

MIT Open Access Articles

Synthetic Human Memories: AI-Edited Images and Videos Can Implant False Memories and Distort Recollection

The MIT Faculty has made this article openly available. **Please share** how this access benefits you. Your story matters.

Citation: AI is increasingly used to enhance images and videos, both intentionally and unintentionally. As AI editing tools become more integrated into smartphones, users can modify or animate photos into realistic videos. This study examines the impact of AI-altered visuals on false memories—recollections of events that didn't occur or deviate from reality. In a pre-registered study, 200 participants were divided into four conditions of 50 each. Participants viewed original images, completed a filler task, then saw stimuli corresponding to their assigned condition: unedited images, AI-edited images, AI-generated videos, or AI-generated videos of AI-edited images. AI-edited visuals significantly increased false recollections, with AI-generated videos of AI-edited images having the strongest effect (2.05x compared to control). Confidence in false memories was also highest for this condition (1.19x compared to control). We discuss potential applications in HCI, such as therapeutic memory reframing, and challenges in ethical, legal, political, and societal domains.

As Published: <https://doi.org/10.1145/3706598.3713697>

Publisher: ACM|CHI Conference on Human Factors in Computing Systems

Persistent URL: <https://hdl.handle.net/1721.1/162835>

Version: Final published version: final published article, as it appeared in a journal, conference proceedings, or other formally published context

Terms of use: Creative Commons Attribution



Synthetic Human Memories: AI-Edited Images and Videos Can Implant False Memories and Distort Recollection

Pat Pataranutaporn
MIT Media Lab, Massachusetts
Institute of Technology
Boston, Massachusetts, USA
patpat@media.mit.edu

Chayapatr Archiwaranguprok
University of the Thai Chamber of
Commerce
Bangkok, Thailand
chayapatr.arc@gmail.com

Samantha W. T. Chan
MIT Media Lab, Massachusetts
Institute of Technology
Cambridge, Massachusetts, USA
swtchan@media.mit.edu

Elizabeth Loftus
UCI School of Social Ecology, UC
Irvine
Irvine, California, USA
eloftus@law.uci.edu

Pattie Maes
MIT Media Lab, Massachusetts
Institute of Technology
Cambridge, Massachusetts, USA
pattie@media.mit.edu

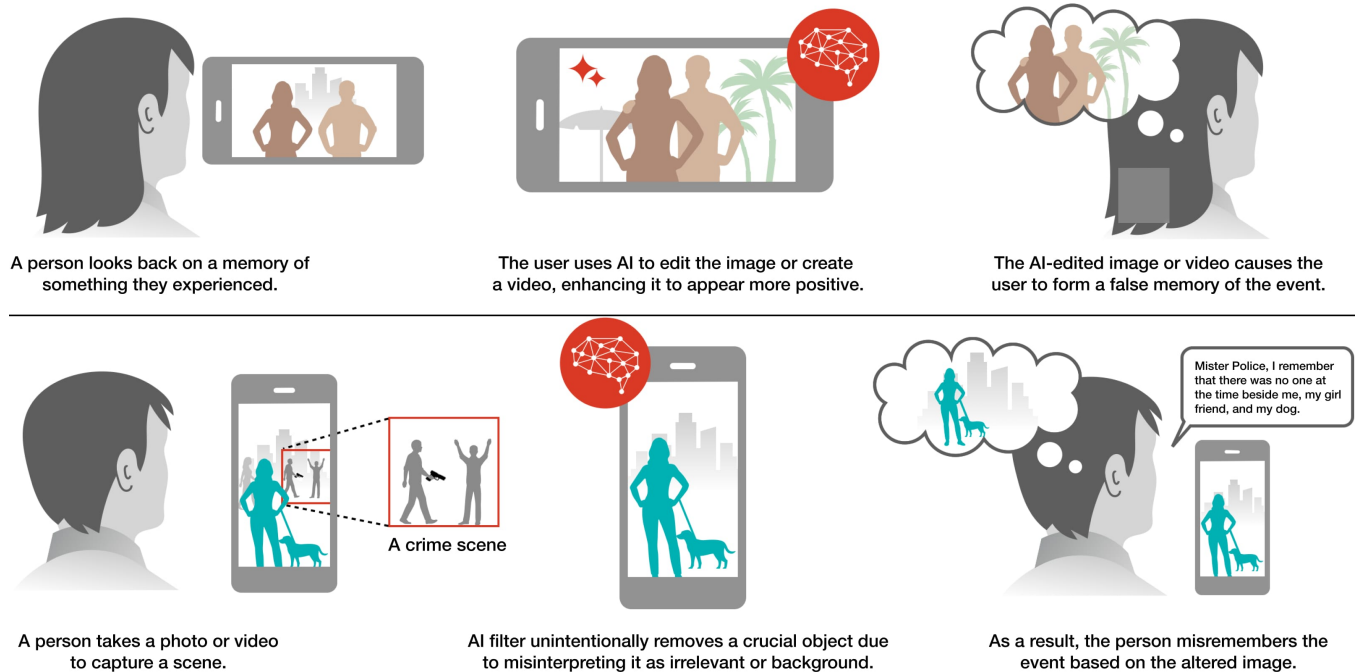


Figure 1: Illustration of how AI-edited media can create false memories. The top row depicts a person using AI to enhance an image or video to make it more positive. Over time, the person revisits the image without recalling that it was edited, leading to the development of a false memory of the event. The lower section depicts a situation where AI inadvertently modifies an image, eliminating bystanders from the frame as part of an automatic filter without retaining the original version (a feature already available in Google Photos and other camera apps). Later, when the individual reviews the photograph—potentially related to a crime scene—they develop a false recollection that matches the edited image rather than the actual event, leading to false witness testimony. This figure highlights the impact of AI-generated edits on human memories, demonstrating how subtle changes can distort recollection.



This work is licensed under a Creative Commons Attribution 4.0 International License.
CHI '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1394-1/25/04
<https://doi.org/10.1145/3706598.3713697>

Abstract

AI is increasingly used to enhance images and videos, both intentionally and unintentionally. As AI editing tools become more integrated into smartphones, users can modify or animate photos into realistic videos. This study examines the impact of AI-altered visuals on false memories—recollections of events that didn't occur

or deviate from reality. In a pre-registered study, 200 participants were divided into four conditions of 50 each. Participants viewed original images, completed a filler task, then saw stimuli corresponding to their assigned condition: unedited images, AI-edited images, AI-generated videos, or AI-generated videos of AI-edited images. AI-edited visuals significantly increased false recollections, with AI-generated videos of AI-edited images having the strongest effect (2.05x compared to control). Confidence in false memories was also highest for this condition (1.19x compared to control). We discuss potential applications in HCI, such as therapeutic memory reframing, and challenges in ethical, legal, political, and societal domains.

CCS Concepts

• **Human-centered computing** → **Interaction design theory, concepts and paradigms; Empirical studies in interaction design; Empirical studies in HCI; HCI theory, concepts and models.**

Keywords

Memory, AI-generated Media, Misinformation, Generative AI, Human-AI Interaction

ACM Reference Format:

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 *CHI Conference on Human Factors in Computing Systems (CHI '25)*, April 26-May 1, 2025, Yokohama, Japan. ACM, New York, NY, USA, 20 pages. <https://doi.org/10.1145/3706598.3713697>

1 Introduction

If a device existed that could help reframe your worst day in a more positive light, would you choose to use it? Memory-editing technologies have been a central theme in science fiction, prominently featured in works such as *Eternal Sunshine of the Spotless Mind*, *Men in Black*, *Total Recall*, and *Inception* [79]. However, techniques for altering human memories are not confined to the realm of fiction, as they represent a heavily studied area within psychology and cognitive science [74].

False memories, which refer to recollections of events that either never occurred or are significantly distorted from reality, have been a major focus in psychology research. The study of false memories is vital because they can distort witness testimonies, disrupt legal processes, and lead to faulty decision-making based on incorrect information. Given these broad implications, understanding how false memories form is a critical area of investigation [29, 58, 59, 62, 86, 107]. Unlike typical forms of misinformation [105], false memories are particularly insidious because the individual genuinely believes they recall accurate events, making them resistant to correction and potentially more influential in shaping beliefs and behaviors [56, 60]. Moreover, false memories can serve as a seed for making people more susceptible to additional false information [42, 92], creating a cascading effect that further distorts perceptions of reality and complicates efforts to establish accurate historical or personal narratives.

Research in cognitive psychology has demonstrated that human memories are remarkably malleable. Landmark studies by Loftus and colleagues [61, 62] revealed how both verbal questioning and visual stimuli can significantly influence and even create false memories of events that never occurred. For instance, experiments showed that subtle changes in wordings during interviews could alter participants' memory of witnessed events, while exposure to manipulated photographs could lead to the formation of entirely false childhood memories [96]. These findings have had profound implications for understanding human memory's susceptibility to external influences

However, these studies have predominantly been conducted in controlled laboratory settings, where images are manually edited by researchers and interviews are carefully planned. The process also involves human intervention in establishing trust, guiding participants, and presenting the manipulated images, which inherently limits the scope and scale of false memory induction. With recent advancements in artificial intelligence (AI), however, these limitations are beginning to change. The automation and accessibility of AI editing tools enable manipulation at unprecedented scale and sophistication, significantly expanding the possible impact of false memories on individuals. Moreover, this study explores new ground by examining how AI-generated videos derived from static images may further amplify memory distortion effects - an increasingly relevant concern as more types of AI tools become widely available.

This unprecedented proliferation of AI-driven image editing and video manipulation technologies has raised significant concerns regarding the integrity of consumed information. We argue that AI-generated content contributes to misinformation by distorting our understanding of the present (e.g., deepfakes) as well as reshaping how we remember the past. AI-generated media can potentially create false memories and lead individuals to recall past events differently than they actually occurred and were initially experienced. The implications of these technologies span both personal and societal domains, as illustrated in figure 2.

On a personal level, there has been a notable trend, particularly on social media platforms such as TikTok, of users employing AI to animate photographs of deceased family members, simulating interactions with departed loved ones. On a broader scale, the potential for AI-generated content to influence collective memory and historical narratives poses significant challenges to societal understanding and cohesion, potentially altering public perceptions of past events and shaping future decision-making processes. For example, AI-edited images of public gatherings or protests could subtly alter the perceived scale or mood of these events, gradually reshaping how participants and observers remember their personal experiences and consequently influencing the collective memory of significant social movements.

A crucial distinction must be made between deepfakes and AI-edits, as both leverage generative AI but differ significantly in their real-world implications and how people encounter them. Deepfakes typically involve the creation of entirely fabricated audio or video content, often for malicious purposes such as spreading disinformation. In contrast, AI-edits modify existing content, subtly altering genuine memories or experiences. This distinction is important, as people may be more vigilant against obviously fake

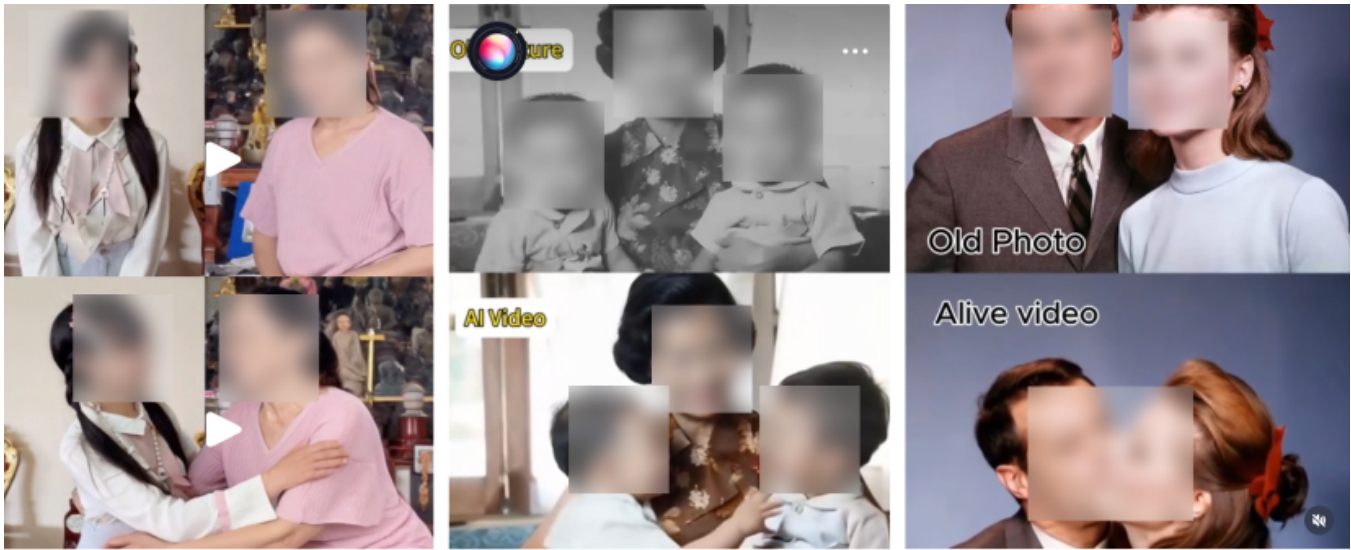


Figure 2: The figure illustrates how AI-generated content can potentially create false memories, particularly through AI-altered videos on social media platforms like TikTok. A recent trend on these platforms involves using AI to animate photos of deceased relatives, creating simulated interactions. These artificial experiences may blur the line between genuine memories and digitally fabricated ones, potentially affecting how people remember their loved ones.

content but less aware of slight modifications to their own experiences or memories. While initial research has primarily explored the effects of deepfakes on false memory formation in political contexts [54, 73, 91], this study specifically focuses on exploring the effects and implications of AI-edited content.

While AI-edited content poses risks for distorting memories, it also holds potential for positive applications for human memory. AI-assisted memory modification could help individuals process traumatic experiences more effectively, potentially reducing symptoms of conditions like post-traumatic stress disorder (PTSD) or depression. In addition, AI could assist in the creation of more comforting or positive recollections by emphasizing moments of joy or success. For example, AI might be used to enhance a photo from a difficult period by adjusting the atmosphere—making a gloomy day look sunny or adding vibrant details to a faded memory.

The use of AI in memory alteration raises significant ethical questions and requires careful consideration of potential long-term psychological effects. The balance between therapeutic benefit and the integrity of personal memories must be carefully weighed in any application of such technology.

Motivated by both the potential risks and promising applications of AI-edited content in memory formation, this study investigates the effects of AI-edited images and videos on the formation of false memories. Specifically, we aim to assess whether exposure to AI-edited visuals can influence individuals' recollections of past experiences. Particularly, this paper explores the following research questions:

- (1) To what extent do AI-edited images influence participants' memories of the original scenario, and how does this compare to participants in a control condition who view unedited images?

- (2) Does the conversion of images into AI-generated videos further exacerbate the formation of false memories, both in terms of the number of false memories reported and the confidence participants have in these memories, compared to static AI-edited images and the control condition?
- (3) How do different types of AI editing (e.g., changes to objects, people, or the environment) affect the severity of false memories?
- (4) What factors, such as familiarity with AI filters, memory efficacy, skepticism, age, gender, and education, moderate the severity and frequency of false memory formation when participants are exposed to AI-edited media?

In this pre-registered experiment involving 200 participants, subjects were initially shown a set of 24 images designed to serve as base memories. After completing a filler task, participants were divided into four groups: one group viewed the original, unedited images (control), another viewed AI-edited images, a third group viewed AI-generated videos of the unedited images, and the final group viewed AI-generated videos of the AI-edited images. The AI modifications included changes such as adding, removing, or altering objects, people, or environmental elements, thereby shifting the original context and meaning of the images (e.g., increasing military presence or removing visible effects of climate change). For video conditions, these images were transformed into dynamic videos using image-to-video AI, further altering their interpretation.

Our results demonstrate that AI-edited images and videos led to a significant increase in false recollections of the original scenario, with participants recalling incorrect details when exposed to the altered visuals, compared to the control condition. Notably, AI-generated videos of AI-edited images had the most profound effect on the number of reported false memories (2.05x compared to

control). The confidence in false memories was also higher for AI-generated videos of AI-edited images (1.19x compared to control). Our study included a weighted score analysis and group homogeneity checks to validate the experimental design.

Additional analyses exploring the effects across different types of image content (daily life, news, documentary) and elements edited by AI (people, objects, environments) consistently showed increased false memories with AI manipulation, with people-related edits having the highest absolute number of false memories (AI-generated videos of AI-edited image condition induced 45.3% false memories of all memories reported) and environmental-related edits gaining the most dramatic increase (2.6x compared to control). A mixed-effects regression model identified age as a small but significant factor in false memory formation, while other demographic and cognitive factors showed no significant relationship.

This study contributes new empirical evidence on how generative AI-based visual editing influences false memory formation, with a particular focus on the novel aspect of AI-generated videos. While previous research has explored how manually edited images can create false memories [62, 96], our work specifically examines the distinct challenges posed by AI-edited media and the transformation of static images into videos. Through our findings, we aim to inform ongoing discussions about the implications of increasingly accessible AI editing tools and their potential effects on memory and perception, as well as touch on the societal, political, and legal aspects of memory alteration through AI technology.

Our research indicates that “externalizing” human memories, a concept long explored in Human-Computer Interaction (HCI) [17], by storing them as digital files like photos or videos, may fundamentally alter how we naturally remember things—especially when AI is involved in modifying these externalized memories. This study highlights the critical responsibility of HCI researchers and practitioners in guiding the design and ethical implementation of AI technologies that could profoundly impact human cognitive functions. As AI increasingly influences how users interact with and interpret digital content, it is essential to understand how these systems can unintentionally distort memory and alter perception and recollection of reality. To summarize, this paper contributes to the discussion of false memories in the context of HCI and AI along the following dimensions:

- **Empirical Evidence on AI-implanted False Memories:** We report on one of the first evidence-based studies showing that AI-generated images and videos significantly increase false memory formation, with AI-generated videos of edited images having the greatest effect.
- **Impact of AI on Memory Recollection:** We demonstrate that exposure to AI-edited media not only increases false memories but also boosts participants’ confidence in these inaccurate recollections.
- **Effect of Different AI Edit Types:** We explore how various AI edits (e.g., changes to people, objects, and environments) influence the severity of false memories, finding that edits involving people have the most pronounced effect.
- **Applications, Implications, and Mitigation Strategies:** We examine both potential benefits (e.g., therapeutic memory

reframing) and risks (e.g., distorted testimony, misinformation) of AI-edited media, while proposing HCI strategies to mitigate these risks through ethical design interventions and public awareness campaigns.

2 Related Work

This study builds upon and extends a rich foundation of interdisciplinary research at the intersection of cognitive psychology, human-computer interaction, and AI. Our work synthesizes and advances knowledge from three primary domains: false memory research, the study of AI-generated misinformation, and HCI research.

2.1 Memory Augmentation and Vulnerability in HCI

HCI research has long explored ways to augment and support human memory through technology, tracing back to Engelbart’s pioneering vision of computers as extensions of human cognitive processes [17]. This work spans from early wearable “remembrance agents” [82] to modern life-logging technologies [13, 32], and extends to specialized applications like memory aids for older adults [11] and AR-based memory cues [2]. Recent advances in AI have enabled more sophisticated memory assistants, from conversational agents that infer memory needs in real-time [108] to AI-assisted journaling tools [46].

However, research has shown that externalizing memories through technology can leave users vulnerable to manipulation [16, 33, 83]. When memories are stored digitally, they become susceptible to both intentional and unintentional alterations. This vulnerability is particularly concerning as AI tools make image manipulation increasingly accessible and automated. The integration of AI into everyday memory augmentation technologies creates a double-edged sword: while offering powerful tools for memory enhancement, these systems may inadvertently make our memories more susceptible to distortion.

2.2 AI-generated Media and Misinformation

In recent years, the rapid advancement of AI technologies, particularly large language models [12] and generative visual models [103], has led to their widespread integration into work processes and daily life. This integration raises critical questions about the potential impact of AI on human cognition, particularly in the area of misinformation (the spread of falsehoods regardless of intent) and disinformation (deliberately misleading content or propaganda).

Researchers have identified an increase in AI-generated disinformation campaigns [28, 44] and the factors that make them disruptive to people’s ability to discern true and false information [6, 18, 27, 30, 31, 47, 75, 85]. These factors include authoritative tone [43], persuasive language [27, 43, 95], and targeted personalization [90]. AI-generated content has also been shown to influence people’s attitudes [38, 45, 95].

The concern about AI-generated misinformation is amplified by the known yet unresolved tendency of AI models to hallucinate or generate false information, either intentionally or unintentionally [15, 35, 102, 105]. Further, initial studies have provided evidence for the potential of AI systems to influence memory formation. In a separate study, a social robot that provided users with incorrect

information before a memory recognition test had an influence comparable to that of humans. The study found that even though the inaccurate information was emotionally neutral and not inherently memorable, 77% of the falsely provided words were incorporated into the participants' memories as errors [36].

The spread of misinformation on social media has become a major concern, prompting extensive HCI research. Studies have explored user interactions with misinformation [64, 66], the role of visual content [53, 98], and various interventions to combat it [34, 37]. AI's role in misinformation detection and mitigation has gained attention. Research has examined AI-based credibility indicators' effects on news perception [65] and the potential of using layperson judgments to combat misinformation [19].

Studies have revealed risks associated with AI-generated content [40] and explored AI-powered tools for fact-checking and misinformation detection [68]. Research has also examined the characteristics of AI-generated misinformation compared to human-created content [106].

Researchers are exploring novel approaches to leverage AI in combating misinformation, including the development of intelligent tools that encourage metacognitive skills "in the wild" [94]. The intersection of AI and misinformation presents both opportunities and challenges for the HCI community. Future research should focus on developing robust, transparent, and user-centered AI systems to support users in navigating the complex information landscape while addressing the implications and potential unintended consequences of AI-driven interventions.

2.3 False Memories Research

Research by Loftus and colleagues has established false memories as a crucial area of psychological research [55, 57, 58, 61, 62]. Their investigations into memory malleability and the misinformation effect have significantly influenced our understanding of memory processes, with far-reaching implications across psychology, law, and education [55, 57, 63].

A seminal study revealed the profound impact of linguistic framing on eyewitness memory, demonstrating that the choice of verbs in questioning could markedly influence participants' speed estimates of a car accident they had witnessed [61]. The "Lost in the Mall" experiment demonstrated the feasibility of implanting entirely fabricated childhood memories [62]. A recent replication, utilizing a larger sample size, corroborated and extended these findings, reporting a 35% false memory rate compared to the original study's 25% [72]. These results not only reinforce the robustness of the initial findings but also underscore the potential ramifications for eyewitness testimony in legal contexts.

In the context of visually induced false memories, visual stimuli can generate false memories of fictitious events [26, 84, 88, 97]. Methods include presenting scenes with omitted elements [71], personal photos [52], and narrative instructions [96]. One study found 50% of participants developed false memories after viewing fake childhood photos and guided imagery, highlighting implications for clinical and legal professionals [96].

Research on false memories has expanded to encompass various technological domains. The advent of immersive technologies has introduced new challenges in memory research, with studies

demonstrating the occurrence of source confusion between reality and VR experiences [4]. Researchers have developed frameworks categorizing XR Memory Manipulations (XRMMs) based on their impact on memory processes, emphasizing the ethical concerns and potential opportunities associated with manipulating perception and memory in XR environments [5]. Additionally, vulnerabilities in chatbot memory mechanisms that allow for the injection of misinformation alongside personal knowledge have been demonstrated [1].

3 Methodology

This study examines how AI-edited images and videos influence the formation of false memories. Specifically, we aim to assess whether exposure to AI-modified visuals affects individuals' recollections of the original scenario. In a pre-registered between-group experiment (AsPredicted #188511 - not yet public for anonymity), 200 participants were initially shown 24 baseline images. After a filler task, they viewed AI-altered versions of the images depending on their randomly assigned condition (50 participants per group). Examples include changing ethnicity, time of day, or military presence, as seen in Figure 4. In two conditions, the static images were converted into 5-second videos using a generative AI tool. Participants then answered 24 questions, one per image, to assess their memory of the originals. The following section details the study design, experimental conditions, and protocols.

3.1 Study Design Rationale

We particularly designed the study to simulate situations where individuals encounter AI-edited media in everyday contexts, particularly through social media and news platforms. In these scenarios, images are often anonymously altered or automatically edited by AI filters, frequently without the user's knowledge. The study also examines cases where others in the user's social circle, such as friends or family members, might edit personal images on behalf of the user without fully informing them of the details. In the "Future Research" section, we recommend exploring the impact of AI on false memories when people edit images themselves. We hypothesize that individuals may eventually forget they made these edits, potentially resulting in an effect similar to the scenarios described earlier. Our study's design incorporates key elements reflective of how people currently interact with digital content, offering insights into the potential risks associated with AI-manipulated media.

3.1.1 AI-Enhanced Media is Ubiquitous. The increasing integration of AI-powered image and video editing tools has spread to widely-used applications like Instagram, TikTok, and news websites, and increasingly to phone cameras themselves. For example, features like Google's "Best Take" [23] can remove unwanted elements or combine the best parts of multiple shots seamlessly. Apps such as Apple Photos, Google Photos, and Samsung's Galaxy AI offer built-in AI tools that allow users to easily edit, remove, or alter photo features with just a few simple steps. Users are frequently exposed to content that has been altered, many times even without their knowledge. Additionally, generative AI models, such as OpenAI's Sora, Luma's Dream Machine, and Kling, are increasingly used to animate static images into realistic-looking videos. Our study reflects this reality by introducing both static (AI-edited images) and

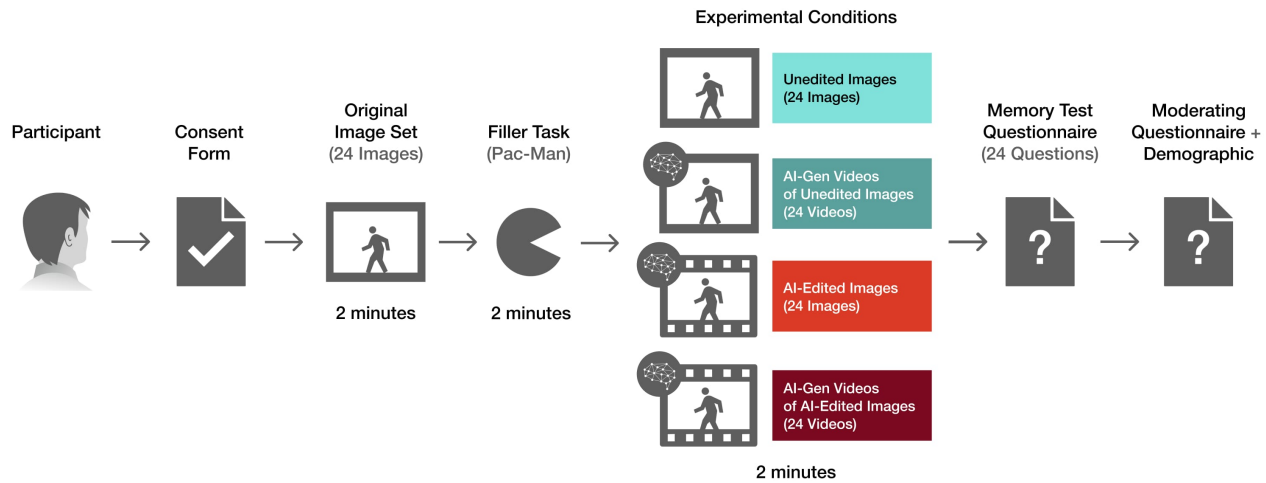


Figure 3: This figure illustrates the study procedure for our experiment examining how AI-generated images and videos can induce false memories. Participants first viewed original images to establish baseline memories, then were exposed to AI-modified versions after a filler task. These modifications included changes like increased military presence or removed climate change indicators. Finally, participants' memories of the original images were assessed through a series of questions, allowing researchers to measure the impact of AI-edited visuals on recall accuracy.

dynamic (AI-generated videos) content, highlighting the various forms in which individuals might encounter altered depictions of events in their daily lives. Additionally, we explore scenarios where AI-generated videos are created based on AI-edited images, representing a multi-layered approach to content generation.

3.1.2 Use of AI Labels in Media. As social media platforms start to implement labels that indicate whether content has been altered or generated by AI, research shows that notification systems can improve user awareness of content authenticity [78]. In our study, we included an "enhanced image" label across all conditions to serve as a baseline for participants' engagement with the visual materials.

3.1.3 Minimal Verification by Users. While in the real world, individuals have the ability to fact-check or search for original images to verify what they see, research indicates that most people do not take the time to do so [69]. Instead, they often trust the media they encounter, especially when it appears realistic and authoritative [22, 70]. Our study mirrors this behavior by presenting participants with edited media without requiring them to verify their authenticity.

3.1.4 Confidence in False Memories. One critical real-world issue is the cumulative effect of false memories. Once individuals have formed a false memory, such as believing a political figure said certain words, they may be more persuasive in influencing themselves and others that the memory is authentic [60, 62] and become more vulnerable to additional misinformation [42, 92]. Our study investigates not only whether false memories form from AI-enhanced media but also how confident individuals are in these false memories.

3.1.5 Risk of False Memories in Sensitive Scenarios. False memories formed through AI-altered content can have dangerous consequences. For instance, an AI-edited video could falsely depict a President committing a crime, leading individuals to believe they witnessed such an event. This kind of misinformation could fuel conspiracy theories or political unrest. While our study uses examples that are representative of what individuals might realistically encounter in daily life, we deliberately avoided using misleading content that could cause harm or incite dangerous reactions. We achieved this by carefully selecting our content: avoiding distortions that could cause reputational damage, using images of political figures from countries with minimal impact on US politics, and excluding any harmful ideologies or extremist content to minimize potential negative influences on participants.

3.2 Stimulus Sets and Experiment Conditions

Our study employed a diverse set of visual stimuli to investigate participants' perceptions and reactions to various types of media. The stimulus set comprised four distinct categories: unedited images, AI-edited images, AI-generated videos from unedited images, and AI-generated videos from AI-edited images:

- **Control (unedited images):** Contains 24 copyright-free images curated by researchers. The image set includes news images, such as politicians shaking hands with other leaders; personal pictures, such as views taken from a trip; and documentary pictures, such as a NASA astronaut portrait and polar bears. An example of the stimulus set is shown in Figure 4.
- **AI-edited images:** From the complete set of 24 images, we designated 12 images for AI editing using Adobe Photoshop AI, where details were either removed, added, or altered.

These edited images served as our primary dataset for analyzing the effects of AI manipulation on memory formation. The remaining 12 images were maintained in their original form across all conditions, serving as internal controls to verify that participants showed consistent response patterns when viewing identical images regardless of their experimental condition as discussed in Section 4.4. As shown in Figure 4, edits are categorized into 3 groups: Targeting people (changing the woman’s expression, changing the runner’s ethnicity, changing the gender of a person in the group), targeting the environment (removing the ice melt, changing the time of day, changing the background setting), and targeting objects (adding a military vehicle, removing military uniforms, and adding a stop sign).

- **AI-generated videos of unedited images:** Luma’s Dream Machine is used to convert the unedited image set into videos, primarily relying on the default auto-generation mode to create realistic motion based on the existing content in the images. For most images, no additional prompts were provided. However, for a few images where the system struggled, we added supplementary prompts to guide the model while maintaining the same intended outcome. The video length is 5 seconds. Example frames of the videos are shown in Figure 4.
- **AI-generated videos of AI-edited images:** Similar to AI-generated videos of unedited images, we used Luma’s Dream Machine to convert the AI-edited image set into videos. The video length is 5 seconds. Example frames of the videos are shown in Figure 4.

Participants were initially presented with the original, unedited image set, followed by one of the stimulus sets from the four conditions. The stimuli were delivered through a customized web interface with "next" and "previous" buttons, allowing participants to navigate through all 24 visuals. However, once they had completed viewing the original set, they could not go back and revisit it. The web interface was embedded within Qualtrics.

We employed a two-by-two design to compare static and dynamic AI-altered content, as well as unedited and AI-edited visuals to investigate their influence on the formation of false memories. Participants were randomly assigned to one of four conditions corresponding to four sets of stimuli.

3.3 Measurement

Participants were presented with a **Memories Test Questionnaire** consisting of 24 questions regarding their memory of the original, unedited images. Each question included a picture with the key detail of interest masked, while still providing enough context for participants to understand which image was being referenced. Example questions are shown in Figure 4.

These questions assessed whether participants recalled specific details from the images, such as objects, people, or environmental elements (e.g., “Did you remember seeing the bride smiling in the original picture?”). The participant could either answer agree, disagree, or unsure. In addition, participants rated their confidence on a 7-point scale from 1 (extremely lacking confidence) to 7 (extremely confident).

For the moderating factors, to measure **AI Filter Familiarity**, participants indicated their familiarity with using image or AI filter technologies on a 7-point scale (1 = Not familiar at all, 7 = Very familiar). **Frequency of Forgetting**, adapted from [104], assessed participants’ general memory performance, with responses ranging from 1 (Major problems) to 7 (No problems). **Memory Efficacy** was measured using a subset of items from [3], where participants rated their ability to remember visual and verbal information, such as recalling names and objects, on a 7-point scale (1 = Strongly disagree, 7 = Strongly agree). Finally, **Skepticism** was assessed using a scale adapted from [21], where participants rated their agreement with statements reflecting distrust in official information (e.g., “The official media provides false information”) on a 7-point scale (1 = Strongly disagree, 7 = Strongly agree).

3.4 Experiment Protocol

The experiment was conducted via Qualtrics with 200 participants (1:1 female:male ratio) recruited from CloudResearch. All participants were U.S. residents, aged 20–73 years ($M=38$, $s.d.=12.25$). After signing a consent form disclosing possible deception, participants were informed they would view images and their "filtered" versions, followed by feedback questions.

Following an initial attention check, participants viewed 24 unedited images for two minutes in a swipeable Instagram-like interface, completed a two-minute Pac-Man filler task, and then viewed a second set of images labeled "AI-enhanced image" based on their randomly assigned experimental condition. The 2-minute viewing time for 24 images was selected to allow participants to process each image while preventing detailed memorization that would be unnatural in real-world scenarios. The 2-minute filler task (Pac-Man game) served to clear working memory without being so long as to induce significant natural forgetting.

A second attention check was administered. Participants then completed 24 memory test questions about the original images, with each question accompanied by a masked version of the image to aid recall without revealing edited features. The questions, balanced between positive (added elements) and negative (removed elements) prompts, assessed whether AI-edited information had been incorporated into participants’ memories (e.g., "Did you remember seeing military presence in the picture?"). The study concluded with demographic questions and a post-survey measuring potential moderating factors including AI filter familiarity, memory efficacy, tendency for forgetting, skepticism, and education level.

3.5 Analysis

In order to evaluate the impact of AI-edited media on participants’ memory, we conducted 3 analyses, each designed to address different aspects of our hypothesis. The analyses were structured as follows:

- (1) **Primary analysis** aiming to test the main hypothesis regarding the overall effect of each condition on memory. Three key metrics were observed: (1) The **Number** of reported false, uncertain, and non-false memories (i.e., how many times participants recalled incorrectly, were unsure, or recalled the original image correctly). (2) The **Confidence** levels associated with each type of memory (false, uncertain, and

Unedited Images	AI-Gen Videos of Unedited Images	AI-Edited Images	AI-Gen Videos of AI-Edited Images	Masked Images	Memory Test Questionnaire	Target	Modification
					Did you remember seeing the bride smiling in the original picture?	People	Changing the woman's expression
					Did you remember seeing two black guys running in the picture?	People	Changing the runner's ethnicity
					Did you remember seeing a woman in this groupshot?	People	Changing the gender of a person in the group
					Did you remember seeing that all the land was covered by snow? (no ocean water)	Environment	Removing the ice melt
					Did you remember seeing the sun in the original picture?	Environment	Changing the time of day
					Did you remember seeing a city in the picture?	Environment	Changing the background setting
					Did you remember seeing military vehicle in the picture?	Object	Adding a military vehicle
					Did you remember seeing military presence in the picture?	Object	Removing military uniforms
					Did you remember seeing a sign on the street?	Object	Adding a stop sign

Figure 4: The stimulus set consisted of four distinct categories: unedited images, AI-edited images, AI-generated videos from unedited images, and AI-generated videos from AI-edited images. The edits were further divided into three subgroups based on the type of change: People, Objects, and Environment. In the questionnaire, masked versions of the images were used to facilitate recall without revealing the edited features.

non-false). (3) A **Weighted score**, calculated by assigning values to each memory type (false: -1, uncertain: 0, true: 1) and multiplying them by the corresponding confidence levels, then summing these products.

- (2) **Subgroup analysis** examining the number of false memory occurrences across conditions in the subgroups separated by two conditions, i.e., **image content** and **subject of edit**, to uncover any category-specific effects of generative content on memory.
- (3) **Moderating Factors** exploring the effect of moderating factors on the number of false memory occurrences using a mixed-effect regression model.

We classified participants' memories as false, uncertain, or non-false based on their responses (agree, disagree, or unsure) to specific questions about details of the original images. If a participant answered a question incorrectly, it was marked as a false memory. A

correct answer was classified as a non-false memory, and if the participant responded with "unsure," it was categorized as uncertain.

For each test, we first assessed if the normality assumption was met for each outcome variable distribution using the Shapiro-Wilk test. If the normality assumption was not met, we performed a Kruskal-Wallis test followed by a post hoc Dunn test using the Bonferroni error correction. If the normality assumption was met, we then conducted a homogeneity test using a Levene test to assess whether the samples were from populations with equal variances. If the samples were not homogeneous, we ran a Welch analysis of variance (ANOVA) and a Tukey's honestly significant difference test (Tukey's HSD). If the samples were homogeneous, we ran a basic ANOVA test with a Tukey post hoc test.

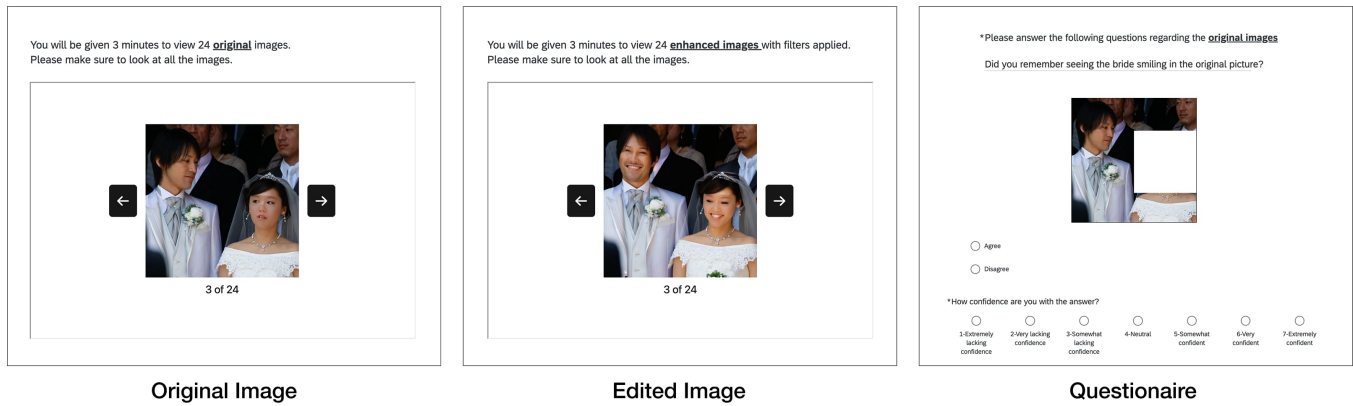


Figure 5: Survey interface components. From left to right: (1) Original image viewing instructions, (2) AI-enhanced image viewing instructions, (3) Questionnaire about original image details and confidence assessment. The sample images depict a wedding scene. The questionnaire prompts participants to recall specific details and rate their confidence in their memory.

3.6 Approvals

This research was reviewed and approved by the [Anonymized] Institutional Review Board, protocol number [Anonymized]. The study was also preregistered at AsPredicted (#188511 - not yet public for anonymity).

4 Result of Primary Analysis

The results show that AI-edited images distort participants' memories of the original images, leading them to report more false memories and higher levels of confidence in the false memories while converting the images into videos further increases the number of reported false memories. The statistics are illustrated in Figure 6.

4.1 Number of Recalled Memories by Categories

Shapiro-Wilk tests revealed non-normality across all four experimental conditions (P-value ranging from $3.238e-5$ to $5.79e-4$). We employed one-way Kruskal-Wallis tests to analyze variation.

4.1.1 Number of False Memories. As illustrated in the first column of Figure 6, the test indicates significant differences in the number of reported false memories between the conditions, $H=34.157$, $P=1.836e-07$, $P<0.05$. The AI-edited images induced significantly more false memories than the unedited images (1.67x), while AI-generated videos of AI-edited images amplified the number of false memories, leading to a 2.05x number of reported false memories. However, AI-generated videos from unedited images result in 1.25x compared to control. Statistics: control, $M=18.878$, $s.d.=14.164$, number of reported false memories out of 12 (#)=2.265; AI-gen videos of unedited images: $M=23.667$, $s.d.=18.055$, #=2.840; AI-edited images: $M=31.373$, $s.d.=15.965$, #=3.765; AI-gen videos of AI-edited images: $M=38.667$, $s.d.=18.838$, #=4.640. Post hoc Dunn test with Benjamini-Hochberg (FDR) correction: control vs. AI-edited images, $P=2.5e-4$; control vs. AI-gen videos of AI-edited images, $P=7.654e-08$; AI-gen videos of unedited images vs. AI-gen videos of AI-edited images, $P=6.153e-05$.

4.1.2 Number of Uncertain and Non-false memories. There were no significant differences in the number of reported uncertain memories between conditions as shown in the second column of Figure 6, $H=0.903$, $P=0.825$, $P>0.05$. Meanwhile, in the third column, the test indicates significant differences in the number of reported non-false memories between the conditions with $H=30.448$, $P=1.111e-06$, $P<.0071$. All three edited sets introduce a lower number of reported non-false memories, i.e., 0.88x, 0.82x, 0.67x, in AI-gen videos of unedited images, AI-edited images, and AI-gen videos of AI-edited image conditions, respectively, compared to control. Statistics: control, $M=66.156$, $s.d.=18.929$, number of reported non-false memories out of 12 (#)=7.939; AI-gen videos of unedited images: $M=58.167$, $s.d.=19.543$, #=6.980; AI-edited images: $M=53.922$, $s.d.=20.032$, #=6.471; AI-gen videos of AI-edited images: $M=43.833$, $s.d.=18.545$, #=5.260. Post hoc Dunn test with Benjamini-Hochberg (FDR) correction: control vs. AI-edited images, $P=0.005$; control vs. AI-gen videos of AI-edited images, $P=5.594e-8$; AI-generated videos of unedited images vs. AI-gen videos of AI-edited images, $P=5.086e-4$; AI-edited images vs. AI-gen videos of AI-edited images, $P=0.0078$.

4.2 Confidence in Recalled Memories

Shapiro-Wilk tests revealed non-normality across all four experimental conditions (P-value ranging from $6.709e-12$ to $1.167e-10$). We employed one-way Kruskal-Wallis tests to analyze variation. The results are illustrated in Figure 7

4.2.1 Confidence of False Memories. As shown in the first column of Figure 7, the results indicate significant differences between the conditions, $H=8.581$, $P=0.0354$, $P>0.05$. AI-gen videos of AI-edited images and AI-edited images, respectively, induce a 1.19x and 1.1x increase in the level of confidence in false memories vs. control. Statistics: control, $M=4.536$, $s.d.=2.041$; AI-gen videos of unedited images: $M=4.383$, $s.d.=2.220$; AI-edited images, $M=5.027$, $s.d.=1.092$; AI-gen videos of AI-edited images: $M=5.412$, $s.d.=1.449$. Post hoc Dunn test with Benjamini-Hochberg (FDR) correction: control vs. AI-gen videos of AI-edited images, $P=0.0139$; AI-gen videos of

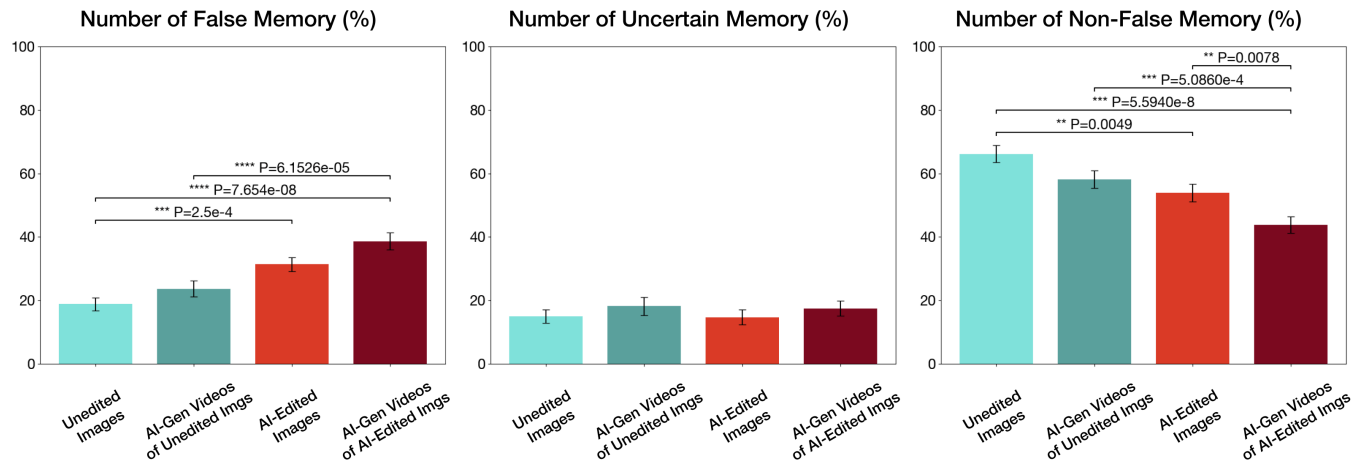


Figure 6: The percentage of reported false, uncertain, and non-false memories (i.e., how many times participants recalled incorrectly, were unsure, or recalled the original image correctly) were analyzed using a one-way Kruskal-Wallis and post hoc Dunn with FDR. P-value annotation legend: **, $P < 0.01$; *, $P < 0.001$; ****, $P < 0.0001$.**

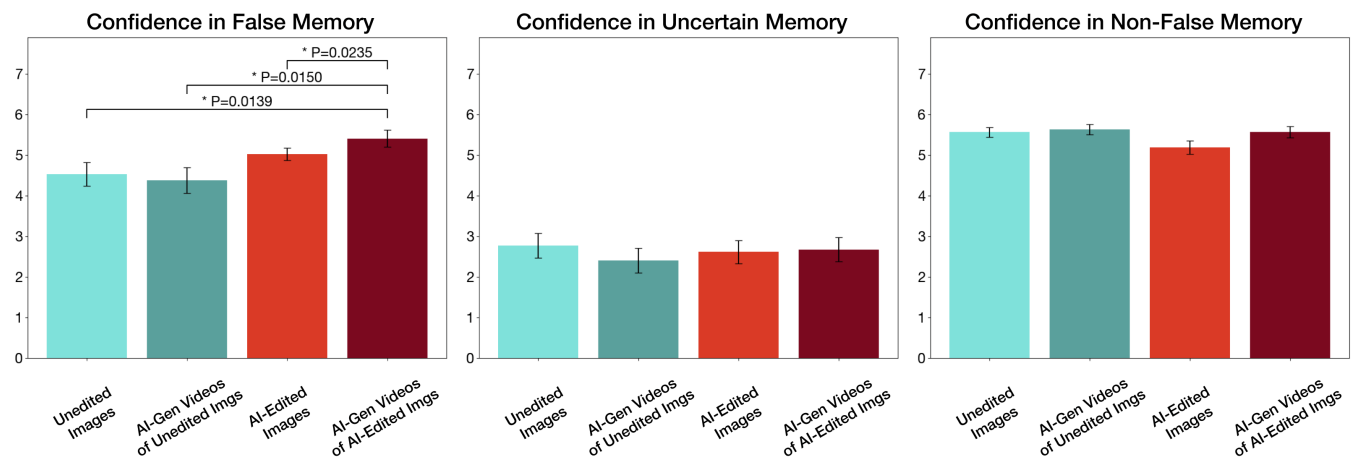


Figure 7: The confidence of recalled memories in three categories (false/uncertain/non-false) were analyzed using a one-way Kruskal-Wallis and post hoc Dunn with FDR. P-value annotation legend: *, $P < 0.05$.

unedited images vs. AI-gen videos of AI-edited images, $P=0.0150$; AI-edited Images vs. AI-gen videos of AI-edited images, $P=0.0235$.

4.2.2 Confidence in Uncertain and Non-False Memories. The tests indicate no significant differences in the confidence in uncertain and non-false memories, the numbers are illustrated in the second and third column of Figure 7. Statistics: Uncertain Memories, $H=0.726$, $P=0.867$, $P > 0.05$; Non-false Memories, $H=5.012$, $P=0.170$, $P > 0.05$.

4.3 Weighted Score

To combine multiple metrics into a single score, we calculated the weighted score focusing on participants' memory recollection performance by multiplying the score assigned to each memory

type ($False = -1$, $Uncertain = 0$ ¹, $Non-False = 1$) by the participant's confidence level for that memory (ranging from 1 to 7). The Shapiro-Wilk test results for all groups show $P > 0.05$, indicating that the data in each group is normally distributed. The Levene test result ($P=0.683$) suggests homogeneity of variances across the groups. The one-way ANOVA test yielded a highly significant result ($F=14.577$, $P=1.312e-08$, $P < 0.001$), indicating substantial differences among the group means. Statistics: control, $M=32.061$, $s.d.=21.989$; AI-gen of unedited images, $M=24.760$, $s.d.=22.801$; AI-edited images: $M=15.803$, $s.d.=21.476$; AI-gen videos of AI-edited images, $M=3.040$, $s.d.=24.933$.

¹While this scoring approach prioritizes definitive memory performance assessment, we acknowledge that uncertain responses may carry valuable information about implicit memory processes such as partial recollection or intuitive recognition, which warrants dedicated investigation in future research.

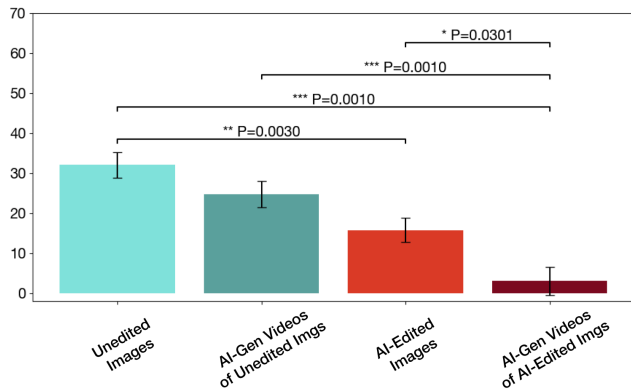


Figure 8: The weighted score of recalled memories based on number and confidence was analyzed using a one-way ANOVA and post hoc Tukey. P-value annotation legend: *, $P < 0.05$; **, $P < 0.01$; *, $P < 0.001$.**

The subsequent Tukey’s post hoc test revealed significant differences between several pairs of groups, particularly between the control and both AI-edited images ($P=0.003$) and AI-gen videos of AI-edited images groups ($P=0.001$), as well as between AI-gen videos of unedited images and AI-gen videos of AI-edited images groups ($P=0.001$), and between AI-edited images and AI-gen videos of AI-edited images ($P=0.03$). These findings suggest that the AI editing interventions, especially those involving static AI image edits and combination with AI video generation, have a significant impact on the measured variable compared to the control condition.

4.4 Group Homogeneity Analysis

Our experimental design incorporated internal controls by maintaining a subset of identical images across conditions. Specifically, in the AI-edited images group, 12 of the 24 images were identical to those in the control group. Similarly, in the AI-generated videos of AI-edited images group, these same 12 images were presented as AI-generated videos, matching those in the AI-generated videos of unedited images group. This design element allowed us to examine participant responses to identical stimuli across different experimental conditions.

Following the same procedure as our main analysis, we conducted statistical tests. The Shapiro-Wilk test results for normality were as follows: control ($W=0.934$, $P < 0.001$), AI-gen videos of unedited images ($W=0.935$, $P < 0.001$), AI-edited images ($W=0.937$, $P < 0.001$), and AI-gen videos of AI-edited images ($W=0.942$, $P < 0.001$). The Kruskal-Wallis tests examining responses across conditions yielded P-values ranging from 0.425 to 0.927. The analysis revealed consistent response patterns across experimental conditions, with no detected differences in how participants processed this subset of identical stimuli, supporting the homogeneity of our test groups. The results of the Kruskal-Wallis test are shown in Table 1.

Condition	H Statistic	P-Value
Number of False Memories	1.359	0.715
Number of Uncertain Memories	1.961	0.581
Number of Non-False Memories	0.502	0.918
Confidence in False Memories	1.073	0.784
Confidence in Uncertain Memories	2.791	0.425
Confidence in Non-False Memories	0.460	0.928

Table 1: Group Homogeneity Checks for Primary Analysis Using Kruskal-Wallis Tests

5 Results of Additional Analysis

5.1 Number of false memories based on different types of image contents

The stimulus set is categorized into three distinct subgroups, representing different domains of visual media consumption. Daily life photos capture contemporary informal situations and objects from personal perspectives, reflecting familiar scenes people encounter regularly. News images are defined as photographs taken primarily for current event reporting by professional photojournalists in recent years, focusing on timely coverage of events and public figures. Documentary and archival materials encompass photographs taken with anthropological or scientific intent or images that serve as historical records.² Image categorization was performed independently by two researchers, with any disagreements resolved through discussion or removal of ambiguous images. Each of these subgroups in our test set contains four images, resulting in a total of 12 images across the three categories. We apply the main analysis procedure to each of these subgroups based on the number of reported false memories, aiming to evaluate the consistency of AI effects across these diverse content types. As illustrated in Figure 9, the results demonstrate a consistent pattern in the effects of AI-generated content. Statistics: Daily Life, $H=12.884$, $P=0.00489$, $P < 0.005$; News, $H=27.712$, $P=4.174e-06$, $P < 0.05$; Documentary and Archive, $H=23.617$, $P=3.003e-05$, $P < 0.001$.

In all three groups, there is a clear trend of increasing false memories from unedited images to AI-generated videos of AI-edited images. This similarity in trends suggests that the impact of AI-generated and AI-edited content on memory recall is robust across different types of visual stimuli, whether they depict everyday scenes, current events, or historical content, indicating that the observed effects of AI manipulation on memory are not limited to a specific type of visual content but appear to be a more general phenomenon.

5.2 Number of false memories based on type of subject edited by AI

As part of an exploratory analysis with descriptive statistics, we categorized the edited content into three subgroups based on the subject of the edit: people (alterations to facial features, changes in race or ethnicity, and adjustments to body shape or posture),

²While some images may share characteristics across categories (e.g., a NASA photograph could be considered both news and documentary), examining these distinct contexts arguably remains valuable as they represent different modes of visual media engagement and information consumption.

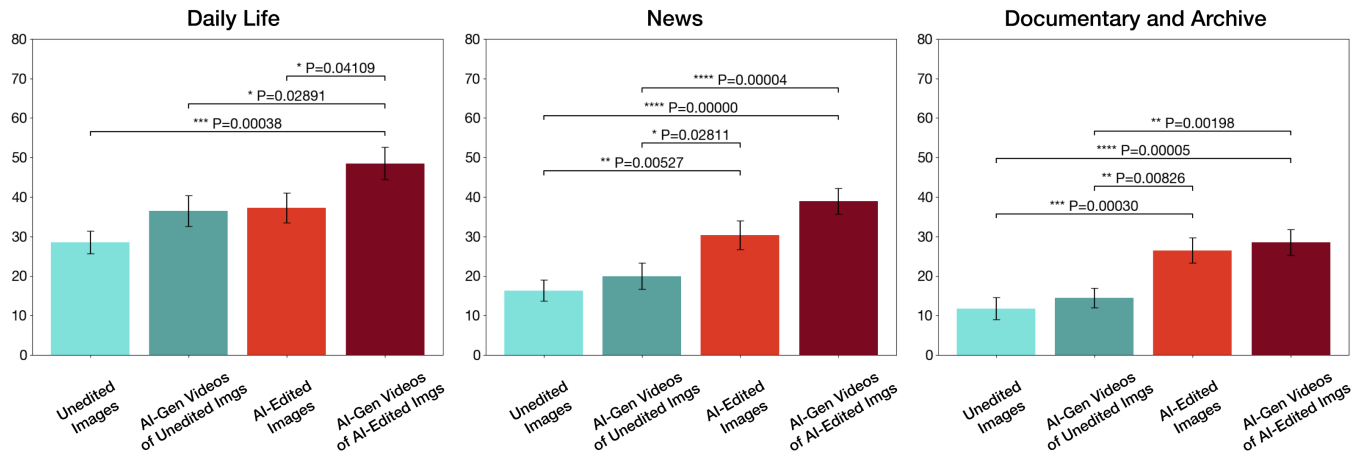


Figure 9: The number of reported false memories in three subgroups were analyzed using a one-way Kruskal-Wallis and post hoc Dunn with FDR. P-value annotation legend: *, $P < 0.05$; **, $P < 0.01$; *, $P < 0.001$; ****, $P < 0.0001$.**

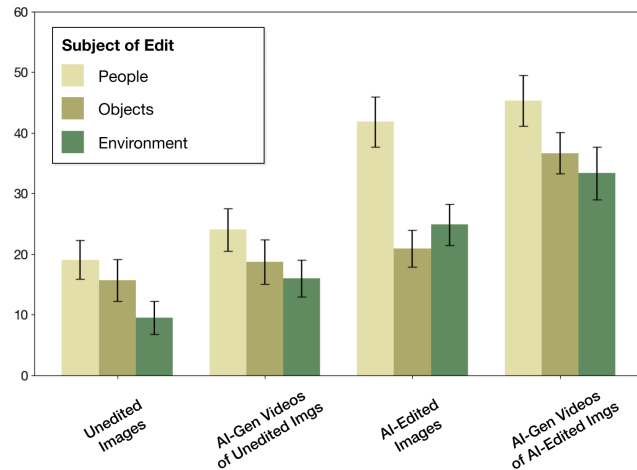


Figure 10: The percentages of reported false memories by subjects of edit.

objects (the addition or removal of items from scenes, or changes to existing objects within the image), and environments (modifications to backgrounds, lighting, weather conditions, or architectural elements). The number of false memory reports across these different categories of AI-generated edits showed a consistent pattern of increasing false memories. Yet, the impact of AI manipulation varies across the different subjects of edit. While people-related edits consistently produced the highest absolute number of false memories, environmental edits showed the most dramatic increase in false memory reported due to image edits, especially in the most complex condition of AI-generated videos of AI-edited images (2.6x of control in AI-gen videos of AI-edited images condition, compared to 2.4x and 2.3x in people and object edits respectively).

Statistics: **People:** control, $M=19.048$, $s.d.=22.335$; AI-gen videos of unedited images, $M=24.000$, $s.d.=24.980$; AI-edited images, $M=41.830$, $s.d.=29.405$; AI-gen videos of AI-edited images, $M=45.333$, $s.d.=29.635$.

Objects: control, $M=15.646$, $s.d.=24.376$; AI-gen videos of unedited images, $M=18.667$, $s.d.=25.957$; AI-edited images, $M=20.915$, $s.d.=21.854$; AI-gen videos of AI-edited images, $M=36.667$, $s.d.=24.267$. **Environment:** control, $M=9.524$, $s.d.=19.048$; AI-gen videos of unedited images, $M=16.000$, $s.d.=21.333$; AI-edited images, $M=24.837$, $s.d.=24.560$; AI-gen videos of AI-edited images, $M=33.333$, $s.d.=30.551$.

5.3 Moderating Factors

We employed a mixed effects regression approach to investigate how variables such as gender, age, education, familiarity with AI-filter technology, and cognitive factors (frequency of forgetting and memory efficiency) relate to the number of false memories. Table 2 provides the result of the regression model. The result suggests that age has a significant negative relationship ($P=0.01$) with the number of false memories, yet the effect is small ($Coef.=-0.031$). Meanwhile, other factors do not show significant relationships in the model.

6 Discussion

6.1 AI-Edited Content Boosts False Memories with Alarming Confidence

Our findings provide evidence suggesting the effect of AI-edited and AI-generated media on human memory distortion. Participants exposed to AI-altered images exhibited a markedly higher propensity to report false memories compared to those who viewed unedited control images. This effect was even more pronounced when participants were presented with AI-generated videos based on AI-edited images, suggesting that dynamic AI-edited media significantly amplified the distortion of memory, effectively embedding false details deeper into participants' recollections.

Perhaps the most disconcerting aspect of the study's results is the high degree of confidence participants reported in their inaccurate recollections. The AI-edited images not only led to the formation of false memories but also instilled a misplaced sense of certainty in these fabricated recollections. This effect is maximized with AI-generated videos of edited images, which caused the

	Coef.	Std.Err.	z	P> z	[0.025	0.975]
Familiarity with AI-Filter Technology	0.040	0.087	0.454	0.650	-0.131	0.211
Age	-0.031	0.012	-2.583	0.010*	-0.055	-0.008
Gender (Male compared to Female)	-0.040	0.291	-0.138	0.891	-0.610	0.530
Gender (Other compared to Female)	0.021	2.063	0.010	0.992	-4.022	4.065
Education Level	0.103	0.119	0.862	0.389	-0.131	0.336
Cognitive Load	-0.133	0.119	-1.121	0.262	-0.366	0.100
Memory Efficiency	-0.069	0.058	-1.194	0.232	-0.182	0.044
Skepticism toward Media and Institutions	-0.021	0.015	-1.376	0.169	-0.051	0.009

Table 2: The mixed effects regression model result. P-value annotation legend: *, $P < 0.05$

most significant increase in both false memory formation and the associated confidence levels.

The implications of these findings are far-reaching and potentially alarming. The combination of false memories and high confidence levels creates a particularly dangerous scenario, as individuals are more likely to believe and act upon incorrect information they perceive as true. This phenomenon could have consequences in various contexts, from eyewitness testimonies in legal proceedings to the spread of misinformation in social and political spheres. Moreover, the study raises important questions about the nature of memory itself and how easily it can be manipulated by advanced AI technology.

6.2 The Impact of Different Types of Edits: People, Environment, and Objects

This study explored the impact of different types of AI-generated edits—specifically changes to people, environments, and objects, on the formation of false memories. The pattern in the results suggests that while people-related details are generally harder to recall, possibly because of their complexity, they are less susceptible to additional distortion from AI manipulation. Environmental manipulations, such as altering the time of day or modifying landscape elements, proved particularly potent in distorting recall when subjected to AI editing and video generation. Object-related edits fell between these two extremes. The data underscores the varying susceptibility of different image elements to AI-induced false memories, with contextual/environmental changes showing the most striking relative increase. These findings highlight the complex nature of AI manipulation techniques and human memory across different aspects of visual information, suggesting that while human subjects may be naturally more challenging to remember accurately, environmental details, despite their lower base rate of false memories, show greater relative vulnerability to AI-induced distortions. These findings on both the prevalence of false memory recall and the susceptibility of memories to manipulation have crucial implications for real-world applications, particularly in eyewitness testimony. This underscores the urgent need for further research to understand and mitigate the negative effects of false memories and memory manipulation, as further discussed in Section 6.5.1.

6.3 Moderating Factors

Interestingly, the mixed-effects regression model shows that age has a statistically significant negative effect on AI-induced false

memories, indicating that younger individuals are more susceptible to such influences. This finding warrants further investigation into the underlying mechanisms. One possible explanation could relate to the higher exposure of younger individuals to technology, including AI-filtered media. The mere exposure effect suggests that people tend to develop a preference for things they are more familiar with [10], which could lead to increased trust or decreased skepticism toward AI-generated content among younger users. However, with older adults increasingly adopting technology and becoming more proficient in its use [20, 41], we may also observe a parallel reduction in skepticism among older users in the near future. Another explanation could involve the different modes of interaction with technology across age groups. Younger individuals may allocate their attention differently when consuming information, potentially making them more susceptible to false memory formation. Nonetheless, it is crucial to emphasize that while the age-related difference is statistically significant, the effect size is relatively small, which suggests caution against overstating its practical implications.

Familiarity with AI-filter technology, while positively correlated, does not have a statistically significant moderating effect, suggesting that mere exposure or understanding of the technology may not provide protection against false memory implantation. This aligns with the earlier explanation of the inverse relationship between familiarity and skepticism, wherein increased trust may lead to greater susceptibility to false memories. Other variables, such as education level, cognitive load, and memory efficiency, do not show significant moderating effects, meaning these individual differences do not strongly influence the likelihood of developing false memories in response to AI-edited media. Interestingly, skepticism toward media and institutions, though negatively correlated, also lacks statistical significance, suggesting that even individuals with a more critical view of media may still be vulnerable to AI-induced memory distortions. This suggests that false memory formation operates beyond a simple “trust vs. mistrust” framework [76, 87]: anyone, regardless of their background or expertise, could potentially fall victim to false memories when exposed to altered media. In practical terms, this means that AI-altered media presents a challenge to all segments of society. Educational campaigns may need to focus not just on technological literacy but also on promoting awareness of how our cognitive systems process and reconstruct information, making us all prone to forming false memories in certain contexts.

6.4 The Impact of Label

All AI-edited visuals in our study were presented with a label indicating AI alteration, similar to notification systems being implemented on social media platforms. Despite these labels, we observed significant false memory formation in the AI-edited conditions. While our study was not designed to directly compare the effectiveness of different labeling approaches, this observation aligns with prior research showing that passive notifications alone may be insufficient to mitigate cognitive biases from manipulated media [19].

The persistence of false memories despite labeling suggests that such notifications, while informative, may not sufficiently alter the cognitive processing of visual material. When individuals encounter plausible altered content, they may integrate it into their existing memory frameworks regardless of accompanying labels. This points to a potential need to shift from purely informative labels to designs that actively provoke critical engagement with content. Future research could explore this direction by investigating how different label designs might impact false memory formation, particularly focusing on approaches that encourage active reflection on content authenticity.

6.5 Negative Implication of AI-implanted False Memories

Our study has shown that AI has the capacity to manipulate human memory by altering images or videos, creating false recollections that may not align with reality. In Human-Computer Interaction, the implications of these AI-implanted false memories are significant, posing both challenges and opportunities, as shown in Figure 2. This section explores the negative implications of AI-implanted false memories in HCI and discusses potential mitigation strategies, while the next section explores the beneficial use cases for such phenomena.

6.5.1 Wrongful Legal Accusations. In the legal domain, AI-generated false memories pose significant risks for justice systems, particularly as altered media can rapidly spread through social platforms and influence witness recollections. The case of Steve Titus, who was wrongfully convicted of rape in 1980 based on a witness's increased confidence in their initially uncertain identification [100], demonstrates how memory distortion can lead to devastating legal consequences. Today, with AI-manipulated content capable of going viral instantly, the potential for distorting public perception and witness memories is even more extreme. Social media platforms can amplify manipulated content through memes, posts, and viral sharing, often simplifying or sensationalizing events in ways that can shape public understanding and potentially influence legal proceedings.

The spread of AI-edited media presents unique challenges in courtroom settings, where witness testimony often plays a pivotal role in case outcomes. Witnesses exposed to manipulated content may unknowingly incorporate false details into their testimonies, with their confidence potentially inflated by repeated exposure to misleading narratives online. Even subtle alterations to images or videos, such as placing individuals in misleading contexts or

manipulating crime scene details, can rapidly circulate online and affect both public perception and witness recollections.

6.5.2 Public Misinformation Spread. AI-edited false memories pose significant risks for manipulating public perception of historical and societal events at scale. For instance, AI could alter footage of a peaceful protest by introducing elements of conflict, potentially causing viewers to falsely remember the event as violent—even if they previously saw the original footage or were present.

The 2024 presidential election campaign illustrates these concerns, where crowd sizes at political events have drawn attention [24, 93] alongside worries about AI image manipulation [7, 89]. Event attendees who later encounter subtly modified visuals might experience shifts in their original memories, potentially affecting their perceptions of candidate popularity or policy support. These distorted recollections could influence voter preferences based on manipulated depictions rather than reality.

The ramifications of such misinformation are profound and far-reaching. It could lead to widespread public confusion, distorting not only individual memories but also our collective understanding of historical events. This manipulation of shared memories has the potential to polarize societal beliefs and exacerbate existing social divisions, especially when disseminated widely on social media platforms and other digital channels. Moreover, the proliferation of AI-edited media is eroding trust in visual content, leading to increased skepticism toward all images and videos, including authentic ones.

6.5.3 Reinforcement of Biases and Stereotypes. AI-generated false memories can reinforce existing biases and stereotypes by subtly altering images or videos to fit preconceived notions. For example, AI editing could inadvertently modify media to emphasize negative stereotypes about specific groups, which can lead to false memories that perpetuate harmful narratives. This is particularly dangerous in the context of racial, gender, or cultural bias, where AI-manipulated media could fuel prejudice and discrimination. The reinforcement of biases through AI-generated false memories is a critical concern that intersects with issues of social justice and equality. AI systems, if not carefully designed and monitored, can inadvertently perpetuate and amplify societal biases. For instance, an AI system trained on biased data might systematically alter images to conform to stereotypical representations of certain groups, leading to a feedback loop that reinforces these biases in human memory and perception.

6.5.4 Mitigating AI-implanted False Memories. Current strategies for addressing AI-implanted false memories—including detection tools, regulatory frameworks, and notification systems—remain vital but underutilized. These foundational measures need to be widely implemented, particularly in high-stakes contexts like political campaigns and legal proceedings. However, these approaches alone may be insufficient. Rather than relying on automated detection and passive alerts, we need to develop mechanisms that actively encourage users' critical engagement with content [78]. This challenge requires combining protective measures with HCI solutions that encourage critical analysis. Solutions should address both individual and societal needs, promoting digital diversity [8, 51, 80] while helping users evaluate content authenticity. Future work must unite psychology, HCI, and AI experts to understand how AI

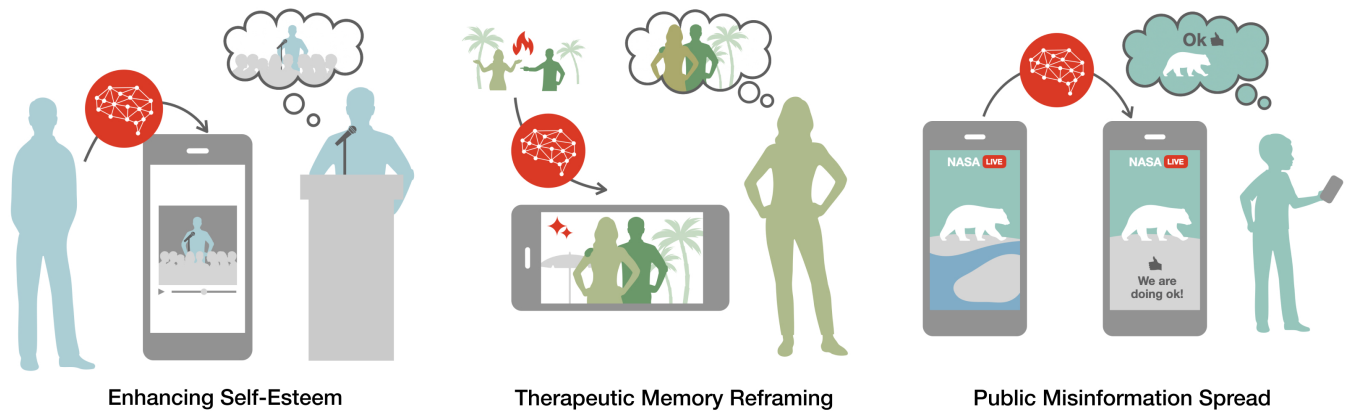


Figure 11: Examples of the potential implications of AI-generated media. The first scenario, “Enhancing Self-Esteem,” shows how AI can be used to improve personal memories, boosting confidence by enhancing positive recollections. The second scenario, “Therapeutic Memory Reframing,” illustrates the use of AI to alter memories of past events, reframing distressing moments in a more positive light for therapeutic purposes. The third scenario, “Public Misinformation Spread,” demonstrates how AI can manipulate information in public media. For instance, a student may have seen an accurate NASA climate change video in a classroom, but later, their memory could be updated with an AI-generated false video from social media showing no evidence of climate change, altering their perception of the issue.

affects perception and develop interfaces supporting active analysis of increasingly complex interactions.

6.6 Positive Use Cases of AI-Generated False Memories

While AI-generated false memories pose significant risks, exploring their potential benefits under controlled, ethical conditions is equally important, particularly in mental health and personal development contexts.

6.6.1 Therapeutic Memory Reframing. AI might assist in altering distressing memories for patients undergoing psychological treatment, providing a powerful tool for mental health professionals [39, 48, 67, 101]. Specifically, AI could enhance therapeutic memory reframing techniques [39, 48, 67, 101]. As shown in Figure 2, AI could assist in modifying distressing elements in photos or videos associated with traumatic memories, helping patients build less emotionally charged associations under professional supervision. This technique might be particularly beneficial for patients suffering from Post-Traumatic Stress Disorder (PTSD), phobias, or other anxiety disorders.

6.6.2 Enhancing Self-Esteem. AI-edited images could also support self-esteem enhancement and personal development. Building on existing research in AR-enhanced confidence for public speaking [50] and creativity [49], AI could subtly enhance images of past achievements or create motivational visualizations, as demonstrated in Figure 2. For example, when an individual recalls this enhanced version of their past public speaking, it could potentially boost their confidence in these future scenarios. This application aligns with visualization techniques used in sports psychology and personal development [14, 81, 99].

While these applications show promise, their implementation demands rigorous ethical consideration and oversight. Any use of AI-generated false memories, must prioritize transparency and informed consent. The goal should remain focused on supporting healing and growth rather than memory manipulation without compromising an individual’s agency or authentic experiences.

6.7 Limitations and Future Research

Our study provides insights into AI-edited media’s impact on memory formation, though several limitations warrant promising directions for future research.

6.7.1 Long-term exposure and Memory Persistence. One primary limitation of this study is its reliance on relatively short-term exposure to AI-altered media. Participants were exposed to the manipulated images and videos for a limited time, which may not fully reflect real-world scenarios where individuals are repeatedly exposed to such content over extended periods. Moreover, as demonstrated in Loftus’ research, false memory implantation often involves gradual and repeated suggestion rather than a single exposure. However, our study reveals that even brief exposure to these types of stimuli induces a significant false memory effect. Previous studies [9, 25, 62, 77] suggest this effect would likely intensify with increased exposure duration or frequency. In addition to memory formation, studying how memories persist over time and resist correction is vital for a more comprehensive exploration of the phenomena.

6.7.2 Expanded Demographic and Sample Sets. Another limitation is the constrained nature of our sample set, both in demographics and dataset size. Demographically, the study was confined to U.S. participants aged 18-100, despite efforts to ensure diversity within these parameters. In terms of stimuli, our study employed only 24 pre-selected images rather than participants’ personal photos,

potentially limiting both the representativeness of AI-generated content and ecological validity, as memories of personal experiences are typically richer and more complex. Future research should address these constraints through a dual approach: expanding to a more globally diverse sample with a larger, more varied image set, while also conducting parallel studies using participants' personal photos, leveraging AI tools' ability to efficiently process personal images at scale.

The study focused primarily on visual media (images and videos), but AI is capable of generating and manipulating other forms of content, including audio and text. Future research should investigate the impact of AI-edited auditory cues (such as fabricated sound environments) and textual content (like AI-edited personal narratives) on false memory formation. This could provide a more holistic understanding of how different modalities of AI-generated content affect human cognition. A follow-up study should also test what difference it makes whether it is the person themselves or some other entity that does the editing. This exploration of agency in AI-assisted editing could reveal important insights into the mechanisms of false memory formation and the role of perceived authorship in memory implantation. Additionally, our study did not fully explore memory failures related to forgetting, particularly in cases where AI editing removed elements from images rather than adding new ones. Future work should examine this distinct type of memory distortion.

Further, our experiment was conducted in a controlled setting, which may not fully capture the complexity of real-world scenarios where individuals encounter AI-generated content. Future studies could employ more naturalistic designs, such as field experiments or ecological momentary assessments, to examine how AI-altered media affects memory in everyday life. This could include studying the impact of AI-generated content on social media platforms or in news consumption.

6.7.3 Ethical Considerations. The ethical implications of AI-generated false memories warrant further investigation. Future research should explore the long-term psychological and social consequences of living in an environment where memories can be easily manipulated by AI. This includes studying informed consent and transparency in AI-altered content exposure, developing robust safeguards against intentional misuse in political and personal contexts, and examining privacy implications for both training data and personalized false memory generation. Additional focus areas should include the psychological impact in therapeutic settings and ensuring equitable access to protective measures.

7 Conclusion

This research demonstrates the significant impact of AI-edited media on human memory distortion. The results reveal that exposure to AI-altered images substantially increases the likelihood of false memory formation, with participants exposed to such content exhibiting a markedly higher propensity to report inaccurate recollections compared to those who viewed unedited control images. This effect was even more pronounced when participants were presented with AI-generated videos based on AI-edited images, suggesting that dynamic AI media significantly amplifies memory distortion. Perhaps most concerning, participants reported high

levels of confidence in their false memories, particularly in the AI-generated video conditions.

These findings have far-reaching implications across various domains, including legal proceedings, political discourse, and misinformation. However, the study also highlights potential positive applications of this technology, particularly in therapeutic contexts. AI-generated content could be used to reframe traumatic memories or enhance self-esteem when applied ethically and under professional supervision. These findings underscore the need for a balanced approach to AI development that maximizes benefits while mitigating risks.

Moving forward, this research calls for increased awareness and stricter regulations regarding the use of AI in media creation and dissemination. Future research should focus on developing more effective interventions to mitigate the risk of AI-induced false memories, including improved content labeling systems and public education campaigns. Additionally, interdisciplinary collaboration between AI researchers, cognitive scientists, ethicists, and policymakers will be crucial in addressing the complex challenges posed by AI's influence on human memory and perception. Ultimately, this study serves as a compelling foundation for better understanding and navigating the intricate relationships between AI and human cognition, both now and in the future.

References

- [1] Conor Atkins, Benjamin Zi Hao Zhao, Hassan Jameel Asghar, Ian Wood, and Mohamed Ali Kaafar. 2023. *Those Aren't Your Memories, They're Somebody Else's: Seeding Misinformation in Chat Bot Memories*. Springer Nature Switzerland, 284–308. doi:10.1007/978-3-031-33488-7_11
- [2] Kilian L Bahnsen, Lucas Tiemann, Lucas Plabst, and Tobias Grundgeiger. 2024. Augmented Reality Cues Facilitate Task Resumption after Interruptions in Computer-Based and Physical Tasks. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24, Vol. 2)*. ACM, 1–16. doi:10.1145/3613904.3642666
- [3] Jane M Berry, Robin L West, and Dierdre M Dennehey. 1989. Reliability and validity of the Memory Self-Efficacy Questionnaire. *Developmental psychology* 25, 5 (1989), 701.
- [4] Elise Bonnal, Julian Frommel, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. 2024. Was it Real or Virtual? Confirming the Occurrence and Explaining Causes of Memory Source Confusion between Reality and Virtual Reality. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24, Vol. 4)*. ACM, 1–17. doi:10.1145/3613904.3641992
- [5] Elise Bonnal, Wen-Jie Tseng, Mark McGill, Eric Lecolinet, Samuel Huron, and Jan Gugenheimer. 2023. Memory Manipulations in Extended Reality. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23, Vol. 4)*. ACM, 1–20. doi:10.1145/3544548.3580988
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [7] Kellen Browning. 2024. Despite Trump's Claims, Footage Shows Large Crowd at Harris's Detroit Rally — nytimes.com. <https://www.nytimes.com/interactive/2024/08/12/us/elections/trump-harris-detroit-rally.html>. [Accessed 12-09-2024].
- [8] Elie Bursztein, Karla J Brown, Leonie M Sanderson, and Patrick Gage Kelley. 2024. Leveraging Virtual Reality to Enhance Diversity, Equity and Inclusion training at Google. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–7.
- [9] 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. *arXiv preprint arXiv:2408.04681* (2024).
- [10] Sam WT Chan, Tamil Selvan Gunasekaran, Yun Suen Pai, Haimo Zhang, and Suranga Nanayakkara. 2021. KinVoices: Using voices of friends and family in voice interfaces. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–25.
- [11] Samantha W. T. Chan, Thisum Buddhika, Haimo Zhang, and Suranga Nanayakkara. 2019. ProspecFit: In Situ Evaluation of Digital Prospective

- Memory Training for Older Adults. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 3 (Sept. 2019), 1â€“20. doi:10.1145/3351235
- [12] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology* 15, 3 (2024), 1–45.
- [13] Yi Chen and Gareth J. F. Jones. 2010. Augmenting human memory using personal lifelogs. In *Proceedings of the 1st Augmented Human International Conference (AH '10)*. ACM. doi:10.1145/1785455.1785479
- [14] Jennifer Cumming and Sarah E. Williams. 2013. Introducing the revised applied model of deliberate imagery use for sport, dance, exercise, and rehabilitation. *Movement & Sport Sciences - Science & Motricité* 82 (2013), 69–81. doi:10.1051/sm/2013098
- [15] Valdemar Danry, Pat Pataranutaporn, Ziv Epstein, Matthew Groh, and Pattie Maes. 2022. Deceptive AI systems that give explanations are just as convincing as honest AI systems in human-machine decision making. *arXiv preprint arXiv:2210.08960* (2022).
- [16] Nigel Davies, Adrian Friday, Sarah Clinch, Corina Sas, Marc Langheinrich, Geoff Ward, and Albrecht Schmidt. 2015. Security and Privacy Implications of Pervasive Memory Augmentation. *IEEE Pervasive Computing* 14, 1 (Jan. 2015), 44–53. doi:10.1109/mprv.2015.13
- [17] Douglas C Engelbart. 2021. Augmenting human intellect: a conceptual framework (1962). (2021).
- [18] Ziv Epstein, Aaron Hertzmann, Investigators of Human Creativity, Memo Akten, Hany Farid, Jessica Fjeld, Morgan R Frank, Matthew Groh, Laura Herman, Neil Leach, et al. 2023. Art and the science of generative AI. *Science* 380, 6650 (2023), 1110–1111.
- [19] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the Crowd Game the Algorithm? Using Layperson Judgments to Combat Misinformation on Social Media by Downranking Distrusted Sources. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–11. doi:10.1145/3313831.3376232
- [20] M Faverio. 2022. Share of those 65 and older who are tech users has grown in the past decade. <https://www.pewresearch.org/short-reads/2022/01/13/share-of-those-65-and-older-who-are-tech-users-has-grown-in-the-past-decade/>
- [21] Rodrigo Ferrer-Urbina, Yasna Ramirez, Patricio Mena-Chamorro, Marcos Carmona-Halty, and Gerald Sepúlveda-Páez. 2024. Naive skepticism scale: development and validation tests applied to the Chilean population. *Psicologia: Reflexão e Crítica* 37 (2024), 6.
- [22] B. J. Fogg. 2003. Prominence-interpretation theory: explaining how people assess credibility online. In *CHI '03 extended abstracts on Human factors in computing systems - CHI '03 (CHI '03)*. ACM Press, 722. doi:10.1145/765891.765951
- [23] Geoffrey A. Fowler. 2008. How Google's Best Take uses AI to edit your photos and fix smiles - The Washington Post. <https://www.washingtonpost.com/technology/2023/10/11/ai-face-google-best-take/>. Accessed: 2024-09-10.
- [24] Kierra Frazier. 2024. Harris tweaks Trump over crowd size ahead of debate-politico. <https://www.politico.com/news/2024/09/10/harris-trump-crowd-size-debate-00178116/>. [Accessed 12-09-2024].
- [25] Peter Frost. 2000. The quality of false memory over time: Is memory for misinformation “remembered” or “known”? *Psychonomic Bulletin & Review* 7, 3 (Sept. 2000), 531–536. doi:10.3758/bf03214367
- [26] Maryanne Garry and Matthew P. Gerrie. 2005. When photographs create false memories. *Current Directions in Psychological Science* 14, 6 (2005), 321–325.
- [27] Josh A Goldstein, Jason Chao, Shelby Grossman, Alex Stamos, and Michael Tomz. 2024. How persuasive is AI-generated propaganda? *PNAS nexus* 3, 2 (2024), pgae034.
- [28] Josh A Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv preprint arXiv:2301.04246* (2023).
- [29] Brian Gonsalves and Ken A Paller. 2000. Neural events that underlie remembering something that never happened. *Nature Neuroscience* 3, 12 (2000), 1316–1321.
- [30] Matthew Groh, Ziv Epstein, Chaz Firestone, and Rosalind Picard. 2022. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences* 119, 1 (2022), e2110013119.
- [31] Matthew Groh, Aruna Sankaranarayanan, Nikhil Singh, Dong Young Kim, Andrew Lippman, and Rosalind Picard. 2023. Human detection of political speech deepfakes across transcripts, audio, and video. *arXiv preprint arXiv:2202.12883* (2023).
- [32] Cathal Gurrin, Alan F. Smeaton, and Aiden R. Doherty. 2014. LifeLogging: Personal Big Data. *Foundations and Trends® in Information Retrieval* 8, 1 (2014), 1–125. doi:10.1561/15000000033
- [33] Linda A. Henkel and Anna Milliken. 2020. The benefits and costs of editing and reviewing photos of one’s experiences on subsequent memory. *Journal of Applied Research in Memory and Cognition* 9, 4 (Dec. 2020), 480–494. doi:10.1037/h0101858
- [34] Hendrik Heuer and Elena Leah Glassman. 2022. A Comparative Evaluation of Interventions Against Misinformation: Augmenting the WHO Checklist. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI '22). Association for Computing Machinery, New York, NY, USA, Article 241, 21 pages. doi:10.1145/3491102.3517717
- [35] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232* (2023).
- [36] Tsung-Ren Huang, Yu-Lan Cheng, and Suparna Rajaram. 2023. Unavoidable social contagion of false memory from robots to humans. *American Psychologist* (2023).
- [37] Farnaz Jahanbakhsh, Amy X. Zhang, and David R. Karger. 2022. Leveraging Structured Trusted-Peer Assessments to Combat Misinformation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 524 (nov 2022), 40 pages. doi:10.1145/3555637
- [38] Maurice Jakesch, Advait Bhat, Daniel Buschek, Lior Zalmanson, and Mor Naaman. 2022. Interacting with Opinionated Language Models Changes Users’ Views. *Arxiv Open Access* (2022).
- [39] Meng Jiang, Qing Zhao, Jianqiang Li, Fan Wang, Tianyu He, Xinyan Cheng, Bing Xiang Yang, Grace WK Ho, and Guanghui Fu. 2024. A Generic Review of Integrating Artificial Intelligence in Cognitive Behavioral Therapy. *arXiv preprint arXiv:2407.19422* (2024).
- [40] Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. 2024. Is Stack Overflow Obsolete? An Empirical Study of the Characteristics of ChatGPT Answers to Stack Overflow Questions. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 935, 17 pages. doi:10.1145/3613904.3642596
- [41] Brittne Kakulla. 2022. Tech Trends and Adults 50-Plus. Washington, DC: AARP Research, December 2021. <https://doi.org/10.26419/res.00493.001>
- [42] Irene P. Kan, Kendra L. Pizzonia, Anna B. Drummey, and Eli J. V. Mikkelsen. 2021. Exploring factors that mitigate the continued influence of misinformation. *Cognitive Research: Principles and Implications* 6, 1 (Nov. 2021). doi:10.1186/s41235-021-00335-9
- [43] Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey Hancock. 2023. Working with AI to persuade: Examining a large language model’s ability to generate pro-vaccination messages. *Stanford Preprint* (2023).
- [44] Katarina Kertysova. 2018. Artificial intelligence and disinformation: How AI changes the way disinformation is produced, disseminated, and can be countered. *Security and Human Rights* 29, 1-4 (2018), 55–81.
- [45] Celeste Kidd and Abeba Birhane. 2023. How AI can distort human beliefs. *Science* 380, 6651 (2023), 1222–1223.
- [46] Taewan Kim, Donghoon Shin, Young-Ho Kim, and Hwajung Hong. 2024. DiaryMate: Understanding User Perceptions and Experience in Human-AI Collaboration for Personal Journaling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. ACM, 1–15. doi:10.1145/3613904.3642693
- [47] Himabindu Lakkaraju and Osbert Bastani. 2020. “How do I fool you?” Manipulating User Trust via Misleading Black Box Explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.
- [48] Jinyoung Lee. 2019. The service design platform for people with dementia: Person-centred reminiscence therapy with Artificial Intelligence in immersive environments. *Temes de Disseny* 35 (2019), 154–169.
- [49] Joanne Leong, Pat Pataranutaporn, Yaoli Mao, Florian Perteneder, Ehsan Hoque, Janet M Baker, and Pattie Maes. 2021. Exploring the use of real-time camera filters on embodiment and creativity. In *extended abstracts of the 2021 CHI conference on Human Factors in Computing Systems*. 1–7.
- [50] Joanne Leong, Florian Perteneder, Muhender Raj Rajjee, and Pattie Maes. 2023. “Picture the Audience...”: Exploring Private AR Face Filters for Online Public Speaking. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–13.
- [51] Sky Leslie, Mirjam Vosmeer, Casper Sterrenburg, Anastasia Maimenscu, Damir Catibovic, and Olico Matsjitadze. 2022. VR for Diversity: A Virtual Museum Exhibition about LGBTQI+. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 203, 4 pages. doi:10.1145/3491101.3519917
- [52] D Stephen Lindsay, Lisa Hagen, J Don Read, Kimberley A Wade, and Maryanne Garry. 2004. True photographs and false memories. *Psychological Science* 15, 3 (2004), 149–154.
- [53] Maxim Lismic, Alexander Lex, and Marina Kogan. 2024. “Yeah, this graph doesn’t show that”: Analysis of Online Engagement with Misleading Data Visualizations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 199, 14 pages. doi:10.1145/3613904.3642448
- [54] Nadine Liv and Dov Greenbaum. 2020. Deep Fakes and Memory Malleability: False Memories in the Service of Fake News. *AJOB Neuroscience* 11, 2 (March 2020), 96–104. doi:10.1080/21507740.2020.1740351

- [55] Elizabeth F Loftus. 1981. Eyewitness testimony: Psychological research and legal thought. *Crime and justice* 3 (1981), 105–151.
- [56] Elizabeth F. Loftus. 1992. When A Lie Becomes Memory's Truth: Memory Distortion After Exposure to Misinformation. *Current Directions in Psychological Science* 1, 4 (Aug. 1992), 121–123. doi:10.1111/1467-8721.ep10769035
- [57] Elizabeth F Loftus. 1996. *Eyewitness testimony*. Harvard University Press.
- [58] Elizabeth F Loftus. 1997. Creating false memories. *Scientific American* 277, 3 (1997), 70–75.
- [59] Elizabeth F Loftus. 2003. Make-believe memories. *American Psychologist* 58, 11 (2003), 867.
- [60] Elizabeth F. Loftus. 2005. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory: Figure 1. *Learning & Memory* 12, 4 (July 2005), 361–366. doi:10.1101/lm.94705
- [61] Elizabeth F Loftus and John C Palmer. 1974. Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of verbal learning and verbal behavior* 13, 5 (1974), 585–589.
- [62] Elizabeth F Loftus and Jacqueline E Pickrell. 1995. The formation of false memories. 720–725 pages.
- [63] Elizabeth F Loftus and Guido Zanni. 1975. Eyewitness testimony: The influence of the wording of a question. *Bulletin of the Psychonomic Society* 5, 1 (1975), 86–88.
- [64] Zhicong Lu, Yue Jiang, Cheng Lu, Mor Naaman, and Daniel Wigdor. 2020. The Government's Dividend: Complex Perceptions of Social Media Misinformation in China. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3313831.3376612
- [65] Zhuoran Lu, Patrick Li, Weilong Wang, and Ming Yin. 2022. The Effects of AI-based Credibility Indicators on the Detection and Spread of Misinformation under Social Influence. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 461 (nov 2022), 27 pages. doi:10.1145/3555562
- [66] Lisa Mekioussa Malki, Dilisha Patel, and Aneesa Singh. 2024. "The Headline Was So Wild That I Had To Check": An Exploration of Women's Encounters With Health Misinformation on Social Media. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 128 (apr 2024), 26 pages. doi:10.1145/3637405
- [67] Antti Mattila et al. 2001. "Seeing things in a new light": Reframing in therapeutic conversation. Ph.D. Dissertation. Citeseer.
- [68] Martino Mensio, Gregoire Burel, Tracie Farrell, and Harith Alani. 2023. MisinfoMe: A Tool for Longitudinal Assessment of Twitter Accounts' Sharing of Misinformation. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization* (Limassol, Cyprus) (UMAP '23 Adjunct). Association for Computing Machinery, New York, NY, USA, 72–75. doi:10.1145/3563359.3597396
- [69] Miriam J Metzger. 2007. Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American society for information science and technology* 58, 13 (2007), 2078–2091.
- [70] Miriam J Metzger and Andrew J Flanagin. 2013. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of pragmatics* 59 (2013), 210–220.
- [71] Michael B Miller and Michael S Gazzaniga. 1998. Creating false memories for visual scenes. *Neuropsychologia* 36, 6 (1998), 513–520.
- [72] Gillian Murphy, Caroline A Dawson, Charlotte Huston, Lisa Ballantyne, Elizabeth Barrett, Conor S Cowman, Christopher Fitzsimons, Julie Maher, Katie M Ryan, and Ciara M Greene. 2023. Lost in the mall again: a preregistered replication and extension of Loftus & Pickrell (1995). *Memory* 31, 6 (2023), 818–830.
- [73] Gillian Murphy and Emma Flynn. 2021. Deepfake false memories. *Memory* 30, 4 (April 2021), 480–492. doi:10.1080/09658211.2021.1919715
- [74] Beate Muschalla and Fabian Schönborn. 2021. Induction of false beliefs and false memories in laboratory studies—A systematic review. *Clinical Psychology & Psychotherapy* 28, 5 (Feb. 2021), 1194–1209. doi:10.1002/cpp.2567
- [75] Sophie J Nightingale and Hany Farid. 2022. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences* 119, 8 (2022), e2120481119.
- [76] Gabriel A Orenstein and Lindsay Lewis. 2022. Eriksons stages of psychosocial development. In *StatPearls [Internet]*. StatPearls Publishing.
- [77] David G Payne, Claude J Elie, Jason M Blackwell, and Jeffrey S Neuschatz. 1996. Memory illusions: Recalling, recognizing, and recollecting events that never occurred. *Journal of Memory and Language* 35, 2 (1996), 261–285.
- [78] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.
- [79] Elizabeth A Phelps and Stefan G Hofmann. 2019. Memory editing from science fiction to clinical practice. *Nature* 572, 7767 (2019), 43–50.
- [80] Daniel Pillis, Pat Pataranutaporn, Pattie Maes, and Misha Sra. 2024. AI Comes Out of the Closet: Using AI-Generated Virtual Characters to Help Individuals Practice LGBTQIA+ Advocacy. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 686–698.
- [81] Fritz Renner, Fionnuala C. Murphy, Julie L. Ji, Tom Manly, and Emily A. Holmes. 2019. Mental imagery as a "motivational amplifier" to promote activities. *Behaviour Research and Therapy* 114 (March 2019), 51–59. doi:10.1016/j.brat.2019.02.002
- [82] Bradley J Rhodes. 1997. The wearable remembrance agent: A system for augmented memory. *Personal Technologies* 1 (1997), 218–224.
- [83] E.F. Risko, M.O. Kelly, P. Patel, and C. Gaspar. 2019. Offloading memory leaves us vulnerable to memory manipulation. *Cognition* 191 (Oct. 2019), 103954. doi:10.1016/j.cognition.2019.04.023
- [84] Frédérique Robin, Emmanuelle Ménétrier, and Brice Beffara Bret. 2022. Effects of visual imagery on false memories in DRM and misinformation paradigms. *Memory* 30, 6 (2022), 725–732.
- [85] Nathaniel Sirlin, Ziv Epstein, Antonio A Arechar, and David G Rand. 2021. Digital literacy is associated with more discerning accuracy judgments but not sharing intentions. *Harvard Kennedy School, Misinformation Review* (2021).
- [86] Scott D Slotnick and Daniel L Schacter. 2004. A sensory signature that distinguishes true from false memories. *Nature Neuroscience* 7, 6 (2004), 664–672.
- [87] Joel R. Sneed, Susan Krauss Whitbourne, and Michelle E. Culang. 2006. Trust, Identity, and Ego Integrity: Modeling Erikson's Core Stages Over 34 Years. *Journal of Adult Development* 13, 3–4 (Dec. 2006), 148–157. doi:10.1007/s10804-007-9026-3
- [88] Christian Stephan-Otto, Sara Siddi, Carl Senior, Daniel Muñoz-Samons, Susana Ochoa, Ana María Sánchez-Laforga, and Gildas Brébion. 2017. Visual imagery and false memory for pictures: a functional magnetic resonance imaging study in healthy participants. *PLoS One* 12, 1 (2017), e0169551.
- [89] Daniel Strain. 2024. AI images abound this election cycle. Here's how you can tell fact from fiction — colorado.edu. <https://www.colorado.edu/today/2024/08/29/ai-images-abound-election-cycle-heres-how-you-can-tell-fact-fiction>. [Accessed 12-09-2024].
- [90] Ben M Tappin, Chloe Wittenberg, Luke B Hewitt, Adam J Berinsky, and David G Rand. 2023. Quantifying the potential persuasive returns to political microtargeting. *Proceedings of the National Academy of Sciences* 120, 25 (2023), e2216261120.
- [91] Rimke Tas. 2023. *Changing memories via Deepfakes*. Ph.D. Dissertation. Ghent University.
- [92] Emily Thorson. 2015. Belief Echoes: The Persistent Effects of Corrected Misinformation. *Political Communication* 33, 3 (Nov. 2015), 460–480. doi:10.1080/10584609.2015.1102187
- [93] Jay Ulfelder. 2024. The Real Numbers: Tracking Crowd Sizes at Presidential Rallies — Ash Center — ash.harvard.edu. <https://ash.harvard.edu/articles/the-real-numbers-tracking-crowd-sizes-at-presidential-rallies/>. [Accessed 12-09-2024].
- [94] Jacqueline Urakami, Yeongdae Kim, Hiroki Oura, and Katie Seaborn. 2022. Finding Strategies Against Misinformation in Social Media: A Qualitative Study. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 242, 7 pages. doi:10.1145/3491101.3519661
- [95] Jan G Voelkel, Robb Willer, et al. 2023. Artificial Intelligence Can Persuade Humans on Political Issues. *OSF Preprints* (2023).
- [96] 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. *Psychonomic bulletin & review* 9, 3 (2002), 597–603.
- [97] Jianqin Wang, Henry Otgaar, Mark L Howe, Felix Lippe, and Tom Smeets. 2018. The nature and consequences of false memories for visual stimuli. *Journal of Memory and Language* 101 (2018), 124–135.
- [98] Yuping Wang, Chen Ling, and Gianluca Stringhini. 2023. Understanding the Use of Images to Spread COVID-19 Misinformation on Twitter. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 108 (apr 2023), 32 pages. doi:10.1145/3579542
- [99] Robert Weinberg. 2008. Does Imagery Work? Effects on Performance and Mental Skills. *Journal of Imagery Research in Sport and Physical Activity* 3, 1 (Feb. 2008). doi:10.2202/1932-0191.1025
- [100] John T Wixted, Gary L Wells, Elizabeth F Loftus, and Brandon L Garrett. 2021. Test a witness's memory of a suspect only once. *Psychological Science in the Public Interest* 22, 1_suppl (2021), 1S–18S.
- [101] Mengxi Xiao, Qianqian Xie, Ziyang Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. HealMe: Harnessing Cognitive Reframing in Large Language Models for Psychotherapy. *arXiv preprint arXiv:2403.05574* (2024).
- [102] Danni Xu, Shaojing Fan, and Mohan Kankanalli. 2023. Combating misinformation in the era of generative AI models. In *Proceedings of the 31st ACM International Conference on Multimedia*. 9291–9298.
- [103] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. 2023. Diffusion models: A comprehensive survey of methods and applications. *Comput. Surveys* 56, 4 (2023), 1–39.
- [104] Elizabeth M Zelinski, Michael J Gilewski, and Cheri R Anthony-Bergstone. 1990. Memory Functioning Questionnaire: concurrent validity with memory performance and self-reported memory failures. *Psychology and aging* 5, 3 (1990), 388.

- [105] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic lies: Understanding ai-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [106] Jiawei Zhou, Yixuan Zhang, Qianni Luo, Andrea G Parker, and Munmun De Choudhury. 2023. Synthetic Lies: Understanding AI-Generated Misinformation and Evaluating Algorithmic and Human Solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 436, 20 pages. doi:10.1145/3544548.3581318
- [107] L Zhuang, J Wang, B Xiong, C Bian, L Hao, PJ Bayley, and S Qin. 2022. Rapid neural reorganization during retrieval practice predicts subsequent long-term retention and false memory. *Nature Human Behaviour*, 6 (1), 134-145.
- [108] Wazeer Deen Zulfikar, Samantha Chan, and Pattie Maes. 2024. Memoro: Using Large Language Models to Realize a Concise Interface for Real-Time Memory Augmentation. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 450, 18 pages. doi:10.1145/3613904.3642450

8 Supplementary Materials

8.1 False Memories Questionnaire

Please answer the following questions regarding the **original images**:

- (1) Did you remember seeing the bride smiling in the original picture?
 - (2) Did you remember seeing the sun in the original picture?
 - (3) Did you remember seeing this woman with blond hair?
 - (4) Did you remember seeing a sign on the street?
 - (5) Did you remember seeing the woman wearing any head-wear?
 - (6) Did you remember seeing a flag in the picture?
 - (7) Did you remember seeing one of the boys using a smart-phone?
 - (8) Did you remember seeing a boat in the picture?
 - (9) Did you remember seeing two black guys running in the picture?
 - (10) Did you remember seeing military presence in the picture?
 - (11) Did you remember seeing a military bodyguard in the picture?
 - (12) Did you remember seeing a military vehicle in the picture?
 - (13) Did you remember seeing only one former US president in the picture?
 - (14) Did you remember seeing this guy in a mask?
 - (15) Did you remember seeing guys in hazmat suits walking with a suitcase?
 - (16) Did you remember seeing coffee mugs on the table?
 - (17) Did you remember seeing that all the land was covered by snow (no ocean water)?
 - (18) Did you remember seeing a woman in this group shot?
 - (19) Did you remember seeing the guy wearing a military uniform?
 - (20) Did you remember seeing a city in the picture?
 - (21) Did you remember seeing an adult in the picture?
 - (22) Did you remember seeing a woman in the picture?
 - (23) Did you remember seeing a table in the middle?
 - (24) Did you remember seeing a small bicycle in front of the boys?
- Agree
 - Unsure
 - Disagree

How confidence are you with the answer?

- 1-Extremely lacking confidence
- 2-Very lacking confidence
- 3-Somewhat lacking confidence
- 4-Neutral
- 5-Somewhat confident
- 6-Very confident
- 7-Extremely confident

8.2 AI filter familiarity

- What is your level of familiarity with using image filter technologies or AI filter technologies? (1-Not familiar at all, 7-Very Familiar)

8.3 Frequency of forgetting

Participants self-reported memory problems, using a scale ranging from 1 (major problems) to 7 (no problems), as adapted from [104].

- How would you rate your memory in terms of the kinds of problems that you have? (1= Major problems to 7= No problems)

8.4 Memory Efficacy

A self-report measure of memory ability (Self-Efficacy Level) taken from a subset of [3].

- (1) If someone showed me the pictures of 16 common everyday objects, I could look at the pictures once and remember the names of 2 of the objects.
- (2) If someone showed me the photographs of 10 people and told me their names once, I could identify 2 persons by name if I saw the pictures again a few minutes later.
 - 1-Strongly disagree
 - 2-Disagree
 - 3-Somewhat disagree
 - 4-Neither agree nor disagree
 - 5-Somewhat agree
 - 6-Agree

- 7-Strongly agree

8.5 Skepticism

The scale to measure naive skepticism in the adult population is taken from [21].

- (1) The official media provides false information
- (2) I distrust the information provided by government authorities
- (3) The World Health Organization (WHO) hides its true interests
- (4) The world press manipulates information
- (5) Social networks call those who tell uncomfortable truths crazy
- (6) The rich manipulate press
- (7) International organizations only deliver information that benefits them
 - 1-Strongly disagree
 - 2-Disagree
 - 3-Somewhat disagree
 - 4-Neither agree nor disagree
 - 5-Somewhat agree
 - 6-Agree
 - 7-Strongly agree