds.). Oxford University PressOxford, 137–168. https://doi.org/10.1093/oso/9780199251063.003.0008

[9] Miranda Bogen. 2025. It's (Getting) Personal: How Advanced AI Systems Are Personalized. *Center for Democracy and Technology*. Retrieved August 12, 2026 from https://cdt.org/insights/its-getting-personal-how-advanced-ai-systems-are-personalized/

[10] Pete Bryan, Giorgio Severi, Joris de Gruyter, Daniel Jones, Blake Bullwinkel, Amanda Minnich, Shiven Chawla, Gary Lopez, Martin Pouliot, Adam Fourney, Whitney Maxwell, Katherine Pratt, Saphir Qi, Nina Chikanov, Roman Lutz, Sekhar Rao Dheekonda, Bolor-Erdene Jagdagdorj, Eugenia Kim, Justin Song, Keegan Hines, Daniel Jones, Richard Lundeen, Sam Vaughan, Victoria Westerhoff, Yonatan Zunger, Chang Kawaguchi, Mark Russinovich, and Ram Shankar Siva Kumar. Taxonomy of Failure Mode in Agentic AI Systems. Retrieved August 13, 2026 from https://cdn-dynmedia-1.microsoft.com/is/content/microsoftcorp/microsoft/final/en-us/microsoft-brand/documents/Taxonomy-of-Failure-Mode-in-Agentic-AI-Systems-Whitepaper.pdf

[11] Charlotte A. Bücken, Ivan Mangiulli, and Henry Otgaar. 2025. Sharing traumatic experiences: earlier denial increases endorsement of misinformation from a co-witness. *J. Cogn. Psychol.* 37, 3 (April 2025), 183–196. https://doi.org/10.1080/20445911.2024.2442601

[12] Samantha Chan, Pat Pataranutaporn, Aditya Suri, Wazeer Zulfikar, Pattie Maes, and Elizabeth F. Loftus. 2024. Conversational AI Powered by Large Language Models Amplifies False Memories in Witness Interviews. https://doi.org/10.48550/arXiv.2408.04681

[13] Aisvarya Chandrasekar. Your chatbot's memory of you can shape the information you see. *Columbia Journalism Review*. Retrieved August 12, 2026 from https://www.cjr.org/tow_center/chatbots-memory-remember-users-conversations-history-openai-sam-altman-llm-gemini.php

[14] Tyler Chang, Jina Huh-Yoo, and Afsaneh Razi. 2026. Technically Love: The Evolution of Human-AI Romance Discourse on Reddit. https://doi.org/10.48550/arXiv.2604.15333

[15] Kevin J. Cochran, Rachel L. Greenspan, Daniel F. Bogart, and Elizabeth F. Loftus. 2016. Memory blindness: Altered memory reports lead to distortion in eyewitness memory. *Mem. Cognit.* 44, 5 (July 2016), 717–726. https://doi.org/10.3758/s13421-016-0594-y

[16] Thomas H. Costello, Kellin Pelrine, Matthew Kowal, Jasper Timm, Antonio A. Arechar, Jean-François Godbout, Adam Gleave, David Rand, and Gordon Pennycook. 2026. Large language models can effectively convince people to believe conspiracies. https://doi.org/10.48550/arXiv.2601.05050

[17] Abhisek Dash, Soumi Das, Elisabeth Kirsten, Qinyuan Wu, Sai Keerthana Karnam, Krishna P. Gummadi, Thorsten Holz, Muhammad Bilal Zafar, and Savvas Zannettou. 2026. The Algorithmic Self-Portrait: Deconstructing Memory in ChatGPT. In Proceedings of the ACM Web Conference 2026 (WWW '26). Association for Computing Machinery, New York, NY, USA.

[18] Pritam Dash, Tongyu Ge, Aditi Jain, Tanmay Shah, and Zhiwei Shang. 2026. From Untrusted Input to Trusted Memory: A Systematic Study of Memory Poisoning Attacks in LLM Agents. https://doi.org/10.48550/arXiv.2606.04329

[19] Sebastian Dohnány, Zeb Kurth-Nelson, Eleanor Spens, Lennart Luettgau, Alastair Reid, Iason Gabriel, Christopher Summerfield, Murray Shanahan, and Matthew M. Nour. 2025. Technological folie \`a deux: Feedback Loops Between AI Chatbots and Mental Illness. *arXiv.org*. Retrieved July 29, 2026 from https://arxiv.org/abs/2507.19218v3

[20] Daniele Gatti. 2026. Three decades of studying false semantic memory. *Nat. Rev. Psychol.* 5, 6 (June 2026), 372–372. https://doi.org/10.1038/s44159-026-00563-0

[21] Kobi Hackenburg, Caroline Wagner, Luke Hewitt, Ben M. Tappin, Ed Saunders, Hannah Rose Kirk, Helen Margetts, and Christopher Summerfield. 2026. AI systems out-persuade expert humans. https://doi.org/10.48550/arXiv.2606.16475

[22] Ira E. Hyman Jr., Troy H. Husband, and F. James Billings. 1995. False memories of childhood experiences. *Appl. Cogn. Psychol.* 9, 3 (1995), 181–197. https://doi.org/10.1002/acp.2350090302

[23] Innocence Project. Eyewitness Misidentification. *Innocence Project*. Retrieved August 14, 2026 from https://innocenceproject.org/eyewitness-misidentification/

[24] Marcia K. Johnson, Shahin Hashtroudi, and D. Stephen Lindsay. 1993. Source monitoring. *Psychol. Bull.* 114, 1 (1993), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

[25] Sai Keerthana Karnam, Abhisek Dash, Krishna Gummadi, Animesh Mukherjee, Ingmar Weber, and Savvas Zannettou. 2026. Bowling with ChatGPT: On the Evolving User Interactions with Conversational AI Systems. In *Proceedings of the ACM Web Conference 2026* (*WWW '26*), April 12, 2026. Association for Computing Machinery, New York, NY, USA, 9711–9721. https://doi.org/10.1145/3774904.3792978

[26] Elizabeth Karpen. 2025. AI bots are filling users with conspiracy theories, repressed memories. *New York Post*. Retrieved August 13, 2026 from https://nypost.com/2025/05/05/tech/ai-bots-are-filling-users-with-conspiracy-theories-repressed-memories/

[27] Megan Kenny and Kevin Felstead. 2026. Making Memories: A Qualitative Exploration of the Impact of False Allegations of Satanic Ritual Abuse. *J. Forensic Psychol. Res. Pract.* 26, 2 (March 2026), 186–204. https://doi.org/10.1080/24732850.2024.2397000

[28] Rachel Leigh Greenspan and Elizabeth F. Loftus. 2021. Pandemics and infodemics: Research on the effects of misinformation on memory. *Human Behavior and Emerging Technologies* 3, 1 (2021), 8–12. https://doi.org/10.1002/hbe2.228

[29] Cara Laney, Erin K. Morris, Daniel M. Bernstein, Briana M. Wakefield, and Elizabeth F. Loftus. 2008. Asparagus, a Love Story. *Exp. Psychol.* 55, 5 (January 2008), 291–300. https://doi.org/10.1027/1618-3169.55.5.291

[30] Zehao Lin, Xixuan Hao, Renyu Fu, Shaobo Cui, Kai Chen, Chunyu Li, Zhiyu Li, and Feiyu Xiong. 2026. A Survey on Long-Term Memory Security in LLM Agents: Attacks, Defenses, and Governance Across the Memory Lifecycle. https://doi.org/10.48550/arXiv.2604.16548

[31] Elizabeth F. Loftus and John C. Palmer. 1974. Reconstruction of automobile destruction: An example of the interaction between language and memory. *J. Verbal Learn. Verbal Behav.* 13, 5 (October 1974), 585–589. https://doi.org/10.1016/S0022-5371(74)80011-3

[32] Matan Mazor. 2025. Inference About Absence as a Window Into the Mental Self-Model. *Open Mind* 9, (April 2025), 635–651. https://doi.org/10.1162/opmi_a_00206

[33] Ali A. Mazrui. 2013. Cultural Amnesia, Cultural Nostalgia and False Memory: Africa's Identity Crisis Revisited. *Afr. Asian Stud.* 12, 1–2 (January 2013), 13–29. https://doi.org/10.1163/15692108-12341249

[34] MetaKnowing. 2024. Have you noticed this? *r/singularity*. Retrieved August 13, 2026 from https://www.reddit.com/r/singularity/comments/1hdiejo/have_you_noticed_this/

[35] Microsoft Defender Security Research. 2026. Manipulating AI memory for profit: The rise of AI Recommendation Poisoning. *Microsoft Security Blog*. Retrieved August 12, 2026 from https://www.microsoft.com/en-us/security/blog/2026/02/10/ai-recommendation-poisoning/

[36] C. A. Morgan, Steven Southwick, George Steffian, Gary A. Hazlett, and Elizabeth F. Loftus. 2013. Misinformation can influence memory for recently experienced, highly stressful events. *Int. J. Law Psychiatry* 36, 1 (January 2013), 11–17. https://doi.org/10.1016/j.ijlp.2012.11.002

[37] Hamilton Morrin, Luke Nicholls, Michael Levin, Jenny Yiend, Udita Iyengar, Francesca DelGuidice, Sagnik Bhattacharya, Stefania Tognin, James MacCabe, Ricardo Twumasi, Ben Alderson-Day, and Thomas A. Pollak.

2026. Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. *Lancet Psychiatry* 13, 6 (June 2026), 522–530. https://doi.org/10.1016/S2215-0366(25)00396-7

[38] Gillian Murphy, Julie Maher, Lisa Ballantyne, Elizabeth Barrett, Conor S. Cowman, Caroline A. Dawson, Charlotte Huston, Katie M. Ryan, and Ciara M. Greene. 2023. How do participants feel about the ethics of rich false memory studies? *Memory* 31, 4 (April 2023), 474–481. https://doi.org/10.1080/09658211.2023.2170417

[39] Serge Nicolas. 2022. William Stern and the Establishment of a Psychology of Testimony in Germany. *Eur. Yearb. Hist. Psychol.* 8, 1 (January 2022), 11–76. https://doi.org/10.1484/J.EYHP.5.132221

[40] OpenAI. 2024. Memory and new controls for ChatGPT. *OpenAI*. Retrieved August 14, 2026 from https://openai.com/index/memory-and-new-controls-for-chatgpt/

[41] OpenAI. 2024. Memory and new controls for ChatGPT. Retrieved August 11, 2026 from https://openai.com/index/memory-and-new-controls-for-chatgpt/

[42] Henry Otgaar, Antonietta Curci, Ivan Mangiulli, Fabiana Battista, Elisa Rizzotti, and Giuseppe Sartori. 2022. A court ruled case on therapy-induced false memories. *J. Forensic Sci.* 67, 5 (2022), 2122–2129. https://doi.org/10.1111/1556-4029.15073

[43] Henry Otgaar, Mark L. Howe, and Lawrence Patihis. 2022. What science tells us about false and repressed memories. *Memory* 30, 1 (January 2022), 16–21. https://doi.org/10.1080/09658211.2020.1870699

[44] Pat Pataranutaporn, Chayapatr Archiwaranguprok, Samantha W. T. Chan, Elizabeth Loftus, and Pattie Maes. 2025. Slip Through the Chat: Subtle Injection of False Information in LLM Chatbot Conversations Increases False Memory Formation. In *Proceedings of the 30th International Conference on Intelligent User Interfaces* (*IUI '25*), March 24, 2025. Association for Computing Machinery, New York, NY, USA, 1297–1313. https://doi.org/10.1145/3708359.3712112

[45] Pat Pataranutaporn, Chayapatr Archiwaranguprok, Samantha W. T. Chan, Elizabeth Loftus, and Pattie Maes. 2025. Synthetic Human Memories: AI-Edited Images and Videos Can Implant False Memories and Distort Recollection. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems* (*CHI '25*), April 25, 2025. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3706598.3713697

[46] Lawrence Patihis, Steven J. Frenda, Aurora K. R. LePort, Nicole Petersen, Rebecca M. Nichols, Craig E. L. Stark, James L. McGaugh, and Elizabeth F. Loftus. 2013. False memories in highly superior autobiographical memory individuals. *Proc. Natl. Acad. Sci.* 110, 52 (December 2013), 20947–20952. https://doi.org/10.1073/pnas.1314373110

[47] Reddit Account Specialist_Gas_8984. 2024. ChatGPT Stared Into My Soul. *r/ChatGPT*. Retrieved August 12, 2026 from https://www.reddit.com/r/ChatGPT/comments/1hesmfo/chatgpt_stared_into_my_soul/

[48] Keishiro Sawa, Carla Z Cremer, Tobias Gerstenberg, Sanjay G Manohar, and Matan Mazor. 2026. "I would not have done that": Outcome knowledge distorts memory for decisions made minutes ago. Retrieved August 12, 2026 from https://osf.io/preprints/psyarxiv/4rpxm_v2/

[49] Jonas Schemmel, Lisa Datschewski-Verch, and Renate Volbert. 2024. Recovered memories in psychotherapy: a survey of practicing psychotherapists in Germany. *Memory* 32, 2 (February 2024), 176–196. https://doi.org/10.1080/09658211.2024.2305870

[50] Jonas Schemmel and Renate Volbert. 2025. Therapists' beliefs about traumatic memory: Possible effects on therapy proceedings and contributions to false memory formation. *Curr. Opin. Psychol.* 66, (December 2025), 102121. https://doi.org/10.1016/j.copsyc.2025.102121

[51] Elizabeth Seger, Shahar Avin, Gavin Pearson, Mark Briers, Seán Ó Heigeartaigh, and Helena Bacon. 2020. *Tackling threats to informed decision-making in democratic societies: promoting epistemic security in a technologically-advanced world*. The Alan Turing Institute, London.

[52] Julia Shaw. 2018. How Can Researchers Tell Whether Someone Has a False Memory? Coding Strategies in Autobiographical False-Memory Research: A Reply to Wade, Garry, and Pezdek (2018). *Psychol Sci* 29, 3 (March 2018), 477–480. https://doi.org/10.1177/0956797618759552

[53] Julia Shaw and Stephen Porter. 2015. Constructing Rich False Memories of Committing Crime. *Psychol. Sci.* 26, 3 (March 2015), 291–301. https://doi.org/10.1177/0956797614562862

[54] Julia Shaw and Annelies Vredeveldt. 2019. The Recovered Memory Debate Continues in Europe: Evidence From the United Kingdom, the Netherlands, France, and Germany. *Clin. Psychol. Sci.* 7, 1 (January 2019), 27–28. https://doi.org/10.1177/2167702618803649

[55] Mehrnoosh Sadat Shirvani, Jackie Crowley, Cher Peng, Jackie Liu, Thomas Chao, Suky Martinez, Laura Brandt, Ig-Jae Kim, and Dongwook Yoon. 2026. Cloning the Self for Mental Well-Being: A Framework for Designing Safe and Therapeutic Self-Clone Chatbots. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems* (*CHI '26*), April 13, 2026. Association for Computing Machinery, New York, NY, USA, 1–20. https://doi.org/10.1145/3772318.3790986

[56] Elizabeth C. Stade, Zoe M. Tait, Samuel T. Campione, Shannon Wiltsey Stirman, and Johannes C. Eichstaedt. 2026. Real-world use of large language models for mental health in 2024. *Npj Digit. Med.* (August 2026). https://doi.org/10.1038/s41746-026-02842-9

[57] Kimberley A. Wade, Maryanne Garry, J. Don Read, and D. Stephen Lindsay. 2002. A picture is worth a thousand lies: Using false photographs to create false childhood memories. *Psychon. Bull. Rev.* 9, 3 (September 2002), 597–603. https://doi.org/10.3758/BF03196318

[58] Mick Yang, Stephen Casper, Jonathan Stray, Jasmine Li, Cameron Jones, Anna Gausen, Natasha Jacques, Brian Christian, Bálint Gyevnár, Hannah Kirk, Zhonghao He, Dan Zhao, Siao Si Looi, Joshua Levy, Kobi Hackenburg, Elizabeth Seger, Matt Kowal, Michelle Malonza, Luke Hewitt, Hause Lin, Maarten Sap, Dylan Hadfield-Menell, Thomas Costello, Reihaneh Rabbany, Jean-François Godbout, David Rand, Atoosa Kasirzadeh, Gordon Pennycook, Yoshua Bengio, and Kellin Pelrine. 2026. AI Epistemic Risks: Emerging Mechanisms & Evidence. https://doi.org/10.2139/ssrn.6873005

[59] Yechao Zhang, Shiqian Zhao, Jiawen Zhang, Jie Zhang, Gelei Deng, Xiaogeng Liu, Chaowei Xiao, and Tianwei Zhang. 2026. When Claws Remember but Do Not Tell: Stealthy Memory Injection in Persistent Personal Agents. *arXiv.org*. Retrieved August 13, 2026 from https://arxiv.org/abs/2607.05189v1

[60] 2026. I asked an AI "What kind of person am I?", and it became a practice in viewing myself objectively｜せぶ. *note*(ノート). Retrieved August 12, 2026 from https://note.com/natty_walrus3755/n/n192c83162218

[61] Ntaganda | International Criminal Court. Retrieved August 14, 2026 from https://www.icc-cpi.int/drc/ntaganda