UNIVERSITY OF CALIFORNIA,
IRVINE

Deepfake Deception: Examining the Effects on False Memory Formation, Digital Media Discernment, and Generation Z

DISSERTATION

submitted in partial satisfaction of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in Informatics

by

Nika Pari Nour

Dissertation Committee:
Professor Constance Steinkuehler, Chair
Professor Elizabeth Loftus
Professor Paul Dourish

2025

DEDICATION

To

my younger self,

thank you for being endlessly curious,

for asking the hard questions,

and for believing that technology can shape the future.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

This dissertation work was made possible with the unwavering support, mentorship, and inspiration of many who shaped my academic and professional journey.

To my committee— Dr. Constance Steinkuehler, Dr. Elizabeth Loftus, and Dr. Paul Dourish—thank you for your guidance, feedback, and understanding of the value of this work. I am especially grateful for your scholarly support, patience, and encouragement throughout the process's highs and lows.

To my family— Sandy, Saum, Niki, Niku, and my incredible in-laws—thank you for your unconditional love, sacrifices, and grounding support.

To my loving husband— Matthew, you reminded me to keep going when it was hard and celebrated every win like it was your own.

I gratefully acknowledge the support of the Google PhD Human-Computer Interaction Fellowship, which made this research possible and provided the freedom to pursue bold, curiosity-driven work.

To the teachers and educators who shaped my academic foundation and critical thinking— Dr. Roderic Crooks, Dr. André van der Hoek, Dr. Melissa Mazmanian, Dr. Kurt Squire, Dr. Rebecca Black, Dr. Bill Tomlinson, Dr. Sean Young, Dr. Braxton Soderman, Katie Salen, Dr. Aure Schrock, and Nguyễn Đăng Tiến Dũng, thank you for the guidance, tutelage, and motivation.

In industry— I'm deeply appreciative to the mentors, colleagues, and collaborators who opened doors and championed this work along the way. Dr. Luke Dicken, Shannon Loftis, Michael Gallagher, Ed Fries, Jen MacLean, Jacob Navok, Christopher Cataldi, Yaprak DeCarmine, Steve Spohn, and Jesse Meschuk, you played a pivotal role in shaping the intersection between my career, academic pursuits, and the games and technology landscape.

VITA

Nika Pari Nour

EDUCATION

Doctor of Philosophy in Informatics **2025**
University of California, Irvine

Masters of Science in Informatics **2024**
University of California, Irvine

Masters of Science in Biodefense **2012**
George Mason University

Masters of Arts in Strategic Public Relations **2011**
George Washington University

Bachelor of Arts in Communication and Public Relations **2009**
California State University, Los Angeles

PUBLICATIONS

Nour, N., Wells, G. & Steinkuehler, C. (2023). *Hireability Gaps and Frictions Between University Game Programs and Industry Hires.* International Game Developers Association.

Nour, N. & Gelfand, J. (2022). Deepfakes: A Digital Transformation Leads to Misinformation. *The Gray Journal 18*(2), 85–94.

Nour, N., O'Leary, S.R., Martinez, R., Gardner, R., Anderson, C.G., & Steinkuehler, S. (2021). Network of Academic and Scholastic Esports Federations Internal Report: Diversifying Student Participation.

CONFERENCE PRESENTATIONS

Nour, N: Ethical Boundaries of the Mediatech Revolution 2024. Media Innovation Xchange. EU Conference on Digital Transformation and Media Innovation. Brussels, Belgium.

Nour, N. AI-Generated Fakery: The State of Deepfake Research and Future Work. Games Learning Society Conference 2022. Irvine, CA

Nour, N. Deepfakes: A Digital Transformation Leads to Misinformation. Conference on Grey Literature 2021. Amsterdam, Netherlands.

ABSTRACT OF THE DISSERTATION

Deepfake Deception: Examining the Effects on False Memory Formation, Digital Media Discernment, and Generation Z

By

Nika Pari Nour

Doctor of Philosophy in Informatics

University of California, Irvine, 2025

Professor Constance Steinkuehler, Chair

The rise of deepfake technology—AI-generated videos that convincingly depict fabricated events—poses growing challenges for digital media discernment, public trust, and memory. While technical detection methods have advanced, little empirical research has analyzed how deepfakes influence human cognition and memory. This dissertation empirically assesses whether deepfake videos can implant false memories in Generation Z adults and whether active debriefing or media literacy interventions mitigate these effects.

Drawing from literature in cognitive science, false memory, and misinformation studies, this research employs an experimental design involving over 1,600 participants aged 18–26. Participants were exposed to a combination of real videos and deepfake videos featuring celebrity figures. Later, they were assessed for memory retention, deepfake detection, and confidence in discernment. Results show that deepfakes can produce false memories, especially when participants are unaware that they are viewing falsified videos. However, active debriefing

(informing participants post-exposure about the presence of deepfakes) significantly reduced the retention of these false memories.

Contrary to expectations, prior knowledge of the featured celebrities, familiarity with deepfake technology, and frequency of social media use did not improve detection accuracy or false memory resilience. Participants frequently relied on intuitive audiovisual cues (e.g., lip-sync mismatches and unnatural eye movements) over factual reasoning, and often expressed high confidence in inaccurate judgments. Emotional responses to deepfakes were also pronounced. Many participants reported anxiety, mistrust, and moral concern about the implications of deepfake technology and the irresponsible use of generative AI tools.

Chapter 1

Introduction

Social networks have increased the dissemination of false, incomplete, or inaccurate information across various forms of media content, including text, images, and animated visuals (Peters, 2017). Specifically, they have broadened the reach of deepfake videos—manipulated videos that convincingly distort truth and portray fabricated events as reality. A public concern with deepfake videos is whether they can implant false memories. These false memories can range from minor inaccuracies, such as misremembering story details, to significant distortions of traumatic or dramatic life events (Newman & Lindsay, 2009). As such, this research investigates whether deepfake videos can implant such false memories, directly or indirectly, thereby enriching our scholarly understanding of their broader psychological and societal effects. Currently, little is known about an individual's ability to discern deepfakes, independent of technical assistance, and whether media literacy— the capacity to access, analyze, evaluate, and create messages in various forms (Livingstone, 2004)—can mitigate its adverse impacts on scientific, political, social, and cultural dynamics. To address this gap, this study introduces research instruments designed to actively debrief and educate individuals on deepfake technologies and assess whether media literacy can effectively prevent the formation of false memories following deepfake exposure. The remainder of this introduction will outline the structure of the dissertation and present a brief overview of subsequent chapters.

1.1 The Rise of Falsified Audiovisual Media

In the new internet era of social media platforms and content manipulation, misinformation has become a central issue in contemporary media discourse. However, the existing body of research on the consequences of digitally altered audiovisual content and deepfakes— highly realistic videos generated by artificial intelligence (AI) depicting people saying and doing things that never happened—is limited (Westerlund, 2019). Evaluating the distribution of deepfakes necessitates considering its numerous determinants: the rapid pace of technology, the recent growth in digitally altered content, and users' varying levels of trust in visual imagery. As emerging technological advances become more sophisticated, the end-user's ability to trust, reason, or discern falsified information has become increasingly evident (Ternovski et al., 2021). Investigating the potential consequences of deepfakes is a significant area of concern within academia, as these AI-generated videos can create credible-looking, realistic videos capable of implanting false memories. Researchers Madine Liv and Dov Greenbaum assert that deepfakes deceive even responsible media consumers, embed false memories, and manipulate end-users (2020). To date, there has been little agreement on how individuals discern the trustworthiness of social media content across multiple modalities, specifically deceptive multimedia videos, and their confidence in their discernment abilities. Thus, understanding the relationship between deepfakes and false memories is central to developing interventions aimed at mitigating the long-term consequences of exposure to such content.

Between social media platforms, messaging apps, and daily internet engagement, today's society represents the most digitally interconnected period in the history of civilization (Volti & Croissant, 2024). Technological change has immersed a generational cohort in an

internet-connected environment (Kardaras, 2016), Generation Z (Gen Z), also referred to as "Zoomers" or the "iGeneration" (Răduț, 2021). As the world's largest consumer demographic, Gen Z consists of individuals born between 1997–2012 (Grigoreva et al., 2021). Gen Z, having grown up in an era of technological convenience, is often assumed to possess high levels of digital literacy and technological competence. However, Bourke (2019) challenges this assumption, suggesting that their technical skills may be more limited than commonly believed. This makes Gen Z a particularly intriguing subject for research into susceptibility to misinformation and deepfakes, as they are considered the most technologically literate adult generation to date. Given their extensive access to diverse media channels, networks, and platforms, Gen Z also tends to have a broader definition of "news" compared to previous generations. Social media platforms like YouTube, Facebook, TikTok, and streaming services distribute a wide array of multimodal content, including videos, memes, graphics, and viral trends (Panagiotou et al., 2022). Consequently, the digital ecosystem of social media is frequently rife with misinformation, which requires users and Gen Z'ers within the United States and internationally to enhance their resilience to falsified media and self-moderate content. With the propagation of AI-generated content, tools, and even bots, the boundaries between real and fake are becoming harder to discern. As such, younger, presumably tech-literate generations require proficiency in media literacy skills to enhance their critical evaluation of information and reduce susceptibility to intentional deceit or malicious social media content (Silveira, 2020).

1.2 Beyond Fake News: The Internet's Credibility Crisis

With the increasing prevalence of social media bots, fake profiles, and a lack of accountability, individuals can propagate fabricated personas to make deepfakes seem believable (Le, 2020). AI-generated bots extend beyond publicly distributing deepfakes on social media

platforms; they can also transmit information in calculated and targeted ways. Deepfake campaigns exist on Twitter as TweepFakes (Fagni et al., 2021), on Facebook as cloaked pages (Farkas et al., 2017), and elsewhere online as “cheap fakes” (Paris et al., 2019). End users can construct seemingly reputable profiles of people who do not exist using AI-generated images, build comprehensive websites, and design social media presences. Creating false personas to distribute content can result in users perceiving these fake personas as credible sources of information without ever confirming their veracity. As a result, internet users’ confidence in online content fades. While deepfakes are increasing in prevalence, concerns about their ability to impact critical societal functions and influence political outcomes are escalating. To these ends, it is essential to rigorously examine how falsified videos can implant memories through familiar, realistic, and credible content (Nash et al., 2009).

Without developing detection strategies, deepfakes can inflict personal and societal harm that threatens the foundations of trust in society. In addition to these moral reasons for exploring the implications of deepfake technologies, there are scholarly concerns about their impact on the end-user's trust and multimodal content integrity. Given these circumstances, existing research does not sufficiently address the ongoing problems for end-users. The tools and creation process of falsified content and information are rapidly evolving, as developing detection capabilities move at a slower pace. Current research on false memories covers instances of falsified autobiographical events (Heaps & Nash, 2001), interview methods (Loftus & Pickrell, 1995), fake news stories (Polage, 2012), court evidence standards (Tortora, 2013), and doctored photographs (Garry & Gerrie, 2005). This dissertation aims to build on these foundations by addressing how deepfakes specifically influence memory formation and retention. It offers a more detailed understanding of how deceptive multimedia content can impact tech-literate

individuals and their collective memories. Though there have been investigations of whether or not deepfakes can distort memories (Murphy & Flynn, 2021), the research lacks comprehensive data on the effects of deepfakes, and researchers are only beginning to delve into the subject matter. The data collected in this dissertation will help determine whether stakeholders should invest in preventative measures to improve media and content literacy, as it is paramount for the global community to learn how to judge the truthfulness of content, question the verification of sources, and reason through misinformation and disinformation literacy.

1.3 What This Study Seeks to Investigate

This dissertation investigates whether deepfakes form false memories among the Gen Z population. To investigate these aims, this study deploys quantitative methods. It utilizes deception, in which recruited Gen Z participants are asked to partake in a social media marketing survey on celebrity influencers promoting products. In order to evaluate whether deepfakes create false memories, participants will be exposed to deepfake content and then take a post-test 24 to 96 hours later to assess memory retention. Next, some participants will be told the study's true purpose before the post-test, while others will remain unaware until after completing the post-test. Using deception is imperative to minimize bias and ensure the validity of the results. The findings will also determine whether or not participants, who are actively debriefed and informed that the videos are deepfakes and that their contents are false, still retain false memories after this disclosure.

Before proceeding to examine how deepfakes influence public trust and perception, it is essential to explore if digital competency, media literacy, and prior knowledge of content enable individuals to differentiate between real and fake social media content, accurately identify misinformation, and effectively verify content sources (Anthonysamy & Sivakumar, 2022).

Deepfake research has yet to determine if the quality of deepfakes matters or affects an end user's ability to determine if the content presents a fictional situation. Thus, to what extent must the quality and resolution of deepfakes reach before average individuals begin to question or verify the authenticity of the content? This study distills the data collected regarding the participants' accuracies in deepfake detection, their levels of confidence in discerning deepfakes, and whether their confidence varies by the accuracy of their discernment. The results provide in-depth insights into the specific cues and clues that participants rely on to discern deepfakes and whether certain cues and clues are more efficacious than others. To these ends, the following questions guided this research:

- RQ1. Can exposure to a deepfake video (Kim Kardashian or Elon Musk) implant false memories?
- RQ2. Can an active debriefing (informing participants that the video is a deepfake) deter the implanting of false memories?
- RQ3. What is the relationship between prior knowledge and the likelihood of a deepfake video (featuring Kim Kardashian or Elon Musk) implanting false memories?
- RQ4. What is the relationship between prior knowledge of the video content (specifically, the celebrity) and the effectiveness of active debriefing?
- RQ5. Is there a relationship between social media use frequency and the effects of deepfakes implanting false memories?
- RQ6. Is there a relationship between prior knowledge of deepfakes or previous exposure to deepfakes and the likelihood of false memories being implanted?
- RQ7. How accurately do people discern whether a video is real or a deepfake?
- RQ8. How confident are they in their discernment?

- RQ9. What cues and clues do individuals use for discernment?
- RQ10. Are some cues and clues more or less efficacious for discernment than others?
- RQ11. How can exposure to information about deepfakes influence individuals' attitudes and perceptions?

1.4 Definitions for Variables and Key Terms

This section provides definitions for the variables and further defines terms used in the study, particularly those with multiple meanings in other fields of research or are considered technological jargon.

Active Debriefing.

Participants are informed whether or not they were exposed to deepfakes, asked to discriminate between falsified and true videos, and then asked to distinguish which videos are deepfakes (Barari et al., 2021).

Age.

An individual characteristic of the information consumer and a demographic variable (Guess, Nyhan, et al., 2020; Grinberg, 2019) that may play a key role in individual susceptibility to misinformation (Pennycook & Rand, 2019b).

Deepfakes.

Using deep learning algorithms, deepfake videos (or colloquially, just “deepfakes”) substitute one person’s visual and acoustic likeness for another, presenting viewers with highly convincing videos in which individuals appear to be performing actions or making statements they never did or said (Vaccari et al., 2020).

Discernment, Accuracy of.

Gauging if users can accurately discern deepfake videos (Van Duyn & Collier, 2018), judge the trustworthiness of the sources (Effron & Raj, 2019), and verify the truthfulness of the content (Fletcher & Park, 2017).

Discernment, Confidence and Judgement.

The individual's judgment, certainty, and feelings in determining whether information is true or false (Salovich et al., 2021).

Discernment, Cues and Clues Used.

The individual's evaluation of the veracity of the information using content-related features, perception of content based on personal biases, deliberation of source credibility (Bago et al., 2020), physical characteristics of the video (such as size, width, audio and visual quality) (Singh et al., 2021).

False Memory.

An erroneous memory that can vary from misremembering specific details of an event to false recollection of live events (Newman & Lindsay, 2009), generally a recollection that can make it difficult to differentiate between what is real and what is not (Loftus & Berstein, 2005).

Generation Z (Gen Z).

The generational cohort referred to as "Zoomers," the "iGeneration," or "digital natives" (Răduț, 2021) born into technological convenience between the years of 1997–2012 (Grigoreva et al., 2021).

Generative Artificial Intelligence (AI).

Advanced machine learning algorithms and large language models that streamline and accelerate the ease and accessibility of deepfake creation (Shoaib et al., 2023).

Media Literacy.

Strategies, such as critical thinking, logical reasoning, technology recognition, attention to detail, and content assessment, designed to develop skills and aid individuals in critically interpreting events, new, and online content beyond how the information is presented by media, distribution channels, and digital platforms, fostering deeper analysis and evaluation of the content's underlying messages and implications (Silveira, 2020).

Prior Knowledge (of Celebrity or Person).

Also known as information literacy, the ability to effectively evaluate and use prior knowledge about the content (celebrities) of the information itself (Karmarkar & Tormala, 2010).

Prior Knowledge (of Deepfakes).

A form of technology literacy. The ability to effectively evaluate and use prior knowledge about deepfake technologies and deepfake video content (Korona & Hutchison, 2023).

Social Media.

Social networking sites, internet applications, and individual websites, such as Instagram, TikTok, and YouTube, allow millions of end-users to easily share information online in multi-modal formats such as text, image, and video (Chatzoglou et al., 2020).

Social Media Use.

The individual’s level of engagement, average time spent per day, and types of content consumption on varied social networking sites, messaging apps, or streaming services (Pérez-Escoda et al., 2021).

1.5 The Social and Scholarly Relevance of Today

The sociotechnical effects of deepfakes remain significantly understudied, and their escalating contributions to misinformation are likely to increase. As such, it is imperative to

understand how falsified videos can influence individuals and shape public perception, particularly given that deepfakes pose a national security threat. Both state and non-state adversaries alter media and promote misinformation to advance their strategic objectives. Currently, there are limited corporate policies, federal and state laws, and repercussions for weaponizing deceptive online content, distributing misinformation, and using identities without permission. The field of informatics requires a rigorous analysis of deepfakes to thoroughly understand their short- and long-term social and cultural consequences on memories. At stake are threats to privacy, national security, the First Amendment, data infrastructure, and information warfare. Though artificial intelligence as a tool is neither classified as good nor bad, the replication of an individual's identity without consent poses true privacy and security concerns.

1.6 Structuring the Investigation

To investigate these questions, this dissertation has been organized into six chapters. The study's research objectives, context, and value are introduced in chapter one. In chapter two, the literature review summarizes the existing studies, research gaps, and unexplored territories in deepfakes, false memories, and content detection. In chapter three, the dissertation includes the methodologies and quantitative instruments employed in the adoption of an experimental design with a deductive approach. Chapter four presents the results of determining whether participants recalled false memories from the experiment. The study included two different treatments, two falsified videos, three real videos, and a delayed post-test. Chapter five presents the summary and discussion of the findings corresponding to each research question. Finally, the conclusion synthesizes the dissertation that encompasses key takeaways, limitations to the study, and potential future work to further this research area.

Chapter 2

Background and Related Work

This chapter serves as the foundation of the dissertation and summarizes what is known about existing research in this field, identifies gaps in the current literature, and provides theoretical foundations for developing my research framework. As noted in the introduction, the goal is to identify how falsified videos can influence people, specifically those in "Generation Z," if active debriefing (informing participants that the video is a deepfake) can deter the implanting of false memories, and whether or not people can detect deepfakes. Thus, the literature review will define what is known about deepfakes, explore media literacy intervention strategies, and synthesize previous research in the false memory research field.

2.1 Defining Deepfakes as a Form of Misinformation

Using deep learning algorithms, "deepfake" videos (or colloquially, just "deepfakes") typically substitute one person's visual and acoustic likeness for another, presenting viewers with compelling videos of individuals doing and saying things they never did or said (Vaccari et al., 2020). In recent events, deepfake videos have been disseminated to exert influence on elections, instigating violence, and altering claims to truths (Paris et al., 2010). The need to verify online content and imagery is not new, yet academic research exploring deepfakes has only recently emerged. The current state of deepfake research necessitates approaches beyond administering empirical studies centering on AI experiments and relying on technical detection methods

(Godulla et al., 2021). Similarly, literature compellingly proposes that scholars delve into the socio-technical effects and ramifications of deepfakes and their moderating effects on false memories, encompassing analysis of opportunities, risks, regulations, and methods to combat deepfakes created with ill intent (Godulla et al., 2021). Due to the rising prevalence of deepfake videos and their ability to promulgate violence and mistrust, it is increasingly vital for academics to deploy their interdisciplinary expertise, including technical and theoretical methods, in combating misinformation online.

What makes deepfake videos especially worrisome is the relative ease and accessibility with which adversarial actors can manipulate moving images (Almars, 2021). With virtually no technical expertise, individuals can generate untraceable, deceptive videos and circulate them across the global digital landscape without geographical constraints. Altered videos have the potential to spread propaganda, recruit new members, incite violence, or reword a politician's speech (Carvajal, 2021). The elements for evaluating non-textual media and social media distribution include the rapid pace of technology, the recent growth in digitally altered content, and users' varying levels of trust in visual imagery. Early scholarly inquiries have analyzed how content influences trust in online information, information presentation, and individual values (Verma et al., 2017). However, these studies have yet to comprehensively utilize these measures to evaluate the verification and trust of altered audio-visual media.

While "algorithmic detection " is a process where computer algorithms are used to identify and flag potential deepfake videos—exists, access to and deployment of these technologies remain uneven, and artificial intelligence blockers have struggled to eliminate harmful online posts and publications (Delort & Paris, 2011). Deepfakes, as a tool, can exacerbate harmful behaviors by allowing bad actors to create falsified, salacious videos that

misrepresent individuals, leading to name-calling, harassment, and other forms of abuse. These manipulated videos can also be used to incite doxxing (the release of personal information) or spread misinformation, all with the intent of damaging reputations and provoking harmful actions against the targeted individuals. In addition, deepfake videos, with the rhetorical persuasiveness of the moving image, represent a significant jump in the ability to inflict social damage or, perhaps, implant false memories. Acknowledging that the algorithmic detection of deepfake videos is mired in dubious reliability, accessibility, and credulity, there is a robust, practical impetus to evaluate—and, where necessary—cultivate the human capacity for discerning between legitimate and deepfake manipulated videos ("Deepfake Videos Easily Fool Face Systems," 2019).

2.2 Defining "Deepfakes" as Media Manipulation

Modifying images and videos is a well-established practice within the domain of visual media. An early example of media manipulation combines a portrait of Abraham Lincoln in 1860 and a print of John Calhoun from 1852 (Chawla, 2019). Though the image seemed real, it was actually multiple photographs of Lincoln's head and John Calhoun's body merged into one portrait (Caldera, 2020). As image and photo manipulation became more common, video editing and manipulation took hold in the 1970s, when computer animation and visual effects became widely adopted in the entertainment industry (Caldera, 2020). Over time, the emergence of hyper-realistic simulations and media surpassed the simple alterations applied to videos and photographs, embodying a significantly more ominous trajectory in recent years.

Another significant aspect of evolving media fake news technologies involves transforming falsified audiovisual content, which increasingly exhibits realistic appearance and sound (Johnson & Diakopoulos, 2021). The innovative process that led to these manipulations

includes using hundreds and possibly thousands of images of the person being deepfaked. These images serve as a dataset to create this type of content (Albahar & Almalki, 2019). Examples of the deepfake phenomenon include videos of celebrities doing things they never did, such as Mark Zuckerberg stating that he would delete Facebook (Kietzmann et al., 2020). While the manipulation of images, audio, and video content is not novel, the phenomenon of deepfakes has precipitated a boom in this new form of deceptive multimedia.

In 2017, an anonymous Reddit user, using the pseudonym "deepfakes," shared the computer code to place famous personalities into pornographic clips (Kietzmann et al., 2020). The anonymous creator of the deepfake method, a software engineer, released a development kit that streamlined the process for others to create their deepfakes (Gardiner, 2019). Consequently, numerous deepfakes were then generated and shared on discussion forums called subReddits, resulting in the platform banning deepfakes in AI porn communities on their website in 2018 (Albahar & Almalki, 2019). The ban encompassed numerous online platforms, including Twitter and Discord; nevertheless, the methodologies for manufacturing deepfakes had already been propagated worldwide. Culminating resources from companies providing open-source software, emerging threats drew the attention of the Defense Advanced Research Project Agency (DARPA), an institution under the U.S. Department of Defense's jurisdiction (Government of Canada, 2002). These threats underscore the accessibility and viability inherent in deepfake creation, further burgeoning the necessity for robust policies and institutional response.

2.3 Types of Deepfake Content

Manipulating visual media is becoming widely accessible, inexpensive, and easier to accomplish (Wade et al., 2002). With software becoming more user-friendly, cell phone cameras increasing in quality, and the rise of technology literacy, most people have the opportunity to

doctor media, an option once reserved for video editing professionals and filmmakers. Though "fake news" and artificial intelligence garner significant attention in the modern news media landscape, sociotechnical scholarly explorations of digitally altered visual content, specifically deepfake videos, are notably underdeveloped. This presents a paramount area for further exploration and discussion. In contemporary digital culture, the most prevalent and widely circulated (nonpornographic) deepfakes encompass a broad spectrum of content, including altered political speeches, celebrities dancing in venues they have never attended, and copyright violations of Hollywood's most prominent franchises and intellectual properties (Langguth et al., 2021). By extension, celebrities and famous individuals—business owners, actors, musicians, and athletes—are consistently the principal targets of deepfake fabrications as their social media networks are primed for widespread distribution (Pérez et al., 2021). As illustrated in Table 2.1, the contemporary milieu of existing deepfake videos reveals various content featuring prominent figures, including celebrities, business leaders, international dignitaries, and elected officials. Moreover, the public's intense interest in celebrities creates a sense of familiarity and connection, making people feel as if they 'know' figures like Taylor Swift, despite never having met them. This perceived intimacy amplifies the impact of false or damaging content, as fans and followers are more likely to engage with and spread such information, leading to real reputational harm.

Table 2.1

Types of Viral Deepfake Video Content Currently Available

Social Roles of Deepfake Subject	Content Type or Category					
	Dancing or Singing	Interview	Speech	Copyright Infringement	Parody or Creative	Adult Content
Elected Official		X	X		X	X
Celebrity	X	X	X	X	X	X
Business Leader		X	X	X		X
International Leader		X	X			X

The advancement of deepfake video technology is progressively simplifying and enhancing accessibility for unskilled end-users through programs such as FakeApp, a face-swapping program, Zao, a Chinese mobile app using movie clips, and Apple's text-to-speech (TTS) editing system (Kietzmann et al., 2020). There is an imperative for research and initiative expansion on deepfakes in the social sciences, hard sciences, and interdisciplinary space, as the types of content, victims, and agendas in deepfake videos extend beyond the typical deepfakes of porn and adult videos. In 2019, Adjer et al. corroborated that 96% of deepfakes were pornographic. Having defined the substantial amount of adult-rated deepfakes, political scientists aim to comprehend the implications and intersections between political and pornographic deepfakes, which function in similar ways to suppress critical speech (Maddocks, 2020). The media, in particular, appears to offer cultural contextual elucidation, compromise the integrity of the political process, and delineate the challenges associated with pornographic and

political deepfakes, particularly those involving high-profile individuals who are non-consensually deepfaked (Gosse & Burkell, 2020).

Due to restrictions by social media platforms and government regulations, pornographic content is not usually widely shared on Facebook, Twitter, YouTube, and Instagram without consequences. Before 2018, pornographic face swapping, hardcore materials, and revenge porn were accessible on Pornhub and Reddit; however, with banning and moderation, these types of content moved over to codebases allowed on GitHub as open-source contributions (Winter & Salter, 2019). With open-source availability and technology, "designer porn" became a reality. Forums began to populate and currently exist, such as MrDeepFake Forums, where people can anonymously request, pay for, and provide clear instructions on the types of detailed deepfakes they would like made (Kikerpill, 2020). Often, these forums and requests entail the pairing of a celebrity with a recognized pornographic performer to fabricate an image or scenario that does not exist in reality (Kikerpill, 2020). Given the escalating prevalence of deepfake videos in the digital sphere, a pressing demand emerges for expanded research initiatives aimed at comprehensively decoding the current and prospective societal ramifications of digital counterfeit content.

2.4 Harms of Deepfakes to the Public

The accessibility of emerging technologies highlights the adoption of deepfakes and their amplified usage in the digital landscape. The superimposition and swapping of faces of individuals into movies and television clips can be an entertaining example of deepfake videos. In the case of a clickbait video entitled "Keanu Reeves Stop A ROBBERY," the content included a stuntman and voice actor with his face replaced with that of celebrity actor Keanu Reeves (Bode, 2021). The video was disseminated across various social media platforms, but the content

was falsified, and Keanu Reeves was never involved. While deepfake content may initially captivate with its entertainment value, the simplicity of its creation highlights a notable absence of accountability, thus blurring the line between amusement and insidious propagation of misinformation. The efficacy of monitoring, enforcing accountability, and implementing regulatory measures against perpetrators is contingent upon legal and financial resources that are frequently disregarded (Langa, 2021).

As illustrated in Table 2.2, deepfakes can precipitate adverse outcomes and erode trust in news and electoral processes, particularly in the absence of adequate oversight (Diakopoulos & Johnson, 2019). In 2017, researchers skillfully synthesized deepfake stills of former President Barack Obama saying things he never actually said but voiced by actor Jordan Peele (Citron & Chesney, 2019). Demonstrating that the same technology could potentially engineer even more harmful deepfake content, inciting plans to carry out political assassinations, seemingly private conversations featuring elected leaders (Citron & Chesney, 2019), or Speaker Nancy Pelosi slurring her speech (Denham, 2020). With deepfakes being the next frontier for fraud, companies like Recorded Future have found multiple examples on the dark web of criminal activity using deepfakes to blackmail, create pornographic videos, and execute identity theft ("Security Firm," 2021). Tom Cruise was the latest viral victim of a series of high-quality, masterful, and convincing deepfakes shared throughout TikTok and YouTube (Bode et al., 2021). A collaborative effort between a Tom Cruise impersonator, Miles Fisher, and a Belgian visual effects artist brought forth skepticism and speculation about whether or not the videos were fake or if Tom Cruise was trolling TikTok viewers (Hern, 2021). Society and government entities need appropriate countermeasures when such actions impact personal liabilities, such as non-consenting individuals, government officials, and organizations (Kietzmann et al., 2020).

Table 2.2

Consequences of Harmful Deepfakes

Type of Impact	Harm Classification		
	Reputational Damage	Financial Consequences	Influence and Manipulations of Decision Making
Individual Impact	● Defamation ● Libel ● Intimidation ● Manipulation ● Career or family life	● Extortion ● Blackmail ● Scams	Opinions of public figures or well-known individuals (celebrities, influencers, leaders, and politicians)
Institutional Impact	● Brand ● Copyright Infringement ● Undermining trust ● Consumer interest	● Fraud ● Stock prices	● Tampering with evidence/jury ● Fake news ● Disrupting electoral campaigns ● Consumer trust and judgment of brands
Public Impact	● Domestic ● International or global ● Electoral processes ● Instilling community-wide fear, panic, or conflict	● Government spending ● GDP	● Geopolitical conflicts and interests ● Globally or nationally known brands, corporations, or spending on products

Note: Impacts of reputational, financial, and decision-making damages from deepfakes adapted from “Forged Authenticity: Governing Deepfake Risks” by A. Collins, 2019, EPFL International Risk Governance Center (IRGC).

Stakeholders play a pivotal role in delineating strategies to tackle legal and regulatory challenges associated with deepfakes. This is part of the comprehensive effort to address the

deficiencies within the current legal framework concerning deepfakes, which also includes further scrutiny and analysis of the First Amendment, general falsity, "intentional infliction of emotional distress and speech," defamation, false light, state legislatures, and statutory solutions (Gibson, 2020, p. 261). Concurrently, additional domains warrant legal examination, include Section 230 of the Communications Decency Act and Publishers' Hyperimmunity, tort law, copyright law, federal legislation, state non-consensual pornography statutes, expired federal legislation, and pending federal legislation (Gieseke, 2020). To these ends, developers and researchers also need to explore what the societal and research levels should enact to better support those being deepfaked without explicit consent and whether or not platforms, websites, and forums are at fault for not correctly vetting or moderating uploaded content (Karasavva & Noorbhai, 2021). To enhance comprehension of how bad actors are penalized, stakeholders must also administer regulatory standards and media literacy programs, and adequately prohibit non-consensual deepfakes (Gieseke, 2020).

2.5 Deepfake Detection and Discernment

Numerous factors influence an individual's susceptibility to believing deepfake video content. A growing body of literature recognizes the need to investigate the individual’s ability to detect and discern a video’s authenticity (Zhang et al., 2020). However, existing accounts fail to identify what characteristics, cues, and clues respondents refer to when trying to discern if a video is real or fake. Using science education videos and implementing deception in their study, Doss et al. conducted a study to assess the vulnerability of educators. They found that not only did 27-50% of the individuals not identify actual authentic videos from false ones, but all age populations in the study also exhibited some vulnerability to deepfakes, which increased with age and their trust in the deepfake content (2023). While short-term studies are progressing in

recognizing the factors that render individuals susceptible to deepfakes, Josephs et al. point out that research efforts typically fail to induce four conditions in deepfake detection and discernment studies that exist in normal internet browsing conditions: low prevalence, brief presentation, low video quality, and divided attention (Josephs et al., 2024). Under such circumstances, researchers have typically underestimated individuals' susceptibility to deepfakes and the ensuing impacts on human emotion and cognition. (Köbis et al., 2021). Accordingly, Table 2.3 delineates the variables, cues, and clues that ongoing research must consider in the end-user's discernment and detection process of deepfakes.

Table 2.3

Determinants of Deepfake Discernment on Social Media

Type of Misinformation	Content and Individual Characteristics		
	Characteristics of the individual information consumer	Characteristics of the information and content	Individual reasoning process
Falsified Videos or Deepfake	Demographic variables: age, gender, education level, political leanings, etc. Content and Technology Access: Mobile, PC, Social media, News, Console, Television, Tablet, Subscriptions, etc. Platform Literacy: Facebook, Twitch, Twitter, Reddit,	Removal or addition of a person from video (Langguth et al., 2021) Manipulation of background or artificial background (Langguth et al., 2021) Distorted pixels in image or video (Liao et al., 2021) Verification of source of content or distributor	Confidence in the truthfulness of content Confidence in judgment of truthfulness Accuracy of recall of content Accuracy of recall of truthfulness Confidence in the recollection of content

	Discord, Google, YouTube, etc. Cognitive traits: delusionality, dogmatism, religious fundamentalism, etc.	Credible URL, digital fingerprints, or hashtag discrepancies (Johansen, 2020) Unnatural coloring, body movement, facial-features, eye movement (Groh, n.d.) Detectable and imperfect artifacts in image patches (Chai et al., 2020)	Ability to verify credibility/source of content Confidence in verification of credibility/source Emotional response to content Change in opinion or incite action Confirmation bias (Moravec et al., 2019)

Social media companies and platforms operate under the assumption that users can independently discern misinformation online and subsequently rectify it (Ahmed, 2023). Yet, a case study revealed that deepfakes present what is termed “diverse misinformation” based on the complicated relationship between the varied human biases and demographics shown in deepfake videos (Lovato et al., 2024). Similarly, Lovato et al. employ deception in their study, wherein participants are not provided with information indicating potential exposure to deepfakes (2024). This approach aims to ascertain whether participants' biases influence their susceptibility to deepfakes and their ability to identify them accurately. The growing body of scholarly work on deepfakes is still broad, with academics working to better understand how biases shape one’s ability to detect deepfakes through reasoning and judgment processes, perceived confidence in them, and later perceptions of truthfulness given the fragility of human memory. The impact of these factors varies for each individual. Given the myriad techniques for fabricating false contexts based on diverse formats, detecting deepfakes and falsified videos on social media can pose considerable challenges, particularly when considering the disparities in human cognition and perception (Tahir et al., 2021).

2.6 Algorithm Detection: Technical Deepfake Detection Strategies

While some scholarly reviews of deepfake algorithms like FakeApp, Adobe VoCo, Lyrebird, and Face2Face have trended towards the ethically agnostic (Gardiner, 2019), other recent research has shifted towards the creation of methods to detect deepfake manipulations, with the assumption that deepfake videos pose social and moral dangers to the general public. A significant concern regarding high-fidelity deepfakes is that these videos are often imperceptible to the human eye, necessitating the search for algorithms or technologies that can accurately determine if videos were altered (Maksutov et al., 2020). Deep learning and machine learning play a significant role in the production of algorithmic attempts to detect deepfakes and, by extension, to challenge the persuasiveness of deepfake videos distributed online.

Image forensics research in the realm of deepfake videos has utilized neural networks (Güera & Delp, 2018; Amerini et al., 2019; Guarnera et al., 2020), classical frequency domain analysis (Durall et al., 2019), facial recognition (Korshunov & Marcel, 2019), and the algorithmic detection of visual artifacts that emerge during deepfake manipulation (Marra et al., 2018; Li et al., 2019). Additionally, Nguyen et al. surveyed algorithmic detection methods, including measuring blinking rates, pre-processing face alignments, extracting and scaling artifacts, analyzing deep classifiers, and reviewing shallow classifiers (2019). However, as pointed out by Nguyen et al. (2019), the same technologies used to detect deepfake videos are the same as those used to create deepfake videos, perpetuating an arms race reliant on the co-advancement of shared algorithms. Other researchers have identified additional algorithm detection methods, such as AI-based video synthesis algorithms, resampling detection, and GAN-generated image/video detection; however, these algorithms have numerous limitations in a society where mobile cameras and apps are constantly becoming more accessible (Li & Lyu,

2018). Therefore, the need for improved detection methods is pressing, as today's deepfakes are created with minimal coding knowledge and require no elaborate hardware.

As deepfakes proliferate on social media platforms and mobile networks, advanced computer algorithms and apps can swiftly generate, produce, and edit videos that are indistinguishable from the original content (Skibba, 2020). This trend, coupled with the use of generated adversarial networks (GANs) for rapid dissemination, creates vulnerabilities for face recognition software (Korshunov & Marcel, 2018). With facial recognition becoming less reliable in these circumstances, this technology poses a new threat to the foundation of trust online, as people may be unable to discern whether these videos are real or not. The potential consequences are far-reaching, prompting institutions like the Defense Advance Research Projects Agency (DARPA) to enhance their detection methods using forensics algorithms from the Media Forensics (MediFor) program (Greengard, 2019). While certain existing deepfake videos may appear benign, it is imperative to acknowledge the potential wide-ranging consequences of this technology on citizenry, social welfare, business, and daily life.

2.7 Human Detection: Can Human Beings Discern Deepfakes?

Research on human (as opposed to machine) deepfake discernment has more strongly figured around issues of literacy (Nightingale et al., 2017; Schetinger et al., 2017) and the social outcomes of deepfake circulation across social media platforms (Vaccari et al., 2020). Specifically, Kasra et al.'s study demonstrated that 19 participants split amongst four focus groups performed poorly at identifying fake online images in their research and were minimally confident in their ability to identify falsified images (2018). Given that the creation of deepfake forensic algorithms consequently advances the ability of deepfake video creators to evade algorithmic detection further, there is an urgent need to cultivate and promote the human

discernment of deepfake manipulations. Moreover, the circumstances in which individuals encounter deepfake videos unwittingly differ dramatically from the large, high-resolution datasets from which algorithmic detection models learn. For instance, in Durall et al.'s work, as image resolution for deepfake videos fell, so did their algorithm's predictive accuracy, suggesting that we cannot entirely rely on machines to detect manipulations in low-resolution videos correctly (2019). Marra et al.'s observations expand on how the circulation of deepfake videos occurs primarily via social media channels through the means of compressing images and videos (2018). In doing so, compression of audiovisual content often obfuscates how the algorithmic detection "signals" or flags falsified content due to the low-resolution, artifactual "noise." It becomes increasingly evident that the need for human discernment further increases, mainly when access to algorithmic detection models is uneven, impractical, and inequitable (Marra et al., 2018).

Shifting to the demand for deepfake identification and discrimination, researchers have initiated the aggregation of datasets and databases to facilitate algorithm training for detection. Simultaneously, efforts persist in evaluating individuals' capacity to discern deepfakes. Specifically, companies like Google, Facebook, YouTube, TikTok, and others have explored the effectiveness of deepfakes in deceiving individuals (Korshunov & Marcel, 2020). In one study, individuals reviewed 60 images and photographs, of which 30 were fake, and 80% of the human subjects successfully identified the fake images (Rössler et al., 2019). An example of this is a study carried out by Korshunov and Marcel in which they investigated whether human subjects could identify deepfakes in a dataset of 120 videos and found that high-quality deepfakes easily convinced individuals that they were real (Korshunov and Marcel, 2020). An additional deepfake experiment further added that context and reasoning skills are necessary for individuals to

discern falsified videos. In a contrasting study, when prominent political figures were the subjects of deepfake videos, human observers surpassed algorithms in discernment accuracy (Groh et al., 2021). Conversely, videos lacking contextual information surrounding the individuals making contentious statements yielded a reduced rate of discernment. (Groh et al., 2021). One notable discovery indicated that experiments specifically investigating audio deepfakes demonstrated that younger participants yielded higher proficiency in identifying false audio deepfakes compared to older participants (Müller et al., 2021). As individuals increasingly rely on voice assistant technology or corporations continue to integrate accessibility tools into hardware and software, the notion that distinct categories of deepfakes may target specific demographic groups raises significant concern. Overall, human discernment deepfake experiments demonstrate that critical thinking, reasoning, and truth discernment (Groh et al., 2021) vary depending on the context, subject, or content of the deepfake videos.

As research delves into the efficacy of deepfake detection and discernment, the literature demonstrates that bias is often a limiting factor for human discernment. Specifically, people tend to utilize mental shortcuts (Pennycook & Rand, 2021) when making judgments about content, overestimate their capabilities to discern or detect AI-generated content (Köbis & Mossink, 2021), or have differing views if the modality of the information varies from text to video (Graber, 1990). For instance, general research trends indicate that people cannot reliably identify or discern deepfakes, which means future experiments need to assess if deepfakes can implicate an individual's trust, values, or beliefs (Köbis et al., 2021). Finally, Dobber et al. analyzed the effects of political deepfakes on attitudes and found that participants' attitudes were significantly more hostile towards the politician after watching the content, suggesting that targeted deepfake campaigns exacerbate the effects on beliefs within small subgroups (2020). Despite the

prevalence of deepfakes, “fake news,” and general formats of misinformation, there is a lack of empirical research examining the long-term and short-term psychological consequences on individuals (Hancock & Bailenson, 2021).

2.8 How False Memories are Implanted

Due to the rising prevalence of deepfakes and their ability to influence critical societal functions, it is essential to research the phenomenon of falsified content and whether deepfakes affect memories. The existing body of literature on memories continues to demonstrate that our ability to recall and retain information is not as steadfast as commonly believed (Loftus & Hoffman, 1989). Our memories can be highly reliable at times, while at others, they can mislead us or be entirely false, blurring the line between reality and imagination (Loftus & Berstein, 2005). Newman and Lindsay define false memory as "a wide variety of memory errors ranging from misremembered word lists to erroneous reports of details in stories to false memories of dramatic life events" (2009, p. 1106). A systematic literature search analyzes and lists the following definitions by Muschalla and Schönborn, identifying the standard various methods and formats of clinical and experimental laboratory research (2021):

- Induction of false memories - processing or bringing about false memories that did not exist or occur before a study
- Imagination inflation - the repeated imagination of events that did not happen
- False feedback - conveying suggestive misinformation through speech
- Memory implantation - utilizing manipulated photographs or false statements through relevant or close relationships with the individual

The goals of decades of complex academic research on memories have ranged from trying to understand the role of hypnosis to social pre-conditioning of emotions and beliefs

(Gow, 1998). In recent years, studies have increasingly focused on how external influences, like media manipulation and digital fabrication, shape memory formation. This analysis further explores the correlation between false beliefs and false memories. In broad terms, the causal role of deepfakes in forming false memories reinforces the adaptive and flexible nature of the memory system, acknowledging that various variables influence how we process and store information. (Newman & Lindsay, 2009). An integral aspect of comprehending the potential for audiovisual content to implant false memories involves the methods utilized by researchers to assess attitudes, beliefs, and judgment. Furthermore, a significant portion of understanding the mechanisms underlying true and false memory processes hinges on the emotional aspects of how individuals recall or depict actual events.

On the other hand, Mather reveals that emotional representations of events can be more memorable than those without emotional context (2009). The distinction is further exemplified by video content eliciting emotions such as despair, anxiety, or stress, which may enhance the recollection of events but also render individuals more prone to errors in recalling events, specifically in court cases (Deffenbacher et al., 2004). In many instances, false and genuine memories are not simply remembered or manipulated solely based on heightened emotion; the content or the specifics of the events recollected can matter. For example, in another study, Thomas and Loftus surveyed how unusual circumstances can be implanted or constructed as false memories by measuring imagination (2002). Finally, defining the extraordinary events as bizarre, accurate, and false memories, the effort determined that after repeated imagination, participants claimed that they had performed strange actions that they had never performed before (Thomas & Loftus, 2002).

These experiments consistently yield robust evidence suggesting that it is possible for individuals to develop false memories and that people are vulnerable to having false memories implanted in their thoughts. This is evident in the case of Loftus' shopping mall experiment, which tested the effect of misinformation in the 1970s (1999). The experiment involved a subject's relative using deception and constructing a false memory of the subject being lost in a large department store or shopping mall (Loftus & Pickrell, 1995). The results revealed that, under certain conditions, it is possible to lead people to recall completely false events that never happened; however, critics questioned the validity of the study, sparking a discussion about the ethics of studying false memories and the potential for contaminating the memories of human subjects (Loftus & Pickrell, 1995). While these findings illuminate the malleability of memory and the potential for false memories to be implanted, the ongoing scrutiny surrounding the ethical considerations of such research underscores the need for meticulous care in the study of human and false memory processes.

2.8.1 False Memories Created Through Imagery

Scholars have ascertained that with the inclusion of photographs alongside narratives, content, or news articles, individuals are more likely to recall them, whether or not the content is accurate (Garry & Gerrie, 2005). As an example, researchers realized after 9/11 that shocking, traumatic events could produce some of the most vivid and clear recollections among individuals, noted as "flashbulb" memories (Yong, 2021). However, as media, photojournalism, and social media increasingly blurred the distinction between factual information and misinformation, the prevalence of viral headlines paired with photographs distorted the events' narrative, complicating the public's ability to discern accurate accounts from fabricated ones (Klomp, 2020). This idea further complicates research behind memory work as emerging

technologies, media, and misinformation techniques progress over the decades. Comprehending how individuals process information adds complexity to the task of fully grasping the ethical concerns involved in interpreting information. Furthermore, research has demonstrated that the presentation of falsified images alone does not significantly impact the distortion of memory, but combining doctored images with an emotion-evoking image can lead to the implanting false memories (Lindsay et al., 2004).

Researchers investigating false memories tend to concentrate on inducing these memories through the use of fabricated narratives, subtly altered videos, and falsified images. Nonetheless, the advent of deepfake technology presents a potentially more significant challenge, suggesting that the implantation of false memories may extend beyond the scope of traditional doctored photographs and videos. The rapid advancement of deepfakes exacerbates their threat to democracy in artificial intelligence, neural networks, and machine learning. One significant drawback of this subject matter is the limitations involved in developing and applying deepfake videos. The process is time-consuming and requires adequate computer processing power to render the images frame by frame (Karnouskos, 2020). Nevertheless, if a technologically savvy user is motivated, has access to powerful computing processors, and has copious amounts of time, they will have minimal issues creating deepfakes (Keitzmann et al., 2020). Falsified multimodal content, such as images, photos, and videos, can span from harmless memes to satirical news articles produced through artificial intelligence, completely bypassing journalistic control (Sanchez-Castillo & Lopez-Olano, 2021).

The possibility that deepfake videos could contribute to forming false memories adds to existing concerns about the challenges in detecting fake visual information and the complexities involved in addressing accountability for those who misuse such technology. As psychologists

examine why and how memories fail, modern technology and internet platforms continue disseminating falsified content, specifically visual content such as photographs and deepfakes (Loftus & Pickrell, 1995). Photos can illustrate events, capture evidence, and amplify memory recall. In contrast, several notable researchers conducted a study in which participants were exposed to a falsified image depicting a childhood hot air balloon ride. In this study, participants overwhelmingly demonstrated the ability to identify the event as fictitious accurately (Garry & Wade, 2005). The conclusive findings of the research revealed that narratives provide false memories or false partial memories and that images, photographs, and visual media can also produce similar consequences. Although falsified photographs may not exert a more decisive influence than fabricated narratives when presented alongside a series of unaltered images from a reliable source, there remains a significant risk of source confusion, highlighting the need for more effective detection and prevention strategies (Garry & Gerrie, 2005).

2.8.2 False Memories and Existing Deepfake Research

While research on the correlation between deepfakes and false memories remains limited, several studies have begun to investigate this relationship. In one instance, Vaccari and Chadwick investigate the role of deepfake videos on people's perceptions and trust in news on social media (2020). The investigation utilized three variations of a BuzzFeed Obama/Peele deepfake video (Good Morning America, 2018): two deepfake versions and one educational version. The hypotheses tested included:

1. Individuals who watch a deepfake video containing a false statement that is not revealed as false are more likely to be deceived.
2. Individuals who watch a deepfake video containing a false statement that is not revealed as false are more likely to experience uncertainty about its content.

3. Exposure to a deepfake video through social media, resulting in uncertainty about its content, may reduce trust in news on social media.

A random sample of 2,005 British participants was assigned to watch one of the three versions of the deepfake video. The results measured subjects' evaluations of the truthfulness of a deepfake video and their trust in news on social media. The research encompassed a direct question about a highly improbable statement in the deepfake video to determine if participants believed it. They also regarded "Don't know" responses as markers of uncertainty. Participants were also asked about their trust in news on social media to gauge their attitudes. The study revealed that approximately half of the participants were not misled by the deepfake content. A smaller proportion of the participants were deceived, while about a third expressed uncertainty, indicating a mixed response to the authenticity of the deepfakes. The responses varied based on the treatment participants received, providing an understanding of the impact of deepfake videos on people's perceptions and trust in news on social media.

Nevertheless, the analysis indicated that the differences in participant responses between the two deceptive videos and the full video accompanied by an educational reveal were not statistically significant. Overall, the research rejected the hypothesis that individuals who watch deepfake political videos containing a false statement are more likely to believe the false statement if it is not revealed as false. The investigation demonstrated that exposure to deceptive deepfake videos increases uncertainty about the content of the video, but it does not necessarily lead to higher levels of deception. Moreover, elevated levels of uncertainty were correlated with low trust in news, suggesting that deepfake videos could undermine people's confidence in news content on social media. This effect is particularly pronounced when individuals are unaware that the video might be a deepfake. The 'treatment' participants received refers to the specific version

of the deepfake video they were assigned to watch, whether it was one of the deceptive versions or the full video with an educational reveal.

In a related study, Dr. Gillian Murphy and Emma Flynn conducted two experiments in which participants were presented with true and false news events utilizing text, photographs, and deepfake videos. The primary aim of this research was to determine whether exposure to deepfake videos would increase false memory rates and to assess the effectiveness of a post-exposure warning in aiding participants in detecting fake stories (2021). In the first experiment, the authors addressed three questions:

1. Does viewing a fake news story as a deepfake video make a false memory for that story more likely, relative to other formats?
2. Even after being warned about the possible presence of fake news, can a deepfake reduce a participant's ability to detect a fake story?
3. How do participants feel about deepfake videos – do they perceive them as dangerous?

Participants were recruited from university student emails and social media platforms to study "memory for recent news events." The initial sample consisted of 377 participants, 70% of whom were female, with an overall age range of 18 to 59 years ($M = 24.87$, $SD = 6.90$). Most participants (84%) had completed at least some college education. All participants viewed all six news stories once in one of the three available formats (text, image, or video). The researchers used existing deepfake video clips created by social media users, as they did not have the resources to create their own library of deepfake videos. For each story, separate analyses were conducted to compare memory retention rates among participants who viewed the video in one of the three different formats. Overall, participants predominantly held negative views of deepfake videos, characterizing them as dangerous and unethical. The majority of participants

also endorsed the regulation of deepfake videos and expressed confidence that their participation in the study would make them less susceptible to such content in the future. In the second experiment, the authors addressed three additional research questions:

1. Are better-quality deepfake videos rated as more convincing by participants?
2. Does viewing a fake news story as a deepfake video make a false memory for that story more likely, relative to other formats?
3. After being warned about the possible presence of fake news, can a deepfake video reduce participants' ability to detect a fake story?

All participants were recruited through Prolific, an online research platform, for a study on "memory for news concerning Kim Kardashian" (Murphy & Flynn, 2021). The final sample consisted of 305 participants aged 18 to 69 years (M = 33.33, SD = 10.49), and 69% were female. Most participants (84%) had completed at least some college education. All participants viewed the same three news stories (two fake and one genuine) once in one of the three available formats (text, text with an image, or text with a video). The two fake stories concerned a fictitious video of Kardashian rapping about her wealth and comments made by Kardashian about her online marketing. The true story concerned a visit made by Kardashian to the White House. After viewing all stories, participants were asked if they remembered each event and could respond (a) "I specifically remember seeing/hearing about this event," (b) "I don't have a specific memory of seeing/hearing about this event, but I remember it is happening," (c) "I don't remember this event, but I believe it happened," (d) "I remember this event differently" or (e) "I do not remember this event." As pre-registered, (a) and (b) were classed as memories, (c) was classed as a belief, and (d) and (e) were classed as not remembering.

Participants were also asked to rate the truthfulness of the stories on a scale from 0 (Definitely Fake) to 100 (Definitely True) and to rate the quality of each deepfake video on a scale from 1 (Not at all) to 100 (Extremely). The results indicated that the higher-quality deepfake video was rated as more convincing and more likely to persuade others compared to the lower-quality deepfake video. However, neither deepfake video increased the false memory rate relative to other formats. The presentation format did not affect participants' assessment of the stories after they were warned about possible fake news. The study found that even high-quality deepfakes may not necessarily increase false memory rates. These findings suggest that the impact of deepfake videos on false memory rates can depend on various factors, including the story's plausibility and the quality of the video.

In the final deepfakes and false memory study, titled "Face/Off: Changing the Face of Movies with Deepfakes," Murphy et al. investigate the consequences of deepfake technology on the film industry (2023). The study involved 436 participants who were presented with real and fake movie remakes and asked to rate their memories and opinions about them. The participatory quantitative research methods recruited undergraduate students to be included as participants in a class project. 570 participants completed the survey; however, 20 participants reported that they searched the internet for help to answer the questions and were removed. An additional 111 participants were removed for not watching the videos, and three were removed for being under 18. This resulted in a final sample of 436 participants, most of whom were female (n = 253), with an average age of 25. Participants were shown both authentic and artificial (deepfake) movie remakes and asked to evaluate their recollections and opinions about the remakes and their opinions on the use of deepfake technology as a substitute for actors in films.

The qualitative analysis revealed six main themes: positive use cases for the technology, its impact on actors, respect for filmmakers' artistic vision, concerns about undermining communal movie experiences, potential legal and social harms, and the negative effects of too many choices. Participants were particularly concerned about the impact on actors, manipulation, misinformation, and copyright issues. The researchers found that participants readily formed false memories of fictitious movies presented as deepfakes. As an example, Captain Marvel was most frequently falsely recalled (73%), followed by Indiana Jones (43%), The Matrix (42%), and The Shining (40%). Interestingly, presenting authentic movies in the format of a deepfake did not significantly increase the false memory rate. The research indicates that although some participants recognize potential advantages in employing deepfake technology in films, many expressed concerns about its negative implications. These concerns center on the authenticity of performances, the integrity of the filmmakers' artistic vision, and the erosion of the communal social experience of watching movies.

2.9 Media Literacies Among Young Adults to Counter Deepfakes

The persistent widespread belief in misinformation in the digital age presents an ongoing challenge. While research to date has mainly covered text-based misinformation such as news articles, headlines, or social media clickbait, there is a growing need for media literacy education to combat deepfake video content and increase discernment of the quality of information presented online (Guess et al., 2020). Potter (2014) demonstrates that nearly all media literacy interventions share four essential characteristics: the agent (the individual or organization delivering the intervention), the target (the audience for whom the intervention is designed), the treatment (the specific content or method of the intervention), and the expected outcome (the measurable evidence of the intervention's effectiveness). However, the rapid proliferation and

evolving nature of deepfakes present unique challenges to this framework. For example, identifying qualified agents with expertise in deepfake detection, tailoring interventions to targets with varying levels of technological literacy, and designing treatments that address the nuanced psychological effects of deepfakes are critical considerations. Finally, the lack of a standardized toolkit underscores the urgency for media literacy interventions that specifically address deepfake-related risks and their potential to manipulate memory and perception.

2.9.1 The Digital Native Dilemma: Benefits and Drawbacks for Gen Z

Gen Z, defined as individuals born between the mid-1990s and early 2010s (now aged approximately 12–27), is maturing in an epoch defined by rapid technological advancements, particularly in social media and content distribution (García-Marín, 2021). Their general comfortability and connectedness to social media platforms, including Instagram, Twitter, TikTok, and Snapchat, have become pervasive in their daily cognitive and behavioral routines, serving as primary channels for communication, entertainment, and information sharing (Bourke, 2019). The extent of social media use among Gen Z is substantial, with Morning Consult's 2022 study surveying 1,002 respondents indicating that a significant proportion of this cohort spends multiple hours daily (He et al., 2024). The prevalence of this engagement is causally linked to the enhanced availability and adoption of smartphones, hardware, and devices, with Gen Z'ers using their smartphones for everything from making phone calls to completing class assignments (Ahmed, 2019). Moreover, this demographic leverages social media to enter the workforce equipped with pre-established ideologies (Desai & Lele, 2017), seeking jobs primarily through social networks (Peter et al., 2020), and building personal, monetizable brands online (Vițelar, 2019). As a cohort, Gen Z is distinguished by their heightened technological proficiency compared to preceding generations and their propensity for mobile addiction stemming from

their reliance on smartphones (Ozkan & Solmaz, 2015). However, despite their extensive technological familiarity, Gen Z end-users may not necessarily have an enhanced ability to detect falsified content. Empirical research is required to investigate whether they are as susceptible, or perhaps more so, to deepfake videos.

Contributing factors to Gen Z's social media proficiency include the use of social media by consumer brands, organizations, and institutions to engage with Gen Z. These entities utilize platforms to meet Gen Z's communication habits and expectations, aiming for higher online engagement. This includes interacting with Gen Z, as per Luoma-aho (2020), and targeting advertisements to them, as Grigoreva et al. (2021) noted. Additional findings demonstrate that Gen Z's susceptibility to misinformation is nuanced; Grigoreva et al. find that Gen Z's digital literacy makes them "highly informed, more pragmatic, and analytical" than those from previous generations (Grigoreva et al., 2021, p. 164). Given the substantial volume and sophisticated nature of falsified content, particularly deepfakes, Panagiotou et al. (2022) analyze Gen Z's adoption of media literacy. The study found that Gen Z lacks sufficient media and news literacy competencies, as evidenced by the fact that 58.3% of survey participants reported having been influenced by disinformation.

Moreover, a study by Pérez-Escoda et al. (2021) examining Gen Z social media users' exposure to online misinformation found that the participants did not necessarily lack media literacy skills. Instead, the issue lay in the inadequacy of traditional media literacy models, which proved ineffective in addressing the challenges in the new online era, thus highlighting the need for media literacy education, particularly to discern and verify the source of information on social media (Veeriah, 2021). Developing the skills necessary to evaluate information on social media categorically presents significant challenges amidst the proliferation of new digital

platforms. Nonetheless, formulating and implementing robust media literacy strategies in both formal and informal settings (Silveira, 2020) is crucial. This is pertinent given the current gap in research regarding the effectiveness of these strategies in combating the formation of false memories induced by online misinformation. Although Gen Z may possess moderate media literacy, researchers have not yet determined whether this cohort can responsibly navigate the complex landscape of deepfakes.

As far as online misinformation is concerned, deepfakes have become a leading global concern (O'Connor & Weatherall, 2019). Journalists and public scholars highlight the potential risk that deepfake technology presents to both public understanding and the functioning of democratic institutions. (Mihailidis & Viotty, 2017). Social media sites like Facebook face the most complex challenges, as they are responsible for balancing freedom of speech for their users with a need for content moderation and media literacy education to ensure the platform does not harm (Guo & Johnson, 2020). Media literacy, defined as the ability to evaluate, analyze, and respond to various forms of communication (Schwarz, 2005), includes both text-based and non-text media.

Research has also explored how media literacy education, alongside active debriefing techniques, can effectively mitigate the impact of misinformation, including deepfakes (Hwang et al., 2021). Active debriefing not only helps participants process their emotional responses to manipulated content but also reinforces their ability to identify the markers of misinformation. By engaging participants in reflective discussions and providing them with corrective feedback, active debriefing can strengthen critical reasoning skills and reduce the likelihood of false beliefs taking root. Moreover, active debriefing allows for real-time clarification of misconceptions, false news stories, and misinformation (Greene & Murphy, 2023) while providing a structured

opportunity to link specific media literacy strategies to observable outcomes. Recent efforts in media literacy and the examination of misinformation's discernment and impact have concentrated on three primary factors contributing to increased susceptibility to false claims: the characteristics of the individual information consumer (Chen, et. al., 2011), the attributes of the information itself (Caled & Silva, 2022), and the individual's critical thinking and reasoning processes (Qiu, 2024). Integrating active debriefing into these efforts may offer a promising avenue for addressing the unique cognitive and possible emotional challenges posed by deepfakes.

In terms of individual characteristics, demographic variables such as age, education level, and political leaning (Guess, Nyhan, et al., 2020; Grinberg, 2019) affect individual susceptibility to misinformation, as do cognitive traits such as delusionality, dogmatism, religious fundamentalism (Bronstein et al., 2019), and "bullshit receptivity" (Pennycook & Rand, 2019b, p. 3). Research also shows that prior exposure to a given false claim (Pennycook & Rand, 2021; Wickens, 2002) has a negative effect while prior knowledge about the content of the information itself (Karmarkar & Tormala, 2010), political knowledge (Vegetti & Mancosu, 2020), media literacy (Amazeen & Bucy, 2019), information literacy (Jones-Jang et al., 2019), and basic science knowledge (Pennycook et al., 2020) have a positive one. Ideological alignment between misinformation and one's pre-existing beliefs also plays a key role (Drummond et al., 2020; Faragó et al., 2020; Kahan, 2013; Van Bavel & Pereira, 2018; Vegetti & Mancosu, 2020).

With regards to characteristics of the information, source credibility (Pehlivanoglu et al., 2019; Pornpitakpan, 2004), the status of the information source (i.e., elite messaging) (Zaller, 1992), social feedback (such as the number of "likes") (Avram et al., 2020) and emotional valence (Crocket, 2017; Kozyreva et al., 2020; Martel et al., 2020) all have a significant

influence on individual discernment of the veracity of information online. In terms of individual reasoning processes, analytical thinking (Pehlivanoglu et al., 2021; Pennycook & Rand, 2019a), deliberation (Bago et al., 2020), and focusing attention on accuracy explicitly (Pennycook et al., 2021) all affect individuals' evaluation of the veracity of information.

2.9.2 Media Literacy: A Viable Intervention for Deepfakes?

A substantial body of case studies and research has implemented media literacy interventions or debriefing treatments in deepfake studies involving deception and human detection. In one quasi-experimental design, one group received an in-depth lecture on misinformation, a second group was actively debriefed and trained on deepfake technology, and a third group received no debrief on media literacy (Mokadem, 2023). Mokadem's findings revealed that media literacy lectures significantly reduced the sharing of false information (2023). The study highlights the importance of integrating comprehensive media literacy training that encompasses general insights into misinformation and detailed education on deepfake technology. This dual approach is recommended to effectively minimize the likelihood of distributing fabricated content. In a separate inquiry, Luitse completed a textual analysis of notable YouTube videos relating to deepfakes and confirmed that users with digital literacy skills were effective at discerning deepfakes and their harmful potential (2019). Subsequently, another experiment explored the efficacy of active debriefing and media literacy training in enabling participants to accurately identify inauthentic information (Dame Adjin-Tettey, 2022). Although the assessment did not specifically examine deepfakes, the results demonstrated that individuals who received media literacy training and instruction were proficient in distinguishing between accurate and inaccurate information.

Despite the current variety of media literacy programs and initiatives, opportunities still exist for validated interventions promoting deepfake literacy. A recurrent challenge in existing programs is the need to expand media literacy strategies beyond traditional text-based or written sources of information. Effective interventions must also find ways to educate individuals to critically evaluate images, audio, and audiovisual media content (Fogt, 2021). To elaborate on this assertion, Hwang et al. conducted a multi-modal experiment in which a subset of participants were informed about the prevalence of deepfake misinformation (2021). The results revealed that active debriefing of deepfakes led to increased levels of correctly diagnosing the "vividness, persuasiveness, credibility, and intent" behind the dissemination of information (Hwang et al., 2021, p. 188). The results then demonstrate that media literacy mitigated the effects of deepfake content. Recommendations currently set encompass a spectrum of literacies, from establishing fact-checking organizations (Garriga et al., 2024) to developing dedicated websites, such as GitHub repositories and YouTube accounts, teaching end-users how accurately individuals can discern deepfakes addressing AI and data harms (McCosker, 2024). As a whole, even though there is not a one-size-fits-all intervention method, research overwhelmingly supports the notion that education and training are effective in enhancing deepfake awareness and detection (Tahir et al., 2021); however, further research is required to determine the extent to which individuals can accurately detect and discern deepfakes.

2.10 Positioning This Study Within the Deepfake Discourse

While machine learning has been leveraged to automate deepfake detection, there remain substantial gaps in understanding individual susceptibility to deepfakes and the potential vulnerability to false memories. Nonetheless, there is optimism within society regarding the potential for advancements in detection methods to outpace the development of malicious

deepfake content, thus contributing to an ongoing information security landscape. Given this context, my dissertation aims to investigate the perceptions of deepfakes among the Gen Z demographic, exploring the potential for deepfakes to induce false memories. By focusing on this specific population, particularly susceptible to large volumes of digital information, this research endeavors to yield valuable insights into the vulnerabilities posed by deepfakes and their potential societal ramifications. This underscores the importance of approaching the issue of deepfakes from both sociological and cognitive perspectives, in addition to technological considerations, to fully comprehend and address the ramifications of deepfakes.

Chapter 3

Study Design and Methodology

The purpose of this chapter is to introduce the research methodology for this experimental study. It will determine whether deepfakes can implant false memories, and if media literacy interventions are employed, can mitigate such false memories being implanted. The study utilizes a deception-based approach, presenting participants with a cover story about the survey's purpose before revealing its true intent. Data is collected through two quantitative surveys—an initial survey and a delayed post-test survey—administered to a diverse sample of Gen Z adults. Additionally, the chapter includes an in-depth discussion of the survey instruments, including questions related to participants' social media consumption, susceptibility to false memory implantation, and the participants' ability to discern between authentic and manipulated videos. Finally, it details the research design, participant recruitment strategies, data collection procedures, analytical methods, ethical considerations, and potential limitations of the study.

3.1 Research Questions and Instruments

These research questions guide the study by exploring whether exposure to deepfakes can implant false memories, assessing the susceptibility of a specific participant cohort to such deepfake influences, and examining key variables that affect the ability to discern deepfakes. To these ends, this study seeks to answer the following research questions, employing the variables and instruments (see Table 3.4).

Table 3.4

Overview of Research Questions, Variables, and Instruments

Research Questions	Variable(s) and Type	Instrument
RQ1. Can as demonstrated among the Gen Z adult participants in this research.exposure to a deepfake video (Kim Kardashian or Elon Musk) implant false memories?	IV: Exposure to Deepfake (Kim/Elon), No Exposure (Control) DV: False Memory Formation (Binary)	Deepfakes: Two Fake Videos Two Real Videos (Independent) One Real Video (Control) False Memories (FM): To be the best of your knowledge, mark all of the claims that are true about Elon M./Kim K./Paris Hilton
RQ2. Can an active debriefing (informing participants that the video is a deepfake) deter the implanting of false memories?	IV: Exposure to Deepfake (Kim/Elon), Active Debrief or No Debrief DV: False Memory Formation (Binary)	The real purpose of this study is to examine the correlation between deepfakes and false memories. Some people who participated in this study may have been shown deepfakes (videos that were entirely fabricated and may have involved doctored images or videos). Below, please rate whether you think the social media clips you saw were true or false.
RQ3. What is the relationship between prior knowledge and the likelihood of a deepfake video (featuring Kim Kardashian or Elon Musk) implanting false memories?	IV: Exposure to Deepfake (Kim/Elon), Prior Knowledge of Celebrity DV: False Memory Formation (Binary)	How knowledgeable are you about the celebrity Kim Kardashian? 1 - Not Knowledgeable at all 2 - Slightly Knowledgeable 3 - Moderately Knowledgeable 4 - Very Knowledgeable 5 - Extremely Knowledgeable
RQ4. What is the relationship between prior knowledge of the video content (specifically, the celebrity) and the effectiveness of active debriefing?	IV: Prior Knowledge of Celebrity, Active Debrief DV: False Memory Formation (Binary)	How knowledgeable are you about the celebrity Elon Musk? 1 - Not Knowledgeable at all 2 - Slightly Knowledgeable 3 - Moderately Knowledgeable 4 - Very Knowledgeable 5 - Extremely Knowledgeable

		How knowledgeable are you about the celebrity Paris Hilton? 1 - Not Knowledgeable at all 2 - Slightly Knowledgeable 3 - Moderately Knowledgeable 4 - Very Knowledgeable 5 - Extremely Knowledgeable
RQ5. Is there a relationship between social media use frequency and the effects of deepfakes in implanting false memories?	IV: Deepfake Exposure (Kim/Elon), Social Media Use DV: False Memory Formation (Binary)	How many hours, on average, do you spend on social media a day (not including messaging apps)? [Open text box]
RQ6. Is there a relationship between prior knowledge of deepfakes or previous exposure to deepfakes and the likelihood of false memories being implanted?	IV: Deepfake Exposure (Kim/Elon), Prior Knowledge of Deepfakes DV: False Memory Formation (Binary)	Prior to this study, did you consider yourself knowledgeable about deepfakes? 1(Not at All) to 10(Extremely) Have you looked at deepfake videos online before? a. Yes b. No c. I'm not sure d. I didn't know what a deepfake was
RQ7. How accurately do people discern whether a video is real or a deepfake?	IV: Video Type (Categorical: Real vs. Deepfake) DV: Accuracy of Discernment (Binary)	Is this video real or fake? Real or Fake.
RQ 8. How confident are they in their discernment?	IV: Accuracy of Discernment DC: Confidence	How confident are you in your answer? (0 - Not Confident At All to 10 - Extremely Confident)
RQ9. What cues and clues do individuals use for discernment?	IV: Video Type (Categorical: Real vs. Deepfake) DV: Cues and Clues Identified (Categorical)	What cues and clues (i.e. features) lead you to believe this video is real or fake? [Open text box]
RQ10. Are some cues and clues more or less efficacious for discernment than others?	IV: Discernment Cues and Clues used (Categorical) DV: Accuracy of Discernment (Binary)	(Same as above)

RQ11. How can exposure to information about deepfakes influence individuals' attitudes and perceptions?	IV: None (Exploratory) DV: Attitudes and Perceptions Toward Deepfakes	After completing this survey, how do you feel about deepfake videos? [Open text box]

To verify both content and criterion validity, the majority of survey questions were developed from previously validated instrumentation from similar studies. These questions assessed participants' pre-existing knowledge of celebrities and deepfakes, thereby gauging overall receptivity to misinformation (Zmigrod et al., 2023). For instance, with regard to the video content, participants were asked, "How knowledgeable are you about the celebrity Paris Hilton/Kim Kardashian/Elon Musk?" using a 1-5 Likert scale (Joshi et al., 2015), where 1 signified "not knowledgeable" and 5 denoted "extremely knowledgeable." Later, participants were queried on their self-assessed expertise in deepfake technology prior to the study, using a sliding scale ranging from 0 ("Not Knowledgeable At All") to 10 ("Extremely Knowledgeable"). Participants were then asked a subsequent question, replicated from Murphy and Flynn's study (2022), asking if they had ever looked at deepfakes before (e.g., "Yes," "No," "I'm not sure," and "I didn't know what a deepfake was").

Participants watched three videos and indicated whether they believed each was real or fake. They also rated their confidence in their judgments and provided open-ended explanations. This approach draws from Ternovski et al. (2021), who employed a 2x2 factorial design to examine the effects of debriefing on deepfake exposure. The discernment task used a modified version of Murphy and Flynn's (2022) instrument, asking: "Is this video real or fake?" Confidence levels were assessed through an adapted uncertainty measure from Vaccari and Chadwick's (2020) work on deepfakes and disinformation. By integrating these methodologies,

the instrument reliably measures whether participants can tell the difference between real and fake content.

To establish the inter-rater reliability of the open-text instruments in the active debriefing portion (McAlister et al., 2017), a structured coding process was implemented for two key questions: (1) "How do you feel about deepfakes after taking this survey?" and (2) "What cues and clues (i.e., features) lead you to believe this video is real or fake? Please be as specific as possible." As the primary researcher, I developed unique codebooks for each question to standardize the categorization and interpretation of responses. To assess the reliability of these codebooks, a second coder was tasked with independently coding a randomly selected sample of 50 responses per question using the predefined categories. The percentage agreement method was employed to measure the consistency between coders, resulting in a percentage agreement of 83.11% for the question "How do you feel about deepfakes after taking this survey?" and 89.69% for the question "What cues and clues (i.e., features) lead you to believe this video is real or fake? Please be as specific as possible." This level of agreement confirmed the reliability of the coding framework and allowed for further refinement to address any ambiguities.

For the false memory assessment, all participants—both those receiving active debriefing and those who did not—completed a delayed post-test survey between 24 and 96 hours after exposure to determine whether AI-generated videos could implant false memories (Pataranutaporn et al., 2024). The design of the false memory test and post-test survey conditions was informed by expert feedback from Dr. Loftus, as well as prior research by Abadie et al. (2024), Blandón-Gitlin and Gerkins (2010), and Wade et al. (2002). After viewing both authentic and deepfake videos, participants indicated which claims about the featured celebrities they believed to be true. Each video contained four genuine claims, two fabricated claims embedded

within the deepfake, and one additional random false claim. Participants who viewed the deepfake video and marked both fabricated claims as true were classified as having formed a false memory.

3.2 Research Approach

The purpose of this study is to examine how deepfake videos affect memory, discernment, and confidence among Gen Z adults. More specifically, the study investigates whether exposure to deepfakes can implant false memories, the efficacy of active debriefing in reducing such effects, and the role of prior knowledge and social media use in moderating these effects. To investigate these questions, the study implements a structured, quantitative experimental approach. Participants are exposed to deepfake videos, and their responses are analyzed to determine how false memories form and what factors may help mitigate them. This approach is designed to test the study's aims and objectives by enabling measurable, controlled comparisons between experimental conditions, helping to identify key factors that may influence deepfake susceptibility and false memories.

The primary aims of this study are:

- To explore whether exposure to deepfake videos results in false memory formation among Gen Z adults.
- To determine the effectiveness of active debriefing (informing participants about the deepfake nature of videos) in reducing false memory implantation.
- To investigate how prior knowledge (familiarity with celebrities, previous exposure to deepfakes, and social media usage frequency) moderates susceptibility to false memories.
- To assess participants' accuracy and confidence in discerning real versus deepfake videos and identify the most and least effective cues for detection.

- To explore participants' attitudes and perceptions toward deepfakes.

To achieve these aims, the study sets out the following objectives:

- Measure the incidence of false memories among participants after exposure to deepfake videos featuring known celebrities (Kim Kardashian or Elon Musk).
- Evaluate the impact of active debriefing on reducing false memory formation.
- Analyze how participants' prior knowledge (celebrity familiarity, awareness of deepfakes, and social media habits) influences the likelihood of false memory implantation and moderates the effectiveness of active debriefing.
- Investigate accuracy rates, confidence levels, and discernment cues used by participants in distinguishing real from deepfake videos.
- Examine participants' feelings, attitudes, and perceptions toward deepfake technology and its broader implications.

3.3 Data Collection

3.3.1 Sampling

The target population for this study consisted of adults aged 18–26, representing Gen Z. This age range was selected due to its relevance to the research questions and the potential influence of generational characteristics on the study variables. The study aimed to recruit between 600 and 2,400 participants born between 1997 and 2012. The target sample size was informed by prior research on false memories, particularly Dr. Gillian Murphy's studies on deepfakes and their effects on memory. Dr. Murphy's (2021) work explored how deepfake videos—highly realistic but fabricated footage—can distort memory for public events. In her study, 682 participants were exposed to fake news stories presented in three formats: text only, text with a photograph, and text accompanied by a deepfake video.

Building on this foundation, a pilot survey for the current study highlighted a significant dropout rate across the screening, initial survey, and posttest procedures. This pilot survey, conducted on Qualtrics, included over 60 participants, with at least 40.00% failing to complete the study or encountering technical problems. To address this attrition, the study set a recruitment goal of 2,400 participants to account for high dropout rates. Ultimately, the minimum target sample size was established at n=150 per subgroup, requiring a minimum of 600 validated participants to ensure sufficient statistical power to detect effects related to false memories.

3.3.2 Recruitment

The study's sampling pool consisted of undergraduate students from the University of California, Irvine (UCI). Before recruitment, the IRB approved specific undergraduate courses for participant selection, including 31 Informatics classes, 29 Information and Computer Science classes, 21 Game Design and Interactive Media classes, and 46 Psychology classes.

Following IRB guidelines, students who completed the full study and met specific survey criteria were eligible for extra credit determined by their professors. To ensure participation was voluntary and free from coercion, the study offered five pre-approved alternative extra credit options, developed in collaboration with and approved by the IRB. These alternatives provided equitable ways for students to earn the same amount of extra credit without feeling obligated to participate in the study. The options included:

- Writing a summary of a research article
- Creating a project or product development pitch
- Writing a research paper and media engagement, and technology
- Developing a questionnaire relevant to the course
- Revising a poorly graded assignment previously submitted

After finalizing the study details, faculty members shared a recruitment flyer and course announcement—both created by the lead researcher—which included a link to the screening survey (see Appendix B). Upon completing the study, students received a confirmation email and were instructed to upload it to a designated Google Form. This form was later reviewed to verify eligibility for extra credit. Students who chose an alternative extra credit assignment submitted their work directly to their course instructor, who graded it according to the class's standards.

3.3.3 Participant Onboarding

Prior to the study, participants underwent a screening survey process to ensure eligibility. This step filtered participants who met the age criterion (18–26 years old) and collected email addresses to facilitate troubleshooting of software or browser issues, match survey data accurately, and improve retention. To protect privacy, all identifiable information, including email addresses, was securely removed from the dataset after data collection. Eligible participants were then randomly assigned to one of four groups and directed to begin the initial survey.

As part of the screening, participants reviewed the Study Information Sheet (SIS), which outlined the study's procedures, including the completion of an initial survey and a post-test survey. To maintain the integrity of the research, the SIS included a cover story titled "Understanding Generation Z's Perception of Celebrity-Endorsed Products on Social Media," which framed the study as an exploration of celebrity influence on social media purchasing behavior (see Appendix A). This deception was necessary to prevent participants from adjusting their responses based on the study's true objectives. The SIS also detailed potential risks, measures to protect participant confidentiality, the voluntary nature of participation, and the right

to withdraw at any time without consequences. Additionally, the SIS included the lead researcher's contact information, allowing participants to ask questions or address concerns.

3.3.4 Participant Completion Analysis

A total of 1,609 participants enlisted in the study. 1,331 participants started the screening survey, but only 1,290 participants successfully completed and passed the screening criteria. Participants were then randomly assigned to one of four groups: Group 1, Group 2, Group 3, and Group 4. Each group varied in its level of retention and completion rates. Out of the total participants who began the study, 800 participants completed both the initial and posttest surveys, while 531 either failed to complete the post-test survey within the required 96-hour window or were otherwise excluded from the analysis due to incomplete, invalid, or multiple responses. The distribution of participants across groups—including how many started, finished, and were excluded—is detailed in the chart below (see Table 3.5). This breakdown provides an overview of participant engagement and highlights areas where attrition was most significant.

Table 3.5

Participant Responses Breakdown by Group

Group	Started	Finished	Excluded	Final Data Set
Screening	1,331	1,290	141	Final Total: 800
Group 1	337	278	76	202
Group 2	330	294	96	198
Group 3	331	319	118	201
Group 4	338	324	125	199

Strict inclusion criteria ensured data validity, with noncompliant responses removed from the final data set. First, participant authenticity was verified through email addresses and unique

ID numbers, with any incorrectly entered or unidentifiable email addresses or mismatched IDs leading to removal. To maintain temporal consistency in responses, participants who failed to complete the post-test survey within 24 to 96 hours after the initial survey were excluded (Riesthuis et al., 2022). Additionally, age eligibility was strictly enforced, with only respondents between 18 and 26 years old included in the final dataset. To prevent external influence, 86 participants who self-reported searching for the study's products or claims ("yes" or "I can't remember") were removed. Additionally, participants who attempted to complete the survey multiple times were excluded to ensure the integrity of individual responses. Incomplete responses were excluded to ensure comprehensive data integrity (Nardi, 2018). Finally, an audio validation test was embedded at the beginning of each survey, requiring participants to listen to a playable video and correctly input a series of numbers spoken in the video. Only those who accurately entered the spoken numbers had their data included.

3.3.5 Adult Consent Process

Participants were approached through a survey format and presented with an overview of the study. They were informed that the survey pertained to a marketing study about celebrities endorsing products on social media. Those who chose to continue and participate clicked "next" to begin the study. During this initial engagement, the study's true purpose, focusing on deepfake videos, was not disclosed. Later, participants were fully informed about the true nature of the study, regardless of whether they fully or partially completed it. If participants had any questions about the study, an email address was provided to gauge and address their concerns. The consent process was finalized without a signature, as continuing the survey was considered implied consent. Disclosure of the study's true purpose as part of the consent process would have biased the research subjects, rendering the results meaningless.

At the end of the study, participants who fully completed the study were given the choice to opt out of having their data included. Participants who partially completed the study did not have their data included in the analysis.

3.4 Research Design

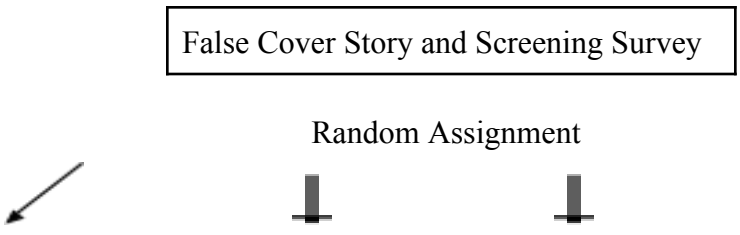
3.4.1 Procedure

This experiment (see Figure 3.1) employed a virtual, pre-post control-group design in which participants were randomly assigned to conditions, comparing the effects of deepfake videos on false memories between those who received an active debrief before false memory assessment and those who did not. Random assignment was key to ensuring internal validity and reducing selection bias between groups. To control for instrument effects, two versions of the treatment materials—a deepfake video featuring a celebrity product endorsement—were used, with one featuring a male celebrity (Elon Musk) and the other featuring a female celebrity (Kim Kardashian), ensuring gender was counterbalanced across participants. This design effectively functions as a 2x2 factorial Solomon Four-Group Design (Sawilowsky, 1994).

In the intervention phase, participants were shown three real or deepfake celebrity product endorsement videos. Half of the participants received an active debrief (see Appendix D) as part of the initial instrument, while the other half did not receive the active debrief until after they completed the delayed post-test survey, which included a false memory test. The active debriefing session, when administered during the initial instrument, lasted 5–7 minutes to ensure clarity and understanding, whereas the remaining participants were debriefed only upon finishing the virtual post-test survey, which took about 2–3 minutes. In the delayed post-test survey, every participant was presented with seven claims for each celebrity and asked to mark all claims they believed to be true.

Figure 3.1

Experimental Design Overview



Condition 1 Group A: Deepfake Kim Kardashian	Condition 1 Group B: Deepfake Elon Musk	Condition 2 Group A: Deepfake Kim Kardashian	Condition 2 Group B: Deepfake Elon Musk
Initial Survey Intervention Variables: ● Prior Knowledge (PK) of Celebrity ● Social Media Use	Initial Survey Intervention Variables: ● Prior Knowledge (PK) of Celebrity ● Social Media Use	Initial Survey Intervention Variables: ● Prior Knowledge (PK) of Celebrity ● Social Media Use Active Debrief Variables: ● Discernment ● Confidence ● Cues and Clues ● PK of Deepfakes ● Feelings About Deepfakes	Initial Survey Intervention Variables: ● Prior Knowledge (PK) of Celebrity ● Social Media Use Active Debrief Variables: ● Discernment ● Confidence ● Cues and Clues ● PK of Deepfakes ● Feelings About Deepfakes
24-96 Hours Later	24-96 Hours Later	24-96 Hours Later	24-96 Hours Later
Post-Test Survey False Memory Test Active Debrief Variables: ● Discernment ● Confidence ● Cues and Clues ● PK of Deepfakes ● Feelings About Deepfakes	Post-Test Survey False Memory Test Active Debrief Variables: ● Discernment ● Confidence ● Cues and Clues ● PK of Deepfakes ● Feelings About Deepfakes	Post-Test Survey False Memory Test	Post-Test Survey False Memory Test

In Condition 1, Groups A and B received an Intervention instrument. Following the intervention, they received four follow-up emails: a thank-you email informing them to look out for a post-test survey in 24 hours, an email with a link to the post-test survey sent 24 hours later, a reminder email to complete the post-test survey sent in 72 hours, and a reminder email to complete the post-test survey instrument sent in 48 hours. The delayed post-test survey instrument included both the false memory test and the active debrief instruments. In this condition, participants in Group A viewed the falsified Kim Kardashian video and took part in an Active Debrief 24–96 hours after completing the post-test survey, while those in Group B viewed the falsified Elon Musk video and received Active Debrief within the same timeframe. After completing the post-test survey instrument within the 24–96-hour window, participants immediately received a thank-you email reminding them of the true purpose of the study and indicating which video was a deepfake.

In Condition 2, Groups A and B received an intervention instrument immediately followed by the active debrief instruments. Following the instrument, they received the same four follow-up emails as in Condition 1: a thank-you email informing them to look out for a post-test survey in 24 hours, an email with a link to the post-test survey sent 24 hours later, a reminder email to complete the post-test survey sent in 72 hours, and a reminder email to complete the post-test survey instrument sent in 48 hours. In this condition, participants in Group A viewed the falsified Kim Kardashian video and took part in an Active Debrief immediately after the intervention, while participants in Group B viewed the falsified Elon Musk video and were debriefed immediately after the intervention. The delayed post-test survey instrument in Condition 2 included only the false memory test as all participants were told the true nature of the study in the initial survey. After completing the post-test survey within 24 to 96 hours,

participants immediately received a thank-you email reminding them of the true purpose of the study and indicating which video was a deepfake.

All participants who were exposed to a deepfake video were sent an email after one week, reminding them of the true nature of the study and which video was a deepfake. Additionally, at the end of the survey, participants were invited to opt out or choose not to have their data included in the study. Participants who completed the treatment involving a deepfake video presented as true, within the context of a marketing study, but did not complete the active debrief or false memory test, were sent debriefing materials one week after their partial participation. The protocol was done to ensure that participants were made aware of which content was false and to mitigate false memories upon study completion.

3.4.2 Piloting

To assess the rigor and reliability of the study, two pilot tests were conducted (Babarović et al., 2021). In the initial pilot, 62 participants, consisting of family and friends recruited through LinkedIn, text, and email, were divided into four groups to systematically identify potential issues in the questionnaire, such as typos, technical problems, or other logistical challenges. This process provided valuable feedback and helped troubleshoot effectively, leading to minor but necessary cosmetic adjustments to the survey design. For the second pilot, 12 email addresses were created to verify that all automatic emails and response mechanisms functioned seamlessly. All data collected during both pilots was promptly discarded upon completion, ensuring participant privacy and data integrity. This thorough testing process ensured the study was both user-friendly and technically sound before full deployment.

3.4.3 Materials and Tools Used

The study employed a combination of real and deepfake videos, survey software, email automation tools, and participant communication methods to investigate the effects of authentic and fabricated content. Real videos featuring Kim Kardashian, Elon Musk, and Paris Hilton were sourced from their social media accounts and interviews, while deepfake videos for Elon Musk and Kim Kardashian were created using the Lipsynthesis platform for video generation, with ChatGPT providing conceptual ideas for hypothetical claims and promotional scripts. All videos were seamlessly integrated into a survey platform designed and hosted with QuestionPro, which facilitated participant group randomization, automated instructions, and efficient data collection. To address limitations in QuestionPro's email automation features, external tools like Zapier were employed to manage follow-ups and reminders, ensuring participants received timely and clear communications. This combination of tools and processes proved to be laborious but ensured that surveys were effectively managed while providing participants with support to maintain engagement and retention throughout the study.

Creation of Celebrity Videos: Real and Deepfakes

The study required producing two distinct types of videos: real videos and deepfake videos. These videos were created and curated with careful attention to detail to ensure consistency in style, quality, and presentation. The following section outlines the methods and considerations involved in their development, including the selection of video materials, considerations in types of content, technical adjustments to the videos, and the generation of deepfake content.

At the start of the video creation and selection process, a wide range of well-known celebrities who actively promote products on social media were considered. Test videos and

deepfakes were created in January 2024 featuring Elon Musk, Kim Kardashian, Cardi B, and Paris Hilton. Each video was reviewed to maintain the same standard of quality, social media relevance, and alignment with the study's objectives as well as ethical considerations. The ethical consideration was to ensure that the subjects of the videos were non-political, sexual, or controversial and that, if the study inadvertently implanted false memories, the materials would be so inconsequential that they would not impact participants in any harmful way.

Ultimately, videos from Elon Musk, Kim Kardashian, and Paris Hilton were selected for inclusion. Kim Kardashian and Elon Musk were chosen due to their widespread recognition and their distinct yet influential public personas. Kim Kardashian, a dominant figure in entertainment, fashion, and social media, represents influencer culture, while Elon Musk, known for his technological innovations and corporate leadership, represents entrepreneurship and futurism. This selection ensures a balance in gender representation, differing audience demographics, and varying content themes. By incorporating both a media personality and a tech entrepreneur, the study examines whether the nature of the celebrity influences participant perception and engagement with deepfake content. Cardi B's videos, while engaging, did not exhibit the same level of technical consistency and her expressive gestures made it significantly more challenging to create convincing deepfakes, leading to her exclusion from the final set of materials.

All videos were designed to reflect the style and format typical of TikTok and Instagram Reels. These platforms feature videos in a vertical portrait mode, optimized for mobile viewing, and emphasize short, engaging content. TikTok videos are known for their dynamic and often fast-paced storytelling, while Instagram Reels highlight visually captivating clips with similar brevity (Vidani, 2024). The videos used in the study were trimmed and edited to ensure uniform

lengths across all samples. This consistency helped maintain participants' attention and minimized potential bias introduced by varying video durations (Koç, 2023).

For the real videos (see Figure 3.2), sources included materials pulled from the personal social media accounts of Kim Kardashian and Paris Hilton. In the case of Elon Musk, an online interview was chosen and subsequently edited into a TikTok-style portrait format to ensure compatibility with the other materials. These real videos were then uploaded to QuestionPro and embedded into their respective experimental groups, accompanied by text headings to provide necessary context and clarity.

Figure 3.2

Real Videos of Paris Hilton, Elon Musk, and Kim Kardashian

Paris Hilton announces Hilton Honors Program coming to Roblox (Hilton, 2024)

Elon Musk advertises White Claw: An alcoholic beverage (MRQUICK, 2022)

Kim Kardashian launches catsuits for SKIMS shapewear (Kardashian, 2023)

In the real Paris Hilton video, the original content was sourced from her TikTok account, with only a 10-second segment used for the study. The video featured Paris announcing a unique

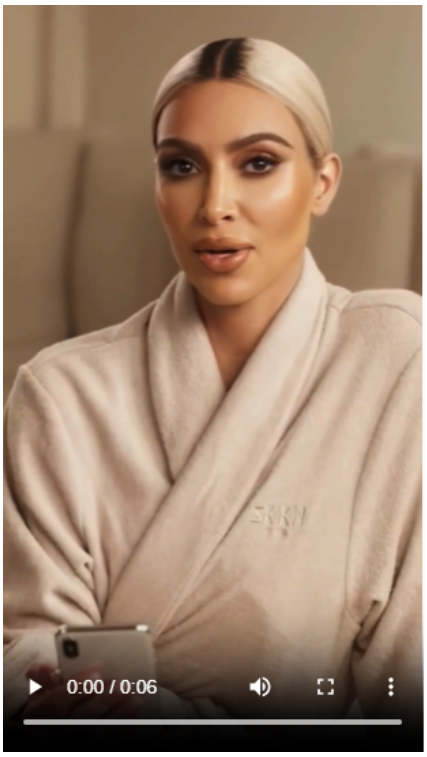
partnership between the Hilton Hotel awards program and Roblox (Hilton, 2024). To align with the study's parameters, the footage was cropped to fit within the designated frame and to remove the original subtitles. Similarly, the real Elon Musk video was taken from MRQUICK's (2022) YouTube clip, in which he discussed his enthusiasm for the alcoholic seltzer White Claw during a podcast appearance. This footage was reformatted from YouTube's standard dimensions to conform to TikTok's vertical format. Lastly, the real Kim Kardashian video was originally posted on her TikTok page (Kardashian, 2023), where she introduced new SKIMS shapewear undergarments from her personal brand. This video was also cropped to remove subtitles and maintain consistency across the study's presentation.

These deepfake videos (see Figure 3.3), featuring Elon Musk and Kim Kardashian, were designed to introduce controlled false information while maintaining a realistic appearance. The development of deepfake content followed a structured process. First, a variety of hypothetical claims about each celebrity were generated using ChatGPT. Prompts such as "Come up with a list of claims about Elon Musk that might be true or completely false" and "Come up with a list of claims about Kim Kardashian that might be true or completely false" were used to produce dozens of potential claims. After reviewing these claims, a few were selected, refined, and adapted into promotional scripts designed for social media (Rao, 2022).

Figure 3.3

Deepfake Videos of Elon Musk and Kim Kardashian

Elon Musk introduces OmniSphere: a venture building Ironman suits (Real Time with Bill Maher, 2023)



Kim Kardashian launches Eco Edibles: Wearable, edible clothing (Nast, 2022)

Once the scripts were finalized, the Lipsynthesis platform, a tool designed for educational and research purposes related to deepfake technology, was used to generate the videos. Lipsynthesis offers a roster of celebrity likenesses, including visual video options for Kim Kardashian (Nast, 2022) and Elon Musk, which were utilized to create the deepfakes. The scripts were then further adjusted to fit each celebrity's tone and style of social media advertisements. The creation of Elon Musk's deepfake video was relatively straightforward, requiring only three iterations to achieve a usable result. In contrast, the process for Kim Kardashian's deepfake was more complex, involving over 31 renderings to synchronize mouth movements and facial expressions convincingly. After the deepfake videos were generated, post-production edits were

conducted to refine their appearance. Sections of the videos that appeared overly artificial or inconsistent with the celebrity's typical mannerisms were removed. To ensure the deepfakes closely resembled the celebrities' authentic promotional styles, over 200 reference videos were reviewed from each of their TikTok and Instagram profiles, as well as online interviews.

Finally, all videos, both real and deepfake, were uploaded to the survey platform. This meticulous process helped achieve a level of realism necessary for the study's objectives (Mirsky & Lee, 2021). Each video was accompanied by text headings to provide context and ensure clarity for participants. The consistent formatting and careful preparation of these videos contributed to the overall validity of the study, ensuring that participants could engage with the content in a standardized manner.

Note on Coincidental Events Following Materials, Pilot, and Data Collection

After the deepfake videos were generated in January 2024 and subsequent survey data collection concluded on November 22, 2024, a notable public event coincidentally occurred involving Elon Musk and AI-generated imagery. On November 23, 2024, Elon Musk shared an AI-generated image of himself in an Iron Man suit, humorously named "Irony Man," via his social media account. In the accompanying caption, Musk joked about using "the power of irony" to defeat villains (Economic Times, 2024). Although this event coincided closely with the conclusion of data collection, it is important to emphasize that all survey instruments, the pilot survey, and associated study materials were finalized, created, and distributed significantly in advance of this event. Additionally, the materials and procedures were also prepared prior to the election and Elon Musk's subsequent governmental involvement. As such, this incident had no bearing on the design, execution, or analysis of the research.

QuestionPro for Survey Design

The study relied on implementing QuestionPro, an online survey software, to design instruments, manage surveys, and collect data. UCI provided a QuestionPro license which included features to build complex surveys, randomize participant groups, and collect responses efficiently, making it an ideal platform for addressing the study's unique requirements. One key feature utilized was QuestionPro's ability to randomize subjects into different survey groups following the screening survey. This ensured balanced distribution and streamlined participation. The software automatically directed participants to the appropriate survey based on their screening responses, allowing for seamless transitions between different stages of the study.

To ensure effective communication with participants, QuestionPro's instant email automation features were utilized. Upon completing the first survey, participants immediately received emails with instructions on the next steps. For those who completed the posttest, the software automatically sent emails disclosing the true purpose of the study along with information about which video was a deepfake and an explanation of what deepfakes are. However, QuestionPro lacked a feature to schedule delayed emails or send automated reminders. To address this limitation, Zapier, a workflow automation tool, was integrated to handle follow-up communications. This integration ensured that all participants received timely and relevant updates.

Managing the "back" button feature presented another challenge. QuestionPro did not allow selective disabling of the "back" button within a single survey, which posed a risk of participants revising their answers after learning the true purpose of the study. To mitigate this, some surveys were split into two parts. The first part enabled the "back" button, allowing participants to navigate naturally through the questions. After the true purpose of the study was

disclosed, the second part of the same survey set had the "back" button disabled, preventing participants from altering their previous responses. QuestionPro's survey chaining feature allowed the two parts to be linked so participants could complete both sections within one sitting. However, this approach introduced a potential issue: if the survey timed out or participants experienced an internet connection failure, they were unable to resume. This required additional troubleshooting with participants to address instances where data collection was disrupted. Unfortunately, troubleshooting could only occur if the subject emailed the lead researcher, and not all subjects were able to complete the survey within the time frame necessary.

Zapier for Email Reminder Automation

Zapier, a workflow automation tool, was utilized to address limitations in QuestionPro's survey software, specifically the absence of a feature allowing for delayed follow-up emails. Zapier connects various applications through automated workflows, enabling seamless communication and task management. For this study, Zapier was integrated with QuestionPro via an API that required administrative approval from the University of California, Irvine (UCI). This integration allowed automated emails to be sent to participants, ensuring timely follow-ups and improving retention rates throughout the study. Due to license restrictions, it was learned that the study could only utilize up to 10,000 API calls; exceeding this limit would result in the Zapier integration being turned off and in turn, compromising the study. If projections suggested the study would exceed this limit, advance notice had to be sent to UCI's Office of Information Technology (OIT) to request permission to increase the limit. To ensure compliance, monthly emails were sent to request reports on the number of API calls used.

In terms of email automation, every subject who successfully completed the screening survey and initial survey received a follow-up email 24 hours later, containing a link to their posttest survey along with instructions on how to complete it. Additionally, all participants who completed the first survey received daily email reminders for four consecutive days, regardless of whether they had already completed the posttest. While this approach ensured no participant missed an email, it introduced a challenge: the reminders confused some participants, leading them to believe they had not properly completed the study or that additional surveys were required. To mitigate this issue, QuestionPro's survey settings were configured to prohibit participants from completing the survey more than once. Despite this safeguard, a small number of participants found workarounds by using a different browser or mobile device, resulting in duplicate survey entries. These duplicates were later identified and cleaned during data analysis to ensure data integrity.

The Zapier platform required a monthly subscription to facilitate the volume of emails necessary for the study. This subscription allowed for a sufficient number of automated emails to be sent to all participants, ensuring consistent communication. By leveraging Zapier's automation capabilities, daily reminders were sent without manual intervention, thereby maintaining participant engagement and addressing the study's retention goals. Although Zapier's automation was effective, its inability to track which participants had already completed the posttest meant that reminders continued for all participants, irrespective of completion status. This decision ensured no participant missed their chance to complete the survey, but required additional data management efforts to handle duplicate entries.

Email Communication with Subjects

The @uci.edu Gmail account served as the main communication channel for participants and was used to address a large volume of inquiries throughout the study. Common types of emails included troubleshooting requests for survey-related bugs or technical issues, such as difficulties playing audio, internet connections timing out, or submitting responses for surveys partially completed. Participants also frequently asked questions about the timing of their posttest emails, seeking more clarification on when they would receive the next step in the study. Additionally, some participants requested early verification that they had successfully completed the surveys, often for extra credit purposes. Other emails ranged from general questions about the study's process to requests for further clarification on survey instructions and questions regarding age or eligibility to participate in the study. Overall, email communication proved critical for resolving participant concerns promptly and contributed to higher retention rates.

3.5 Additional Considerations

3.5.1 Ethical and Moral

Incomplete disclosure was necessary to preserve the integrity of the study, as revealing the presence of deepfakes upfront would have influenced participants' recall and skewed the results (Levine, 1982). The study aimed to assess how subjects recalled falsified content under two different conditions, which would not have been possible if they were aware of the deception from the outset. To these ends, the study prioritized ethical considerations to ensure minimal risk to participants while utilizing deepfake technology. As an example, the falsified videos of Kim Kardashian and Elon Musk focused exclusively on immaterial and non-harmful content about hypothetical products, deliberately avoiding any material that could be political, ideological, financial, adult, or harmful. Careful measures were implemented to ensure participants'

well-being and to mitigate potential concerns surrounding false memory research. These measures included:

Exclusion of Sensitive Content

All real and falsified videos excluded references to political issues, harmful ideologies, adult content, or any content deemed likely to cause distress (Liamputtong, 2007). This ensured that the study's materials remained neutral and inoffensive, mitigating the risk of psychological or emotional harm.

Education and Disclosure of Falsified Content

When the true purpose of the study was revealed, participants were educated about deepfake technology and informed about the falsified nature of the videos during active debriefing. To further reinforce transparency, participants were reminded immediately after the study about which videos were falsified, and all participants received a follow-up email seven days later that disclosed the study's true purpose and clarified the nature of the deepfakes.

Post-Study Follow Up

To prevent the formation of false memories, participants who fully completed the survey received an immediate follow-up email informing them that they had been exposed to deepfake videos. This email explicitly identified the fabricated video and clarified that it was not real. To further mitigate potential false memories, all participants, regardless of whether they fully or partially completed the study, received a second email seven days after the screening survey. This email disclosed the true purpose of the study and specified which videos were deepfakes (Mara et al., 2012), ensuring transparency and addressing any post-study concerns. Additionally, participants who had questions about the study were provided an email address for

correspondence and were given an opportunity to opt out of their data being included in the final data set for analysis.

Preservation of Rights and Welfare

Although the full purpose of the study was not disclosed initially, this did not affect participants' rights or welfare. The cover story was relevant to the presented content, and follow-up measures maintained transparency, ensuring participants felt respected and informed (Hertwig & Ortmann, 2008). Through these carefully designed safeguards, the study adhered to ethical standards, protected participants, and addressed potential risks while providing meaningful insights.

3.5.2 Potential Societal and Scientific Benefits

This study provides a foundation for future interdisciplinary research and addresses the current landscape of deepfake technologies, with minimal socio-technical research available. Firstly, the study can improve digital literacy, equip participants to critically evaluate online content (Kruijt et al., 2022), and recognize manipulated videos and deepfakes on social media platforms. Additionally, the active debriefing instruments can help participants develop critical thinking skills (Greene & Murphy, 2023) that allow individuals to question the authenticity of information and audiovisual content. Increasing awareness of hyper-realistic, synthetic deepfakes (George & George, 2023) and the tools used to create them provides end-users with the skills to better navigate and mitigate misinformation and disinformation (Vraga et al., 2021), thus contributing to a more social media-savvy, informed, and resilient society.

Beyond individual benefits, this research can mitigate and protect society against falsified content. As deepfake technology continues to evolve and outpace deepfake detection (Gil et al., 2023), the potential to spread false narratives, scams, criminal or illegal activity, sexual

harassment, and more underscores the need for heightened awareness and more informed mitigation strategies. Thus, highlighting efforts towards increasing awareness, preparing for future technological challenges, and making additional contributions towards deepfake technology research (Mallinson et al., 2024). This dissertation can serve as an early framework for policymakers, educators, and researchers to expand upon the research findings and develop regulatory frameworks, strengthen media literacy initiatives, and improve tools to reduce social risks from deepfakes.

3.5.3 Limitations

This study faced several limitations and challenges that affected the quality of survey responses, leading to the exclusion of some data. One key issue was participant behavior and data reliability. Some participants entered their email addresses incorrectly or completed the survey multiple times, often due to receiving frequent reminder emails intended to boost survey completion rates. Although the survey software included a feature to prevent multiple submissions, some participants bypassed it by clearing their browser cache and cookies. To mitigate this, only the first completed response from each participant was included in the analysis.

Retention and completion rates also presented challenges. Some participants timed out and could not retake the survey, while others were excluded for failing to meet eligibility criteria, not finishing within four days, or leaving the survey incomplete. Participants who reached out to troubleshoot technical errors were granted permission to complete the study in order to receive extra credit; however, the data submitted was excluded if the participant had already learned the true purpose of the study or if they completed the study after the deadline.

The study environment may have influenced participants' ability to detect deepfakes. Participants knowing they were part of a research study prompted a number of individuals to approach the task with greater caution and analytical focus than they might in casual settings, such as scrolling through social media. This heightened awareness could have artificially improved their ability to detect deepfakes or affect their confidence and accuracy. Additionally, while participants were instructed to close all tabs and browsers to avoid googling the video content, which could reveal whether it was real or fake, compliance with this instruction was assessed only through self-reports, leaving some uncertainty about the reliability of the responses. Additionally, despite being instructed not to discuss the study, there is a possibility that some participants shared details with their peers or other participants, potentially influencing their responses.

Technical difficulties added further complications. Participants using Vivaldi browsers experienced audio issues that prevented them from completing the survey. To address this, an audio test was required to ensure only participants who could hear the content were included. While this measure improved data reliability, it may have excluded some participants who could not resolve the issue. Only those who contacted the lead researcher were able to complete the survey successfully. Finally, logistical challenges in managing the study were notable. Tools like QuestionPro and Zapier were employed to streamline communication and retention efforts, but their limitations required careful adjustments to workflows.

3.6 Summary of Methodological Approach

In conclusion, this chapter has outlined the steps the researcher has taken to investigate how deepfake videos might implant false memories and whether actively debriefing participants about deepfake videos can mitigate that effect. Under the guise of a cover story, the study showed a mix of real and fake celebrity videos to Gen Z adults. Following a deception design, participants were told the study

was about social media marketing, which helped capture their natural reactions. Later, they learned the true purpose of the study, allowing the researcher to measure if false memories can be formed and whether media literacy can mitigate false memory formation. The research was designed with ethical integrity, from robust recruitment efforts and the development of clear survey instruments to timely follow-up communications and meticulous data management. By detailing each procedural step in a structured manner, this chapter has established a solid foundation for investigating the effect of deepfakes on the formation of false memories.

Chapter 4

Research Analysis and Findings

This chapter presents the study's findings and analyzes research questions 1 through 11, focusing on three key areas: the effects of active debriefing and deepfake exposure on false memory formation, the influence of prior knowledge and social media usage on these relationships, and discernment accuracy, confidence levels, cues and clues, and individuals' perceptions of deepfakes. The study followed a mixed-methods approach, integrating quantitative and qualitative analyses to examine key patterns and relationships. The quantitative component applied descriptive statistics to summarize central tendencies, distributions, and variability across all research questions, using visual representations, interaction plots, frequency distributions, thematic categorization, and binary coding. Inferential statistics also explored relationships between independent and dependent variables, providing deeper insights into predictive associations and group differences.

The inferential analysis included binomial logistic regression (Peng et al., 2022), Chi-Square tests (Singhal & Rana, 2015), and pseudo-R-squared measures (Allison, 2013), which estimate how well the logistic regression model fits the data compared to a model with no predictors. Logistic regression models were used for research questions 1 through 8 and 10 to estimate predictive relationships for binary outcomes. To ensure model fit, inferential analysis

included binomial logistic regression, chi-squared tests, pseudo-R-squared measures, and frequency analysis (Neuendorf, 2018).

Model fit, how well the model explains or predicts the outcome, was evaluated using Cox & Snell R^2, Nagelkerke R^2, and McFadden R^2, while Chi-Square tests assessed how well the models fit the data by comparing expected and observed frequencies. R-squared values determined how much of the variation in the dependent variable was explained by the independent variables. Statistical significance for all tests was set at $p < .05$ to determine meaningful relationships. No alpha corrections were applied, as each research question was tested using a single statistical method, eliminating the risk of inflated Type I error or false positives from multiple comparisons. Finally, post-hoc comparisons (Williams & Abdi, 2010) were conducted to identify group differences and refine interpretations.

The qualitative component involved analyzing open-text responses for Research Questions 9 through 11 using content analysis. Themes were then systematically categorized and quantified through descriptive statistics, enabling a structured frequency analysis to examine participant perceptions. To enhance analytical rigor, the study employed protocol analysis (Chi, 1997) to identify patterns in participants' reasoning and decision-making processes, then iteratively refined the coding schema and models to better align with the established protocol framework. Each research question presents a hypothesis, results from both statistical and thematic analyses, and a summary of findings.

Results for the Elon Musk and Kim Kardashian deepfake videos were kept separate to account for important content-specific differences that could influence participant responses. Not all deepfakes are created equal—variations in subject matter, tone, and contextual plausibility can significantly shape memory formation and viewer interpretation. By analyzing them

independently, the analysis can preserve the integrity of these content-based nuances and avoid masking distinct patterns of false memory formation tied to each video's unique characteristics.

4.1 Demographics and Participant Overview

The study included a diverse sample of young adults aged 18-26, representing a range of gender identities and educational backgrounds (see Table 4.6). The gender distribution was relatively balanced between male (53.1%) and female participants (44.0%), with a small percentage identifying as nonbinary (1.8%) or choosing not to disclose their gender (0.1%). In terms of education, most participants had completed at least some higher education (60.4%), with varying levels of attainment ranging from high school diplomas (39.6%) to bachelor's degrees (8.1%). No participants reported holding advanced degrees.

Table 4.6

Sociodemographic Characteristics of Participants at Baseline

Baseline Characteristic	*n*	%
Gender		
Female	352	44.0
Male	425	53.1
Nonbinary	14	1.8
Self-describe	1	0.1
Prefer not to disclose	8	1.0
Highest Education Level		
High school diploma or equivalent	317	39.6
Some college without completing a degree	258	32.3
Associate degree	159	19.9

Bachelor's Degree	65	8.1
Master's degree	0	0.0
Doctorate	0	0.0
Other	1	0.1

4.2 Research Questions 1 and 2 (RQ1 & RQ2): Effects of Deepfake Exposure on False Memory Formation and the Role of Active Debriefing

4.2.1 Research Questions and Hypotheses

RQ1. Can exposure to a deepfake video (Kim Kardashian or Elon Musk) implant false memories?

H_0: Exposure to deepfake videos can not significantly increase the likelihood of false memories being implanted.

H_1: Exposure to deepfake videos induces false memories.

RQ2. Can an active debriefing (informing participants that the video is a deepfake) deter the implanting of false memories?

H_0: Active debriefing can not significantly decrease the likelihood of false memories being implanted.

H_1: Active debriefing can significantly decrease the likelihood of false memories being implanted.

H_2: There is an interaction between active debriefing and exposure to deepfake videos in influencing the likelihood of false memories being implanted.

A binary logistic regression was conducted to examine the relationship between exposure to a deepfake video (Kim Kardashian vs. Elon Musk) and the influence of active debriefing on false memory formation. Binary coding was applied to key variables for analysis. False memory

retention was coded as 0 for participants who did not report any false memories and 1 for those who retained at least one. Similarly, the active debriefing variable was coded as 0 for participants who did not receive an active debrief and 1 for those who did. This approach allowed for assessing how deepfake exposure and active debriefing interact to influence false memory formation.

4.2.2 RQ1 and RQ2 Descriptive Statistics

A total of 355 participants reported or retained at least one false memory (see Table 4.7). Among participants who viewed the Elon Musk deepfake video, 58.3% reported false memories, whereas 44.3% who viewed the Kim Kardashian deepfake video reported false memories. In total, 412 false memories were reported, with 273 false memories related to the Elon Musk deepfake and 139 false memories related to the Kim Kardashian deepfake.

Table 4.7

False Memory Formation by Deepfake Condition

False Memory Formation	Kim Kardashian Deepfake With Active Debriefing		Kim Kardashian Deepfake Without Active Debriefing		Elon Musk Deefake With Active Debriefing		Elon Musk Deepfake Without Active Debriefing	
	n	%	*n*	%	*n*	%	*n*	%
At Least One False Memory	12	5.8%	122	63.9 %	19	9.2%	135	64.6%
No False Memory	194	94.2%	69	361. %	194	90.8 %	74	35.4%

Notably, active debriefing appeared to have a substantial impact on false memory formation (see Figures 4.4 and 4.5). Among participants who viewed the Kim Kardashian deepfake video, those who did not receive an active debriefing were more likely to report false

memories, with 122 individuals (63.9%) doing so, compared to only 12 participants (5.8%) who received an active debriefing. A similar pattern was observed among those who watched the Elon Musk deepfake video, where 135 participants (64.6%) who did not receive an active debriefing reported false memories, whereas the number dropped to 19 participants (9.2%) among those who were actively debriefed (see Figure 4.6). These findings suggest that informing participants about the nature of deepfake content after exposure can help mitigate false memory formation.

Figure 4.4

False Memory Scores by Group (Deepfake Viewers vs. Real Video Viewers)

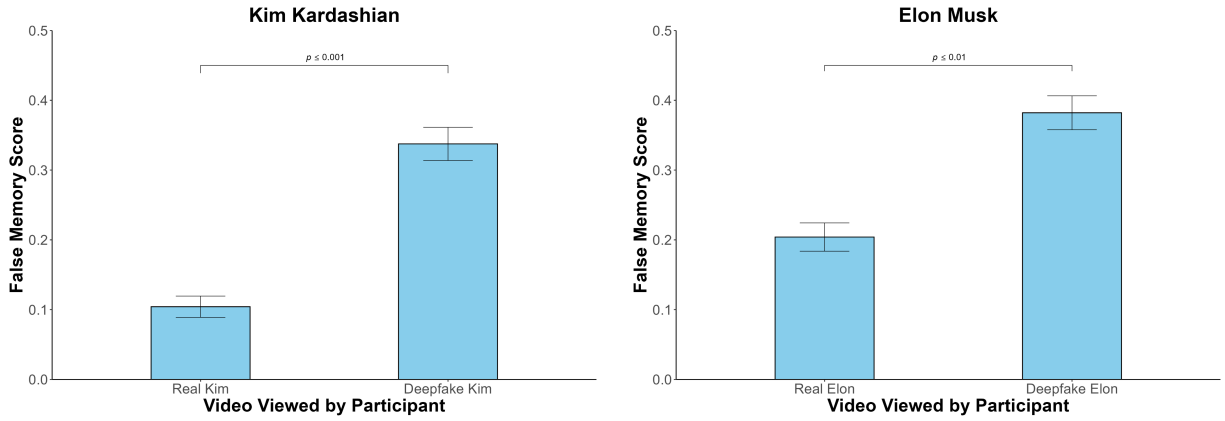


Figure 4.5

False Memory Mean Scores of Deepfake Participants: Impact of Active vs. No Active Debrief

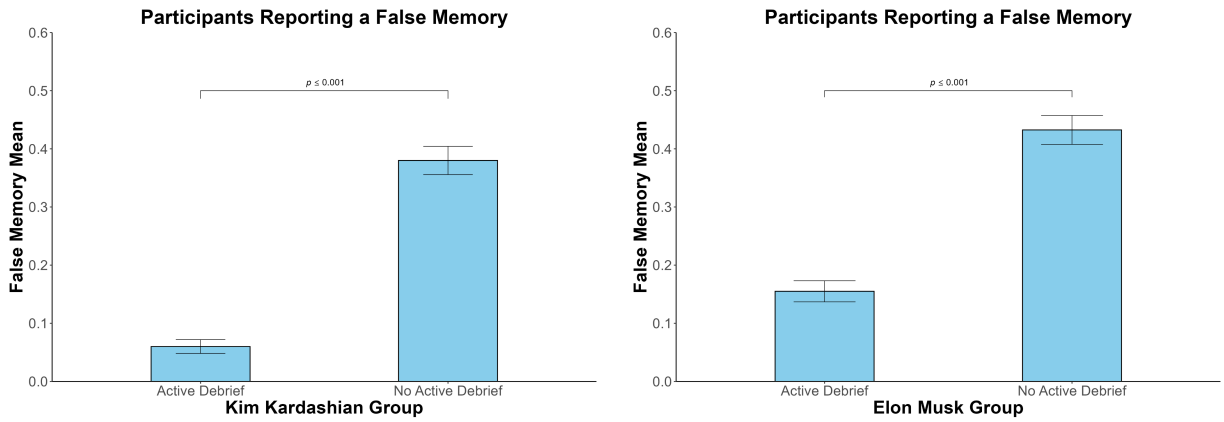
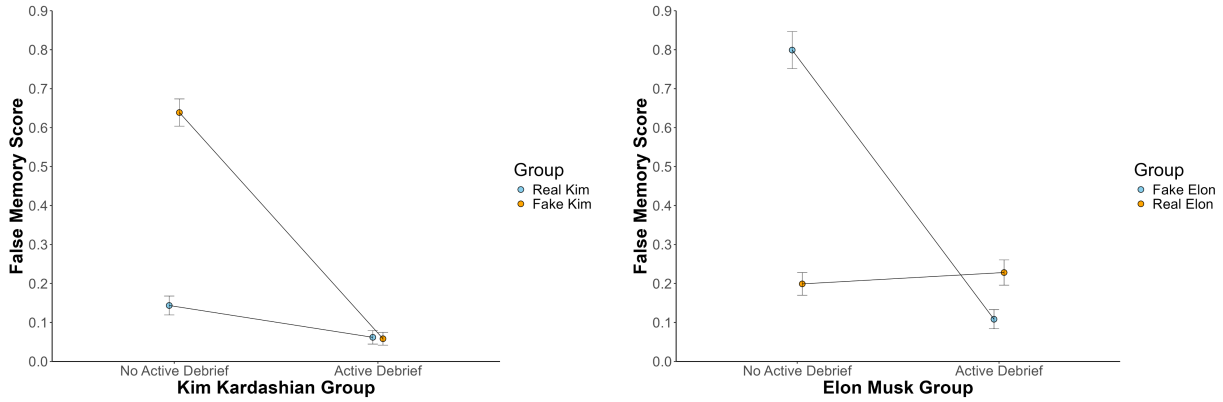


Figure 4.6

Interaction Plots: False Memory Scores by Group and Debriefing



4.2.3 RQ1 and RQ2 Binary Logistic Regression Results

The analysis focused on whether participants retained at least one false memory, rather than the total number of false memories reported. Independent variables included exposure to the Kim Kardashian deepfake (with the real Kim Kardashian video as the reference category), exposure to the Elon Musk deepfake (with the real Elon Musk video as the reference), whether participants received an active debriefing, and an interaction term between the deepfake condition and active debriefing.

Elon Musk Deepfake Model

The logistic regression model was conducted to examine the relationship between exposure to each unique deepfake video (Kim Kardashian or Elon Musk) or no exposure to a deepfake video and the likelihood of false memory formation, measured as a binary outcome. The model examining false memories related to the Elon Musk deepfake was statistically significant, $\chi^2(3, N = 800) = 171.07$, $p < 0.001$, R^2 (Cox & Snell) = 0.19, R^2 (Nagelkerke) = 0.27, and correctly classified 78.3% of cases.

Participants who had been exposed to the real Elon Musk video instead of the Elon Musk deepfake were less likely to develop false memories about Elon Musk (OR = 0.14, 95% CI = 0.09 - 0.21, $p < 0.001$). Additionally, those who received an active debrief, where they were explicitly told about the possibility of deception, were also less likely to develop false memories (OR = 0.06, 95% CI = 0.03 - 0.10, $p < 0.001$). Furthermore, there was a strong interaction effect between which deepfake participants saw and whether they received a debrief (OR = 17.85, 95% CI = 8.64 - 37.88, $p < 0.001$).

Kim Kardashian Deepfake Model

The model examining false memories related to the Kim Kardashian deepfake also showed statistically significant results, $\chi^2(3, N = 800) = 239.68$, $p < 0.001$, R^2 (Cox & Snell) = 0.23, R^2 (Nagelkerke) = 0.34, and correctly classified 81.5% of cases.

Exposure to the Kim Kardashian deepfake increased the likelihood of forming false memories (OR = 10.55, 95% CI = 6.56 - 17.40, $p < 0.001$). Receiving an active debrief significantly reduced false memories (OR = 0.09, 95% CI = 0.03 - 0.23, $p < 0.001$). There was also a significant interaction effect (OR = 0.39, 95% CI = 0.19 - 0.77, $p < 0.01$) indicating that the effectiveness of debriefing varied depending on whether participants saw the real Kim

Kardashian video or Kim Kardashian deepfake, with debriefing being more effective at reducing false memories for one deepfake over the other.

4.2.4 Summary of Findings

Overall, results suggest that exposure to a deepfake video can significantly influence the likelihood of forming false memories, with notable differences emerging when comparing the deepfake and real versions of the same individual, either Elon Musk or Kim Kardashian, rather than across all four videos simultaneously. Exposure to deepfake videos demonstrated a significant increase of false memory formation compared to participants who were not exposed to deepfakes, with those viewing the Elon Musk deepfake video demonstrating slightly higher rates of false memory formation than those viewing the Kim Kardashian video.

Active debriefing played a critical role in reducing false memory formation across both conditions, and its effectiveness differed based on which deepfake participants saw. While descriptive statistics indicate that false memories were more frequently reported in both deepfake conditions, but logistic regression analysis revealed that deepfake type and debriefing interacted in complex ways, influencing memory retention differently across conditions. The moderate explanatory power of both models (as indicated by R^2 values) suggests that deepfake exposure and debriefing are important factors, though other unmeasured variables may also contribute to false memory formation.

4.3 Research Question 3 (RQ3): How can the impact of exposure to a deepfake video (Kim Kardashian or Elon Musk) on implanting false memories vary with the level of prior knowledge of celebrity?

H_0: Prior knowledge of video content can not influence the effect of exposure to deepfake videos on false memories.

H_1: Prior knowledge of video content influences the effect of exposure to deepfake videos on false memories.

To analyze RQ3, a logistic regression model was used to assess whether prior knowledge of Kim Kardashian and Elon Musk influenced the likelihood of false memories being implanted after exposure to deepfake videos. The dependent variable (false memory: present or absent) was analyzed using a binomial logistic regression with interaction effects between deepfake exposure and prior knowledge.

Prior knowledge was measured by asking participants to rate how much they knew about Kim Kardashian or Elon Musk on a 5-point scale, ranging from "Not knowledgeable at all" to "Extremely knowledgeable." For analysis, responses were consolidated into three broader categories: participants who reported little to no knowledge about the celebrity (ratings 1-2) were coded as 0, those who were somewhat knowledgeable (rating 3) were coded as 1, and those who were highly knowledgeable (ratings 4-5) were coded as 2. Reducing the 5-point scale to three broader categories helped simplify the analysis by minimizing noise and enhancing interpretability, particularly given the sample size and the focus on broader trends rather than granular distinctions. This also made the models more stable and allowed for clearer comparisons between participants with low, moderate, and high levels of familiarity.

4.3.1 RQ3 Descriptive Statistics

The study categorized participants based on their prior familiarity with Kim Kardashian and Elon Musk (see Table 4.8 and Figure 4.7). A larger portion of participants had limited knowledge of Kim Kardashian (42.6%), while a moderate group reported some familiarity (38.5%), and a smaller group identified as highly knowledgeable (18.9%). In contrast, familiarity with Elon Musk followed a different pattern, with fewer participants reporting little to no

knowledge (19.1%), a substantial portion having some familiarity (50.0%), and a notable number identifying as highly knowledgeable (30.9%).

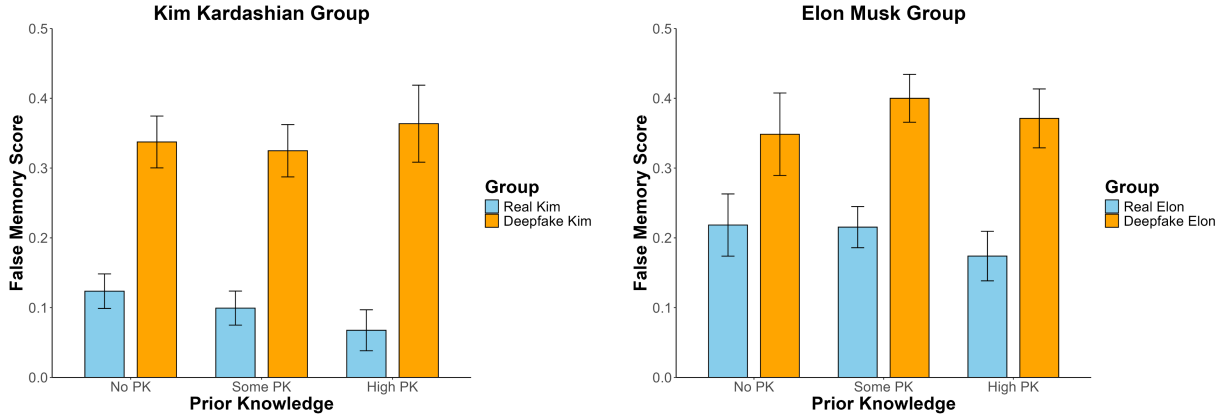
Table 4.8

Categorization of Prior Knowledge: Kim Kardashian vs. Elon Musk

Prior Knowledge	of Kim Kardashian		of Elon Musk	
	n	%	*n*	%
Little to No Knowledge About the Celebrity	341	42.6	153	19.1
Somewhat Knowledgeable	308	38.5	400	50.0
Highly Knowledgeable	151	18.9	247	30.9

Figure 4.7

False Memory Scores, Grouped by Prior Knowledge Level and Type of Video



Kim Kardashian Model

The Kim Kardashian model was statistically significant $\chi^2(5, N = 397) = 71.02$, $p < .001$, R^2 (Cox & Snell) = 0.08, R^2 (Nagelkerke) = 0.13, and correctly classified 78% of cases. Overall, exposure to a deepfake video of Kim Kardashian significantly increased false memory implantation (OR = 3.61, 95% CI = 2.02 - 6.75, $p < .001$), indicating that participants who viewed the manipulated video were more likely to form false memories. However, having some

prior knowledge level (OR = 1.12, 95% CI = 0.56 - 2.28, p = 0.742) or high level of prior knowledge (OR = 0.98, 95% CI = 0.37 - 240, p = 0.973) did not have a significant main effect on false memory formation. These findings suggest that prior knowledge of Kim Kardashian did not significantly moderate the relationship between deepfake exposure and false memory formation.

Elon Musk Model

The Elon Musk model was statistically significant ($\chi^2(5, N = 403) = 37.32$, $p < .001$), R^2 (Cox & Snell) = 0.05, R^2 (Nagelkerke) = 0.06, and correctly classified 70.6% of cases. In contrast to Kim Kardashian, prior knowledge of Elon Musk was a significant predictor of false memory formation. Participants who reported being somewhat (OR = 1.82, 95% CI = 1.03 - 3.29, p = 0.043) or highly knowledgeable (OR = 1.97, 95% CI = 1.09 - 3.68, p = 0.028) about Elon Musk were more likely to form false memories. This suggests that pre-existing familiarity with Elon Musk did directly contribute to false memory formation, exerting a stronger influence than the deepfake exposure itself. However, overall was no significant interaction between prior knowledge and the effect of deepfake exposure on false memory formation.

4.3.2 Summary of Findings

The results highlighted that participants viewing the deepfake video of Kim Kardashian were likely to form false memories, regardless of how familiar they were with the celebrity beforehand. On the other hand, for Elon Musk, participants who had somewhat prior knowledge or were highly knowledgeable were more likely to develop false memories, while simply viewing the deepfake was not as influential. This suggests that the impact of prior knowledge on false memories could vary depending on the celebrity, how well-known they are to the public, or potentially other bias measures.

4.4 Research Question 4 (RQ4): What is the relationship between prior knowledge of the video content (specifically, the celebrity) and the effectiveness of active debriefing?

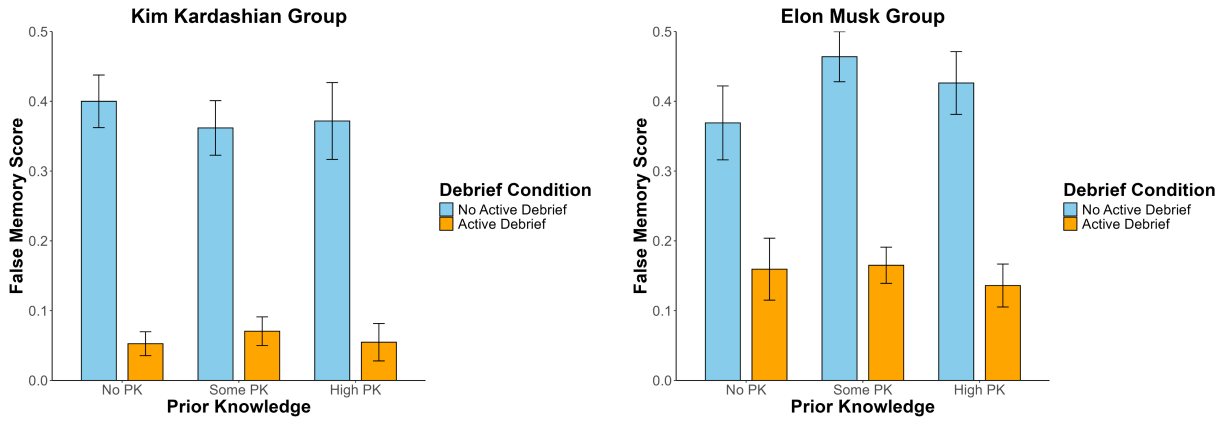
H_0: Prior knowledge of the video content has no effect on the effectiveness of active debriefing in reducing false memories.

H_1: Prior knowledge of video content affects the effectiveness of active debriefing in reducing false memories.

A logistic regression model was used to assess whether active debriefing interacted with prior knowledge of the video content (specifically, the celebrity) to reduce false memories (see Figure 4.8). The independent variables included prior knowledge of the celebrity (categorized as little to no knowledge, somewhat knowledgeable, or highly knowledgeable) and active debriefing (binary: 0 = no active debrief, 1 = received active debrief). The dependent variable was false memory formation, coded as 0 for participants who did not report any false memories and 1 for those who retained at least one false memory.

Figure 4.8

False Memory Scores by Prior Knowledge Level, Type of Video, and Debrief Condition



4.4.1 RQ3 Binary Logistic Regression Results

Kim Kardashian Model

The Kim Kardashian model was statistically significant ($\chi^2(5, N = 400) = 71.02$, $p < .001$), R^2 (Cox & Snell) = 0.08, R^2 (Nagelkerke) = 0.13, and correctly classified 78.0% of cases. Participants who had some prior knowledge of Kim Kardashian had statistical significance (OR = 1.65, 95% CI = 1.05 - 2.61, p = 0.032) as opposed to those with high prior knowledge (OR = 1.10, 95% CI = 0.63 - 1.90, p = 0.732), suggesting that those with some prior knowledge compared to those with no knowledge were slightly more prone to false memories. However, there was no significant interaction between prior knowledge and the effectiveness of active debriefing for either group, including participants with some prior knowledge (OR = 0.61, 95% CI = 0.22 - 1.65, p = 0.325) or high prior knowledge (OR = 0.64, 95% CI = 0.13 - 2.44, p = 0.545), indicating that active debriefing was equally effective regardless of how much participants knew about Kim Kardashian.

Elon Musk Model

The Elon Musk model was statistically significant ($\chi^2(5, N = 400) = 80.11$, $p < .001$), R^2 (Cox & Snell) = 0.10, R^2 (Nagelkerke) = 0.14, and correctly classified 70.6% of cases. The odds ratios for participants with some prior knowledge (OR = 1.43, 95% CI = 0.83–2.51, p = 0.206) and high prior knowledge (OR = 1.66, 95% CI = 0.92–3.01, p = 0.093) demonstrate that these effects were not statistically significant. This indicates that familiarity with Elon Musk did not meaningfully predict false memory formation in this context. Additionally, there was no significant interaction between some prior knowledge (OR = 0.75, 95% CI = 0.30 - 1.94, p = 0.543), high prior knowledge (OR = 0.81, 95% CI = 0.30 - 2.21, p = 0.672), and the impact of active debriefing.

4.4.2 RQ4 Summary of Findings

The results demonstrate that active debriefing can be a strong intervention for preventing false memories related to deepfake videos of celebrities such as Kim Kardashian and Elon Musk. However, the analysis did not show any significant effect or interaction between prior knowledge and the effectiveness of active debriefing. Therefore, active debriefing can effectively reduce false memories regardless of participants' prior familiarity with the celebrity.

4.5 Research Question 5 (RQ5). Is there a relationship between social media use frequency and the effects of deepfakes implanting false memories?

H_0: There is no significant relationship between social media use frequency and the likelihood of deepfakes implanting false memories.

H_1: There is a significant relationship between social media use frequency and the likelihood of deepfakes implanting false memories.

To examine whether social media usage frequency is related to the effects of deepfakes in implanting false memories, two binomial logistic regression models were conducted. The independent variable was social media usage frequency, which participants self-reported in an open-text numerical format. For analysis, responses were categorized into quartiles: Low Usage (Q1: 0–1 hours/day; n = 97), Moderate Usage (Q2: 2–3 hours/day; n = 330), High Usage (Q3: 4–6 hours/day; n = 304), and Very High Usage (Q4: 7+ hours/day; n = 69). The dependent variable was false memory formation, coded as 0 for participants who did not report any false memories and 1 for those who retained at least one false memory.

This quartile-based categorization provided a structured approach to assessing the potential relationship between social media engagement and susceptibility to deepfake-induced

false memories. The analysis explored whether increased exposure to social media corresponded with higher or lower rates of false memory formation when viewing deepfake content.

4.5.1 Descriptive Statistics

Participants were asked to self-report which social media platforms they use, revealing a diverse range of engagement across various platforms (see Figure 4.9). Participants were also asked to report the number of hours they spend on social media each day to assess whether overall usage influenced memory accuracy or deepfake discernment (see Figure 4.10). Instagram emerged as the most commonly used platform, with 646 participants (80.8%) indicating active use, followed closely by YouTube (602 participants, 75.3%). Discord (457 participants, 57.1%) and TikTok (395 participants, 49.4%) were also frequently reported.

Reddit (273 participants, 34.1%) and X (formerly Twitter) (207 participants, 25.9%) were used by a smaller but still substantial portion of participants. Professional and niche platforms saw lower engagement, with LinkedIn (155 participants, 19.4%), Pinterest (123 participants, 15.4%), and Snapchat (134 participants, 16.8%) ranking in the mid-range. Facebook was reported by only 66 participants (8.3%). A small number of participants (15 participants, 1.9%) selected "Other," with responses including platforms such as Tumblr, WeChat, Xiaohongshu, and other Chinese social media, as well as Rednote. Notably, only three participants (0.4%) indicated that they do not use social media at all.

Figure 4.9

Preferred Social Media Platforms Among Participants

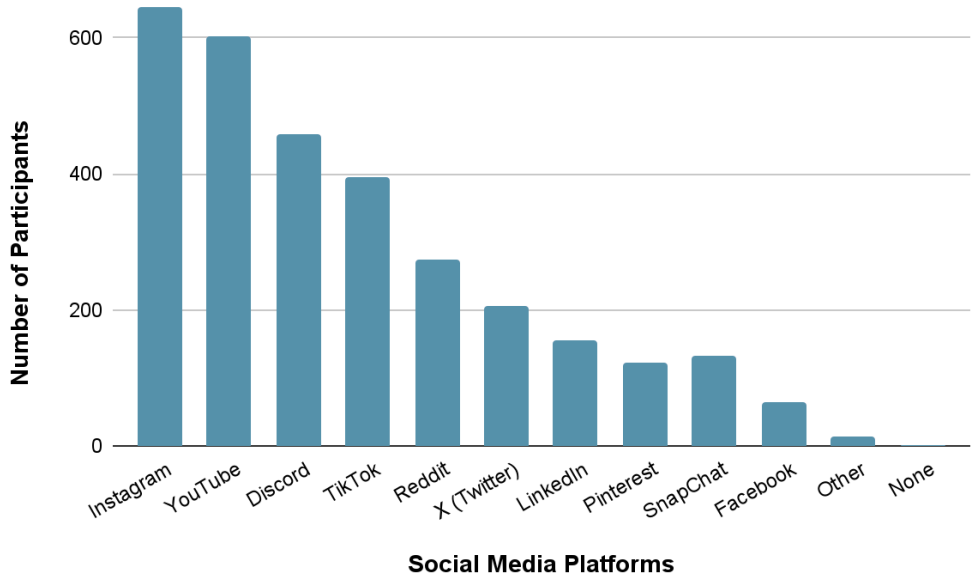
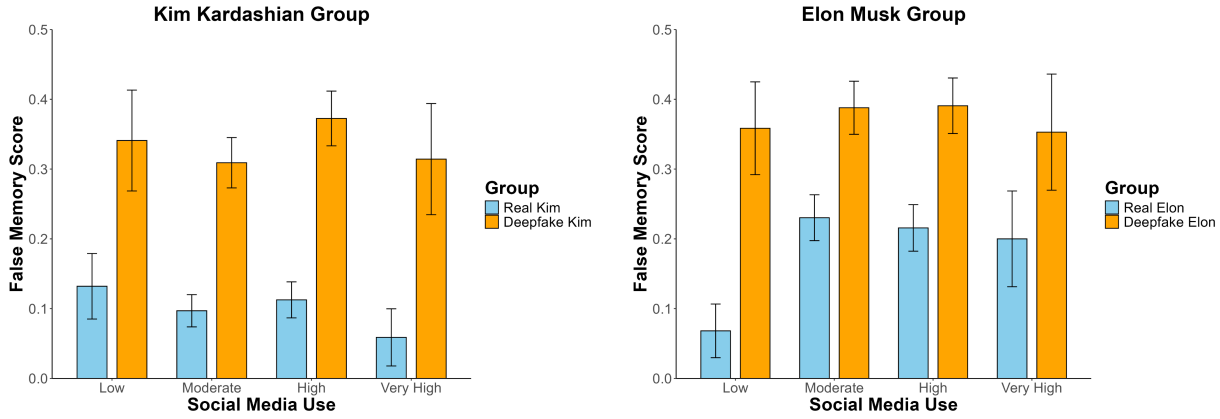


Figure 4.10

False Memory Scores by Social Media Usage (Hours per Day) for Real vs. Deepfake Videos



4.5.2 RQ5 Binary Logistic Regression Results

Kim Kardashian Model

The Kim Kardashian model was statistically significant ($\chi^2(7, N = 397) = 75.26$, $p < .001$), R^2 (Cox & Snell) = 0.09, R^2 (Nagelkerke) = 0.14, and correctly classified 78.4% of cases. Social media usage alone did not significantly predict false memories, as none of the usage categories showed significant effects: moderate (OR = 0.82, 95% CI = 0.30 - 2.62, p = 0.709), high (OR = 1.12, 95% CI = 0.42 - 3.61, p = 0.816), or very high usage (OR = 1.39, 95% CI = 0.36 - 5.39, p = 0.628). Similarly, the interaction terms between social media usage and group membership were not significant: moderate social media use (OR = 1.87, 95% CI = 0.48 - 6.63, p = 0.343), high social media use (OR = 1.33, 95% CI = 0.35 - 4.67, p = 0.659), and very high social media use (OR = 2.67, 95% CI = 0.52 - 13.91, p = 0.237). These results suggest that the frequency of social media usage may not moderate the relationship between deepfake exposure and the formation of false memories.

Elon Musk Model

The Elon Musk model also reached statistical significance, ($\chi^2(7, N = 403) = 41.66$, $p < .001$), R^2 (Cox & Snell) = 0.05, R^2 (Nagelkerke) = 0.07, and correctly classified 70.6% of cases. Social media usage again did not significantly predict false memory formation, and no interaction effects were observed. Again social media usage alone did not significantly predict false memories, as none of the usage categories showed significant effects: moderate (OR = 0.60, 95% CI = 0.31 - 1.17, p = 0.131), high (OR = 1.19, 95% CI = 0.62 - 2.32, p = 0.610), or very high usage (OR = 1.00, 95% CI = 0.41 - 2.40, p = 1.000). Similarly, the interaction terms between social media usage and group membership were not significant: moderate social media use (OR = 2.65, 95% CI = 0.90 - 8.40, p = 0.084), high social media use (OR = 1.44, 95% CI =

0.49 - 4.54, p = 0.519), and very high social media use (OR = 1.22, 95% CI = 0.30 - 5.89, p = 0.703).

4.5.3 Summary of Findings

The results suggest that social media usage frequency alone is not a significant factor in determining susceptibility to false memories implanted by deepfake videos. While social media platforms play a critical role in the spread of misinformation, frequent social media users in this study were not more likely to develop false memories than infrequent users. This finding challenges assumptions that higher exposure to digital media automatically could lead to greater susceptibility to misinformation, suggesting instead that while exposure may be higher, vulnerability is not necessarily increased.

4.6 Research Question 6 (RQ6). Is there a relationship between prior knowledge of deepfakes or previous exposure to deepfakes and the likelihood of false memories being implanted?

H_0: There is no significant relationship between an individual's prior knowledge of deepfakes or previous exposure to deepfake content and the likelihood of false memories being implanted.

H_1: There is a relationship between an individual's prior knowledge of deepfakes or previous exposure to deepfake content and the likelihood of false memories being implanted.

To examine whether prior knowledge of deepfakes or previous exposure to deepfake videos is related to the likelihood of false memory implantation, logistic regression models were conducted for both the Kim Kardashian and Elon Musk conditions. The independent variables included prior knowledge of deepfakes (categorized as little to no knowledge, somewhat

knowledgeable, or highly knowledgeable) and previous exposure to deepfake videos (binary: 0 = no prior exposure, 1 = prior exposure). The dependent variable was false memory formation, coded as 0 for participants who did not report any false memories and 1 for those who retained at least one false memory.

Participants responded to the question, "Prior to this study, did you consider yourself knowledgeable about deepfake technology?" using a likert scale from 0 (Not Knowledgeable at All) to 10 (Extremely Knowledgeable). For analysis, responses were grouped into three categories: participants with little to no understanding of deepfakes (0–3) were coded as 0, those with some familiarity but not deeply informed (4–6) were coded as 1, and those with a strong understanding or expertise in deepfakes (7–10) were coded as 2. Based on this classification, participants were distributed across all three levels, reflecting varying degrees of awareness and familiarity with deepfake technology (see Table 4.9 and Figure 4.11).

Table 4.9

Categorization of Participants' Prior Knowledge of Deepfakes

Prior Knowledge of Deepfakes		
	n	%
Little to No Understanding of Deepfakes	207	25.9
Some Familiarity, But Not Deeply Informed	305	38.1
Strong Understanding or Expertise in Deepfakes	288	36.0

Participants were also asked, "Have you ever looked at deepfake videos before?" with multiple-choice response options: Yes, No, I'm not sure, and I didn't know what a deepfake was. For analysis, responses were consolidated into three categories: participants who reported no

exposure to deepfakes were coded as 0, those who had previously viewed deepfake videos were coded as 1, and those who were unsure or unaware of what a deepfake was were coded as 2.

Based on this classification, participants varied in their prior exposure to deepfakes. Some reported having no prior exposure, while a majority indicated they had seen deepfake videos before. Others were either unsure or unaware of what a deepfake was. This distribution highlights differences in prior experience, which can influence discernment and susceptibility to false memories (see Table 4.10 and Figure 4.12).

Table 4.10

Categorization of Participants' Prior Exposure to Deepfakes

Prior Exposure	to Deepfakes	
	n	%
No: No Exposure to Deepfakes	100	12.5
Yes: Had Exposure to Deepfakes	529	66.1
Unsure: Didn't Know What a Deepfake Was or the Participant Wasn't Sure If They Were Exposed	171	21.4

Figure 4.11

Prior Deepfake Knowledge on False Memory Formation: Real vs. Deepfake (Kim and Elon)

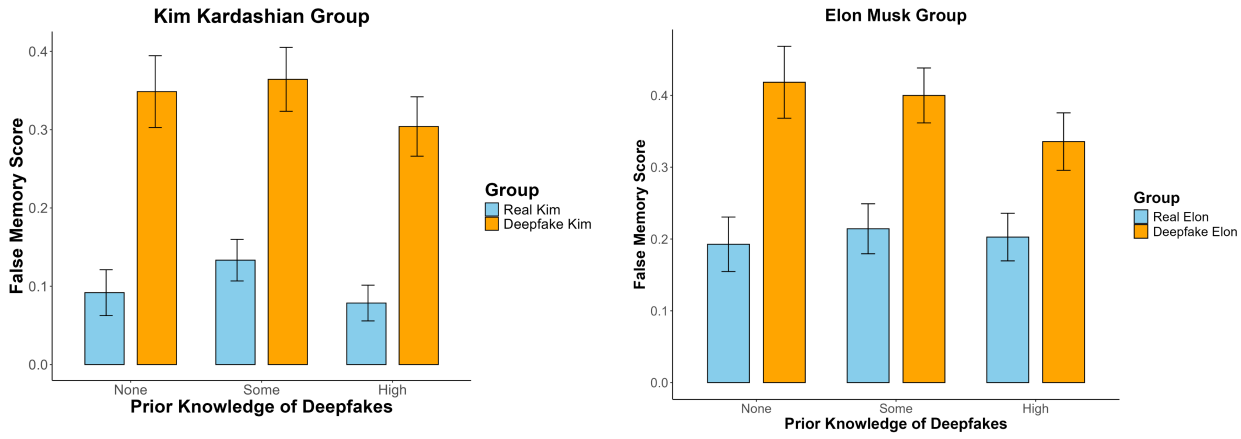
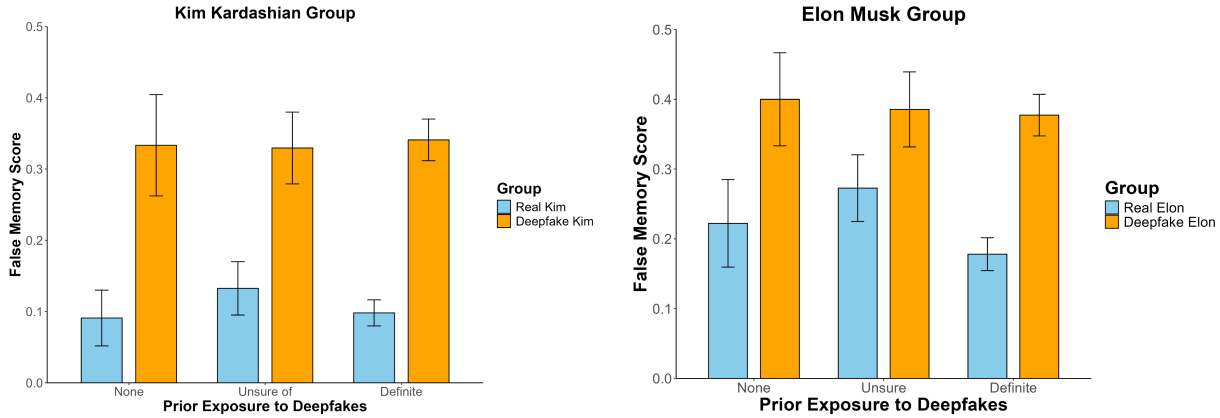


Figure 4.12

Prior Deepfake Exposure on False Memory Formation: Real vs. Deepfake (Kim and Elon)



4.6.1 RQ6 Binary Logistic Regression Results

The deepfake prior knowledge Kim Kardashian model was also statistically significant ($\chi^2(5, N = 397) = 68.67$, $p < .001$), R^2 (Cox & Snell) = 0.08, R^2 (Nagelkerke) = 0.13, and correctly classified 78.0% of cases. In the Kim Kardashian model, neither some prior knowledge (OR = 0.58, 95% CI = 0.25 - 1.37, p = 0.21) nor high prior knowledge (OR = 0.86, 95% CI = 0.40 - 1.92, p = 0.71) reached statistical significance. Additionally, the deepfake prior knowledge Elon Musk model was statistically significant ($\chi^2(5, N = 403) = 32.08$, $p < .001$), R^2 (Cox & Snell) = 0.04, R^2 (Nagelkerke) = 0.06, and correctly classified 70.062% of cases. Similarly, in the Elon Musk model, prior knowledge was not a significant predictor, neither some prior knowledge (OR = 0.81, 95% CI = 0.47 - 1.37, p = 0.43) nor high prior knowledge (OR = 0.93, 95% CI = 0.55 - 1.56, p = 0.77). Overall, the analysis did not reveal a statistically significant relationship between prior knowledge of deepfakes and the likelihood of false memory formation.

For prior exposure to deepfake videos, the results followed the same pattern, with no significant effects detected. The prior exposure Kim Kardashian model was statistically

significant ($\chi^2(5, N = 397) = 66.91$, $p < .001$), R^2 (Cox & Snell) = 0.08, R^2 (Nagelkerke) = 0.12, and correctly classified 78.0% of cases. In the Kim Kardashian model, some prior exposure (OR = 0.92, 95% CI = 0.38 - 2.55, p = 0.854) and definite prior exposure to deepfakes (OR = 0.72, 95% CI = 0.21 - 2.43, p = 0.587) were not significantly related to false memory formation. Additionally, the prior exposure Elon Musk model was statistically significant ($\chi^2(5, N = 403) = 35.92$, $p < .001$), R^2 (Cox & Snell) = 0.08, R^2 (Nagelkerke) = 0.12, and correctly classified 70.62% of cases. The Elon Musk model produced comparable results, with some prior knowledge (OR = 0.91, 95% CI = 0.50 - 1.66, p = 0.753) and definite prior exposure to deepfakes (OR = 0.94, 95% CI = 0.47 - 1.9, p = 0.875) not being significantly related to false memory formation.

4.6.2 Summary of Findings

The findings in this study indicate that neither prior knowledge of deepfakes nor prior exposure to deepfake videos significantly affected the likelihood of forming false memories. While participants varied in their familiarity with deepfake technology and their previous engagement with deepfake content, these factors did not appear to alter their susceptibility to false memories when presented with deepfake stimuli. These results suggest that prior exposure or knowledge may not serve as a protective factor against the implantation of false memories induced by deepfake videos; thus challenging the assumption that greater awareness or experience with deepfakes could improve discernment or reduce false memories.

4.7 Research Question 7 (RQ7). How accurately do people discern whether a video is real or a deepfake?

After completing the initial video assessments, participants were informed of the true purpose of the study and were told that they may have been exposed to deepfakes. Following this

debriefing, they were presented with two real videos and one deepfake video and asked to determine whether each video was real or fake. The independent variable in this analysis was the type of video shown (real vs. deepfake), while the dependent variable was discernment accuracy, recorded as a binary outcome (0 = incorrect classification, 1 = correct classification) based on participants' responses. Responses were recorded using a multiple-choice format, where participants selected either “real” or “fake” for each video.

4.7.1 RQ7 Descriptive Statistics Results

The study examined participants' ability to discern real and deepfake videos across three celebrity conditions: Kim Kardashian, Elon Musk, and Paris Hilton. Participants demonstrated varying levels of accuracy depending on the video (see Table 4.11 and Figure 4.13). While many correctly identified the real (70.9%) and deepfake (81.6%) versions of Kim Kardashian, accuracy was lower for the real Elon Musk video (48.6%), with a higher rate of misclassification. In contrast, the findings demonstrate that participants were more successful in recognizing the deepfake version of Elon Musk (84.6%). For the real Paris Hilton video, accuracy fell (59.5%), with a notable portion of participants misidentifying the real video. These results highlight differences in discernment ability across different celebrity deepfakes.

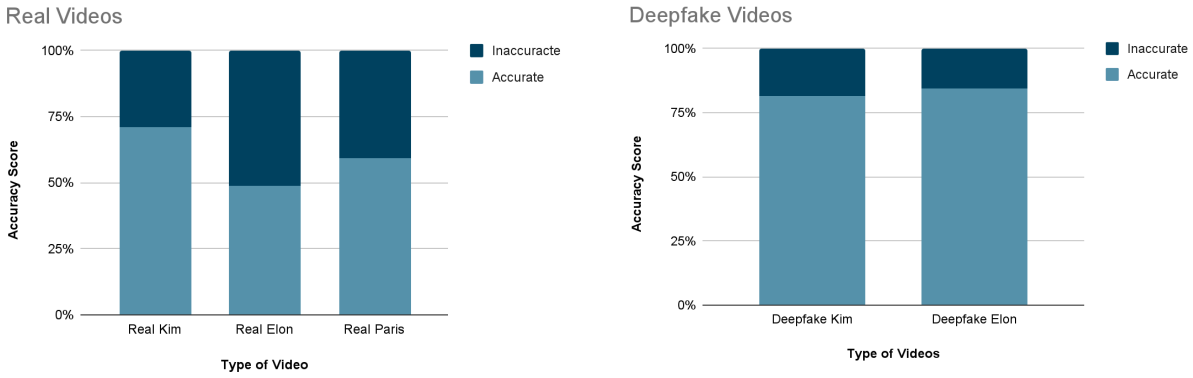
Table 4.11

Participant Accuracy in Identifying Real and Deepfake Videos

Accuracy in Discernment	of Deepfake Kim		of Deepfake Elon		of Real Kim		of Real Elon		of Real Paris	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Accurate	324	81.6	341	84.6	286	70.9	193	48.6	476	59.5
Inaccurate	73	18.4	62	15.4	117	29.1	204	51.4	324	40.5

Figure 4.13

Participant Accuracy in Identifying Real and Deepfake Videos



4.7.2 Summary of Findings

The findings suggest that participants demonstrated stronger discernment abilities when identifying deepfake videos compared to real ones. The deepfake Kim Kardashian and deepfake Elon Musk videos highlighted the highest accuracy rates, suggesting that participants were more adept at recognizing manipulated audiovisual media. However, participants seemed to struggle more with identifying real videos accurately, particularly in the case of the real Elon Musk video, where misclassification was more frequent. This suggests a greater tendency to mistake real content for deepfakes.

These results also highlight potential biases in deepfake detection. While participants were relatively successful at identifying deepfake videos, they struggled more with verifying authentic content, particularly with Elon Musk and Paris Hilton. The lower accuracy in real video identification suggests that exposure to deepfakes or skepticism toward digital media could lead to false positives, where genuine content is mistakenly perceived as artificial.

4.8 Research Question 8 (RQ8). How confident are individuals in their ability to discern real versus deepfake videos, and can their confidence align with their actual accuracy?

H_0: Confidence levels in discernment of real and deepfake videos are not directly correlated with accuracy.

H_1: Confidence levels in discernment of real and deepfake videos are directly correlated with accuracy, meaning individuals with higher confidence are more accurate, while those with lower confidence are less accurate.

This research question examines the relationship between individuals' confidence in their ability to discern real versus deepfake videos and their actual accuracy in doing so. Specifically, it seeks to determine whether participants' self-assessed confidence, the independent variable, aligns with their objective ability to correctly classify deepfake videos, the dependent variable. To evaluate this relationship, a logistic regression analysis was conducted to assess the extent to which confidence levels influenced accuracy.

Participants were shown three video clips drawn from a set of five videos: two deepfakes and three authentic videos featuring well-known celebrities Kim Kardashian, Elon Musk, and Paris Hilton. Depending on their randomly assigned group and condition, all participants ultimately viewed one authentic Paris Hilton video, along with either an authentic or deepfake Kim Kardashian video and either an authentic or deepfake Elon Musk video. This ensured that each participant saw exactly two authentic videos and one deepfake. After viewing each clip, participants were asked to classify the video as real or fake. Their accuracy was recorded as a binary variable, with incorrect classifications coded as 0 and correct classifications coded as 1.

Following their classification, participants rated their confidence on a 0-10 Likert scale in response to the question, "How confident are you in your answer?" where 0 indicated no

confidence at all and 10 indicated extreme confidence. To facilitate analysis, confidence scores were categorized into three distinct levels. Low confidence (0-3) represented individuals who expressed minimal certainty in their ability to differentiate real from deepfake videos. Moderate confidence (4-6) included those who demonstrated some level of confidence but were not entirely certain in their assessments. High confidence (7-10) reflected individuals who strongly believed in their ability to discern real videos from deepfakes.

4.8.1 Descriptive Statistics

The number of participants in each confidence category varied across the five video stimuli, which may have influenced accuracy trends (see Table 4.12).

Table 4.12

Number of Participants by Video Stimuli, Sorted by Confidence Level

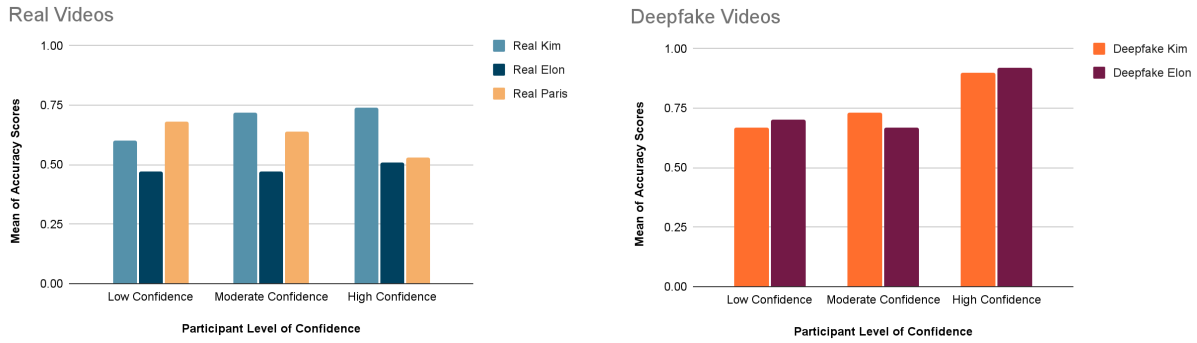
Confidence Level	of Real Kim		Of Real Elon		of Real Paris		of Deepfake Kim		of Deepfake Elon	
	n	%	*n*	%	*n*	%	*n*	%	*n*	%
Low	62	15.4	60	15.1	132	16.5	54	13.4	30	7.4
Moderate	144	35.7	143	36.0	297	37.1	128	31.8	91	22.6
High	197	48.9	194	48.9	371	46.4	221	54.8	282	70.0

These figures illustrate that more participants expressed high confidence when discerning deepfake videos, particularly in identifying manipulated content (see Figure 4.14). More participants who watched the real videos reported low confidence in their answers compared to those who viewed the deepfakes, suggesting they were less certain about recognizing real content. Confidence levels appeared to influence participants' ability to identify deepfakes and real videos, though the relationship was not consistent across all clips. While confidence played a

role, its impact varied depending on which specific real or fake video participants viewed. While higher confidence was often associated with greater accuracy in detecting deepfake videos, it did not always translate to better performance when recognizing real videos, as shown by the relatively low accuracy rates for identifying real content, 48.6% for the Real Elon video and 59.5% for the Real Paris video (see Table 4.11).

Figure 4.14

Accuracy Scores by Level of Confidence, Real vs. Deepfake Videos



4.8.2 RQ8 Binary Logistic Regression Results

A logistic regression analysis was conducted to examine the relationship between confidence and accuracy in discerning real and deepfake videos. The results indicate varying findings across different conditions.

Real Kim Kardashian Video Model

The real Kim Kardashian video model was not statistically significant and only accounted for a small proportion of variance, ($\chi^2(2, N = 403) = 5.40$, $p = .067$), R^2 (Cox & Snell) = 0.01, R^2 (Nagelkerke) = 0.02, and correctly classified 71.0% of cases. Individuals with some confidence did not show a statistically significant effect on accuracy (OR = 0.91, 95% CI = 0.50 - 1.66, p = 0.753); similarly, participants in the highest confidence category also did not significantly impact accuracy (OR = 0.94, 95% CI = 0.47 - 1.90, p = 0.865). The results demonstrated that highly

confident participants viewing the real Kim Kardashin video had a modestly significant association with accuracy.

Deepfake Kim Kardshian Model

The deepfake Kim Kardashian model was statistically significant ($\chi^2(2, N = 397) = 8.31$, $p = .061$), R^2 (Cox & Snell) = 0.02, R^2 (Nagelkerke) = 0.03, and correctly classified 82.4% of cases. In contrast to the real Kim Kardashian video, participants with moderate confidence had a significant negative effect on accuracy (OR = 0.48, 95% CI = 0.25 - 0.89, p = 0.023), while highly confident participants did not significantly differ from the baseline (OR = 1.09, 95% CI = 0.54 - 2.20, p = 0.803). Furthermore, despite reaching statistical significance, the model accounted for only a small proportion of the variance.

Real Elon Musk Video Model

The real Elon Musk video model was not statistically significant ($\chi^2(2, N = 397) = 3.00$, $p = .223$), R^2 (Cox & Snell) = 0.01, R^2 (Nagelkerke) = 0.01, and correctly classified 54.4% of cases. Neither moderate (OR = 1.12, 95% CI = 0.69 - 1.84, p = 0.643) nor high confidence (OR = 1.52, 95% CI = 0.92 - 2.50, p = 0.100) were meaningful predictors, demonstrating that confidence did not significantly predict accuracy here.

Deepfake Elon Musk Model

The Elon Musk deepfake video model was statistically significant ($\chi^2(2, N = 403) = 25.90$, $p < 0.001$), R^2 (Cox & Snell) = 0.06, R^2 (Nagelkerke) = 0.11, and correctly classified 84.6% of cases. Participants with moderate confidence levels showed significantly improved accuracy (OR = 5.38, 95% CI = 2.75 - 10.92, p < 0.001), whereas highly confident participants did not demonstrate a statistically significant effect (OR = 1.60, 95% CI = 0.77 - 3.37, p =

0.208). Thus, for the Elon Musk deepfake video, having at least some confidence was a strong predictor of accuracy.

Real Paris Hilton Video Model

For the real Paris Hilton video, the model fit remained low ($\chi^2(2, N = 800) = 19.84$, $p < 0.001$), R^2 (Cox & Snell) = 0.02, R^2 (Nagelkerke) = 0.3, and correctly classified 59.5% of cases. Highly confident participants were significantly more accurate and statistically significant (OR = 2.07, 95% CI = 1.43 - 3.00, $p < 0.001$), and moderately confident participants demonstrated no statistical significance (OR = 1.10, 95% CI = 0.77 - 1.57, $p = 0.608$). Overall, the model indicates that while the relationship for highly confident viewers was statistically significant, confidence explained only a small fraction of the variance in accuracy.

4.8.3 Summary of Findings

The findings highlight a complex relationship between confidence and accuracy in discerning real and deepfake videos. Moderate levels of confidence significantly predicted accuracy in identifying the Elon Musk deepfake, supporting H_1 in this context. However, confidence can not reliably predict accuracy for authentic videos of Kim Kardashian and Elon Musk, indicating that individuals could overrestimate their ability to recognize genuine content. For real videos, confidence illustrated minimal impact on discernment ability, with participants across different confidence levels performing similarly. However, for deepfake videos, individuals who expressed higher confidence tended to perform better, demonstrating a stronger ability to identify manipulated content. Interestingly, in some cases, confidence may not correlate with improved accuracy. For example, in certain real video conditions, participants with lower confidence performed better than those with high confidence, suggesting the possibility of overconfidence bias. This indicates that while confidence could be a useful predictor of

discernment ability for deepfakes, it may not always translate to better accuracy when assessing real content.

Moreover, the negative association between moderate confidence and accuracy for the Kim Kardashian deepfake suggests a possible hesitancy effect, where moderately confident individuals might second-guess their judgments. In contrast, the pattern observed for the Paris Hilton real video indicates that confidence can, in certain scenarios, enhance accuracy rather than contribute to errors driven by overconfidence. Taken together, these results imply a partial rejection of H_0, as confidence correlates with accuracy under specific conditions but may not consistently predict performance across all video types.

4.9 Research Questions 9 and 10 (RQ 9 & 10): Cues & Clues for Deepfake Discernment

4.9.1 (RQ 9) What cues and clues do individuals use for discernment?

A frequency analysis was conducted to examine the cues and clues individuals used for discernment across five video stimuli: a real Kim Kardashian video, a deepfake Kim Kardashian video, a real Elon Musk video, a deepfake Elon Musk video, and a real Paris Hilton video. Participants were asked to identify whether a video was real or fake and then describe the cues and clues they used to make their decision in an open-text response. Open-text responses were coded at the phrase level, where a phrase was defined as a portion of a sentence, a meaningful string of words, or sometimes even a single word, depending on the context.

To quantify qualitative data, the responses were binary-coded within each theme, assigning 1 if the theme was present and 0 if it was absent. This ensured that each response was categorized based on its interpretative meaning rather than just individual words. Initially, responses amounted to six themes and 24 subthemes. These variables were then consolidated into five overarching categories for analysis purposes: Visual, Audio, Video and Production

Quality, Decision-Making Biases, and Psychological Effects (see Table 4.13 and Figure 4.15).

The visual category included references to observable features such as facial expressions, eye movements, lip synchronization, body movements, and mannerisms, which participants used to assess video authenticity. Audio cues and clues captured less frequently mentioned elements, such as voice clarity, tone, and artifacts. Video and production quality encompassed technical indicators of manipulation, including resolution, glitches, editing seams, and frame rate issues. The decision-making biases category reflected participants' reliance on prior knowledge, expectations, or familiarity with public figures. Lastly, psychological effects included responses related to realism, unnaturalness, and discomfort, particularly references to the "uncanny valley," which influenced authenticity judgments on a perceptual or emotional level.

Table 4.13

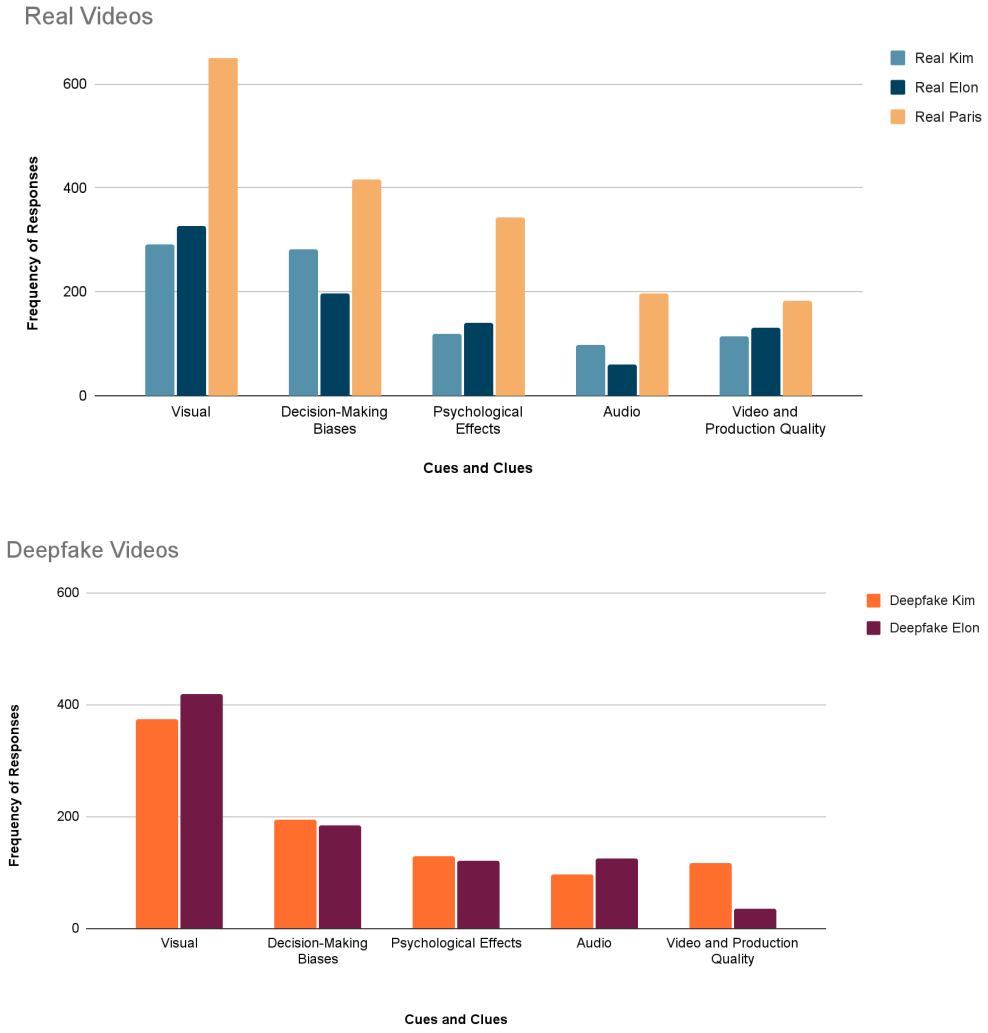
Measures of Central Tendency for Categorical Discernment

Categorical Discernment	of Real Kim	of Real Elon	of Real Paris	of Deepfake Kim	of Deepfake Elon	Mean
Visual	292	326	653	374	419	412.8
Decision-Making Biases	282	196	417	194	184	254.6
Psychological Effects	119	141	342	128	121	170.2
Audio	98	60	198	97	125	115.6
Video and Production Quality	114	131	183	116	34	115.8

Note. The mean represents the average score of the five video stimuli

Figure 4.15

Frequency of Cues and Clues Responses by Category by Video Stimuli



4.9.2 RQ9 Descriptive Statistics Results

Participants primarily relied on visual cues and clues to evaluate video authenticity. This category was most referenced in the real Paris Hilton video, suggesting that visual (643) attributed in affirming authenticity. Similarly, the deepfake Elon Musk video also received considerable attention regarding visual discrepancies (419), highlighting participants' sensitivity to observable inconsistencies even in manipulated content.

Decision-making biases emerged as the second most influential category. This bias was particularly present in responses to the real Paris Hilton (417) and real Kim Kardashian (282) videos, reflecting how familiarity with celebrity personas significantly informed authenticity judgments. Confirmation bias and plausibility of the claim were prominent subthemes, suggesting participants often trusted or doubted authenticity based on their pre-existing beliefs about a celebrity's typical behavior.

Psychological effects was notably present in responses to the real Paris Hilton (342) and real Elon Musk (141) videos, where participants frequently mentioned feelings of unease or noted that the videos seemed "too perfect." The emphasis on the uncanny valley suggests that even genuine content can trigger psychological discomfort, complicating the discernment process.

Audio cues and clues and video and production quality were the least influential categories overall, they had slightly higher relevance in the real Paris Hilton (198) and deepfake Elon Musk (125) videos. Similarly, issues related to production quality were mentioned more frequently in the real Paris Hilton (183) and real Elon Musk (131) videos, while the deepfake Elon Musk (34) video received the fewest mentions, suggesting participants seldom depended on technical inconsistencies alone when stronger visual or cognitive cues were present.

4.9.3 Summary of Findings

The analysis revealed that the frequency with which participants used each cue type across different video stimuli was weighted differently depending on the celebrity and video condition. The findings also suggest that physical and visual may be the most critical factors in determining whether a video is real or fake. Decision-making biases and psychological effects also play a significant role, particularly when participants have prior knowledge of the celebrity.

The results show that audio cues and clues were the least used, reinforcing that visual elements may often dominate the discernment process. Video and production quality are considered, but are also generally a secondary factor. This analysis highlights the complex interplay of perceptual, deception-making, and technical elements in how individuals assess video authenticity.

4.9.4 (RQ 10) Are some cues and clues more or less efficacious for discernment than others?

H_0: There is no significant relationship between the use of discernment cues and an individual's ability to accurately identify false or misleading information.

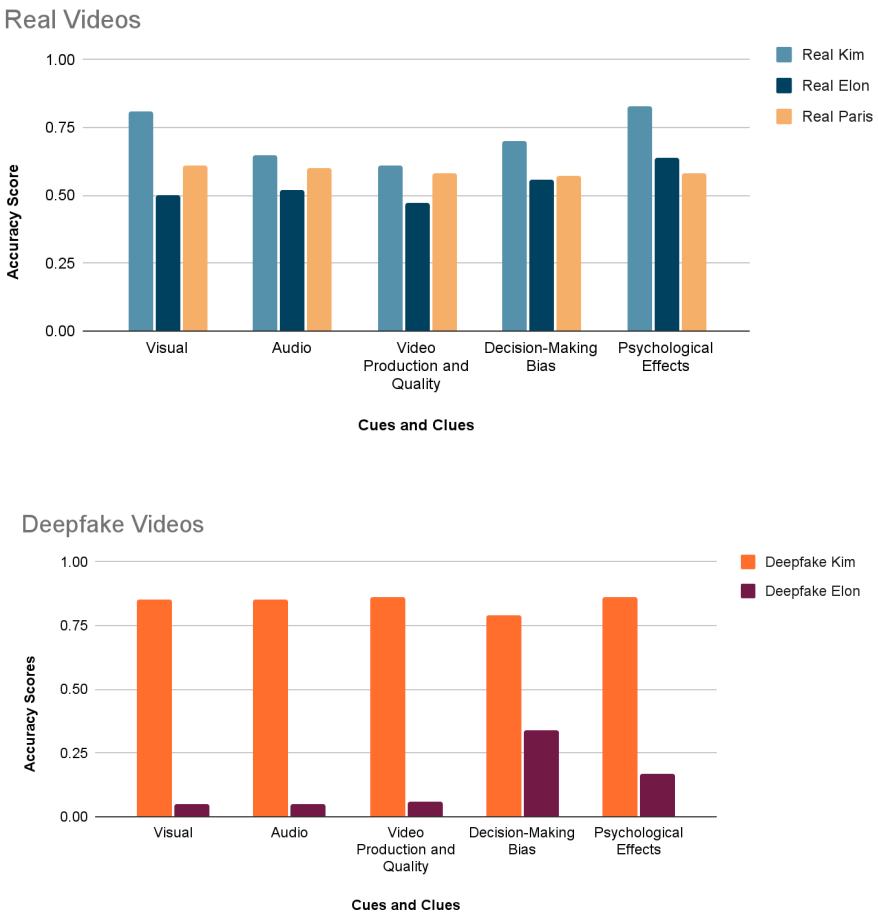
H_1: There is a significant relationship between the use of discernment cues and an individual's ability to accurately identify false or misleading information.

To examine the efficacy of various discernment cues and clues in accurately identifying real versus deepfake videos, the responses for five video stimuli were analyzed: a real Kim Kardashian video, a deepfake Kim Kardashian video, a real Elon Musk video, a deepfake Elon Musk video, and a real Paris Hilton video (see Figure 4.16). To evaluate the role of discernment cues, participant responses were categorized into five overarching groups: Visual, Audio, Video and Production Quality, Decision-Making Biases, and Psychological Effects.

Responses were recorded using a binary coding system, where 0 indicated that a particular cue or clue was not reported, and 1 indicated that the cue or clue was self-reported by the participant. A binary logistic regression was applied to each video condition to determine the relationship between the usage of these discernment cues and participants' accuracy in correctly identifying real and deepfake videos.

Figure 4.16

Accuracy Scores for Discernment Categories Across Video Stimuli



4.9.5 RQ 10 Binary Logistic Regression Results

Real Kim Kardashian Video Model

For the real Kim Kardashian video, the model was statistically significant ($\chi^2(5, N = 403) = 41.84$, $p < 0.001$), R^2 (Cox & Snell) = 0.10, R^2 (Nagelkerke) = 0.14, and correctly classified 73.45% of cases. Psychological effects (OR = 2.91, 95% CI = 1.67 - 5.27, $p < 0.001$) were strong predictors of accuracy, while video and production quality (OR = 0.54, 95% CI = 0.33 - 0.89, $p = 0.016$) and and visual (OR = 0.54, 95% CI = 0.33 - 0.89, $p < 0.001$) had a negative effect,

indicating that reliance on production factors was counterproductive. Audio (OR = 0.67, 95% CI = 0.39 - 1.13, p = 0.13) and decision-making bias (OR = 1.30, 95% CI = 0.79 - 2.15, p = 0.298) demonstrated no statistical significance.

Deepfake Kim Kardashian Model

For the deepfake Kim Kardashian video, the model was not statistically significant ($\chi^2(2, N = 397) = 7.42$, p = .192), R^2 (Cox & Snell) = 0.19, R^2 (Nagelkerke) = 0.03, and correctly classified 82.37% of cases. Furthermore, none of the cues and clues categories significantly predicted accuracy, including visual (OR = 1.66, 95% CI = 0.93 - 2.94, p = 0.084), audio (OR = 1.19, 95% CI = 0.64 - 2.35, p = 0.598), video production and quality (OR = 1.45, 95% CI = 0.79 - 2.79, p = 0.246), decision making biases (OR = 0.91, 95% CI = 0.51 - 1.64, p = 0.750), and psychological effects (OR = 1.40, 95% CI = 0.78 - 2.64, p = 0.274). This indicates that these factors did not meaningfully influence participants' ability to discern real from deepfake content.

Real Elon Musk Video Model

For the real Elon Musk video, the model was statistically significant ($\chi^2(5, N = 397) = 32.49$, p < 0.001), R^2 (Cox & Snell) = 0.08, R^2 (Nagelkerke) = 0.10, and correctly classified 62.97% of cases. Psychological effects (OR = 2.95, 95% CI = 1.89 - 4.66, p < 0.001) and decision-making biases (OR = 2.33, 95% CI = 1.47 - 3.72, p < 0.001) were significant predictors of accuracy, whereas visual (OR = 1.45, 95% CI = 0.92 - 2.29, p = .109), audio (OR = 1.15, 95% CI = 0.64 - 2.05, p = 0.644), and video production and quality (OR = 1.22, 95% CI = 0.76 - 1.95, p = 0.413) did not reach statistical significance.

Deepfake Elon Musk Model

For the deepfake Elon Musk video, the model was statistically significant ($\chi^2(5, N = 403) = 120.45$, p < 0.001), R^2 (Cox & Snell) = 0.26, R^2 (Nagelkerke) = 0.45, and correctly classified

89.08% of cases. However, visual (OR = 0.07, 95% CI = 0.03 - 0.15, $p < 0.001$) and audio (OR = 0.17, 95% CI = 0.06 - 0.42, $p < 0.001$) negatively predicted accuracy, meaning reliance on these cues led to incorrect discernment. Conversely, decision-making bias (OR = 2.46, 95% CI = 1.10 - 5.43, $p = 0.026$) had a positive effect. Video production and quality (OR = 0.52, 95% CI = 0.07 - 2.30, $p = 0.447$) and psychological effects (OR = 1.66, 95% CI = 0.79 - 3.49, $p = 0.177$) demonstrated no statistical significance in accuracy.

Real Paris Hilton Video Model

For the real Paris Hilton video, the model was statistically significant ($\chi^2(5, N = 800) = 51.24$, $p < 0.001$), R^2 (Cox & Snell) = 0.06, R^2 (Nagelkerke) = 0.08, and correctly classified 63.9% of cases, indicating that discernment cues had a measurable impact on accuracy. Psychological effects (OR = 2.32, 95% CI = 1.71 - 3.17, $p < 0.001$), visual (OR = 1.79, 95% CI = 1.30 - 2.48, $p < 0.001$) were the strongest predictors of accuracy, while video production and quality (OR = 1.61, 95% CI = 1.11 - 2.36, $p = 0.014$) and audio (OR = 1.45, 95% CI = 1.02 - 2.08, $p = 0.040$)demonstrated. some statistical significance. Decision-making biases (OR = 1.33, 95% CI = 0.95 - 1.87, $p = 0.095$) had no statistical significance.

4.9.6 Summary of Findings

The results suggest that certain cues and clues are more effective than others in identifying real versus deepfake videos. Psychological effects and visual cues and clues were the most consistent predictors of accuracy across multiple videos, particularly for real videos (e.g., the real Paris Hilton video, the real Kim Kardashian video, and the real Elon Musk video). Decision-making biases also played a role, particularly in the real Elon Musk video. However, reliance on audio cues and clues or video production quality often led to incorrect discernment, especially for deepfake videos. In some specific videos, visual cues also led participants to

incorrect discernment. Overall, these findings support the hypothesis (H_1) that some cues and clues can significantly impact accuracy, with some cues being more efficacious than others.

4.10 Research Question 11 (RQ 11). How do individuals feel about deepfakes after being exposed to information about them?

To examine how exposure to information about deepfakes influences individuals' emotions, attitudes, and perceptions, this study employed a descriptive analysis utilizing frequency analysis and sentiment analysis. Participants were asked, "After completing this survey, how do you feel about deepfakes?" and provided an open-text response. To systematically quantify the qualitative data, responses were binary-coded into themes and subthemes, assigning a value of 1 if a theme was present and 0 if it was absent (see Table 4.14). Open-text responses were coded at the phrase level, where a phrase was defined as a portion of a sentence, a string of meaningful words, or, in some cases, a single word, depending on the context.

Table 4.14

Measures of Central Tendency for Categorical Responses

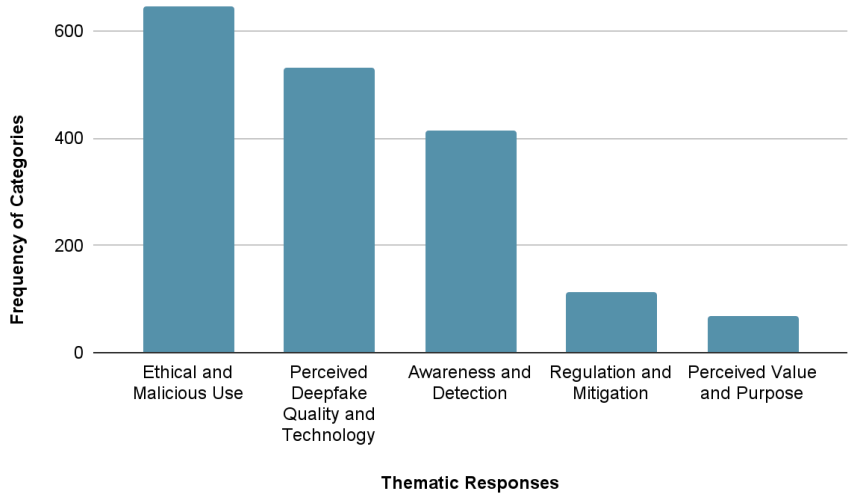
Categorical Responses	Total Frequency
Ethical and Malicious Use	647
Perceived Deepfake Quality and Technology	531
Awareness and Detection	414
Regulation and Mitigation	113
Perceived Value and Purpose	67

4.10.1 RQ 11 Frequency Analysis Thematic Results

The frequency analysis of participants' responses revealed five overarching themes regarding their emotions, attitudes, and perceptions of deepfakes (see Figure 4.17). The most frequently reported theme was ethical and malicious use (647 mentions). Subthemes reflected widespread concerns about the potential for deepfakes to facilitate consumer deception, spread misinformation, enable malicious intent and manipulation, cause reputational harm, support fraud and scams, contribute to sexual exploitation, pose societal threats, and foster perceived trickery. Perceived deepfake quality and technology (531 mentions) represented responses related to deepfakes advancing in quality, realistic-looking deepfakes, ease of creation, underwhelming quality of deepfakes, technological advancements, and the use of AI. Awareness and Detection (414 mentions) encompassed subthemes such as prior knowledge of deepfakes, confidence in detection, and challenges in discerning authenticity. Regulation and mitigation (113 mentions) included discussions about government regulation, legal implications, disclosure requirements, and prevention systems. The least frequently mentioned theme was perceived value and purpose (67 mentions), which collapsed responses about whether deepfakes serve a meaningful function or are primarily used for entertainment. Each theme was categorized based on coded responses, allowing for a structured assessment of participants’ emotions, attitudes, and perceptions.

Figure 4.17

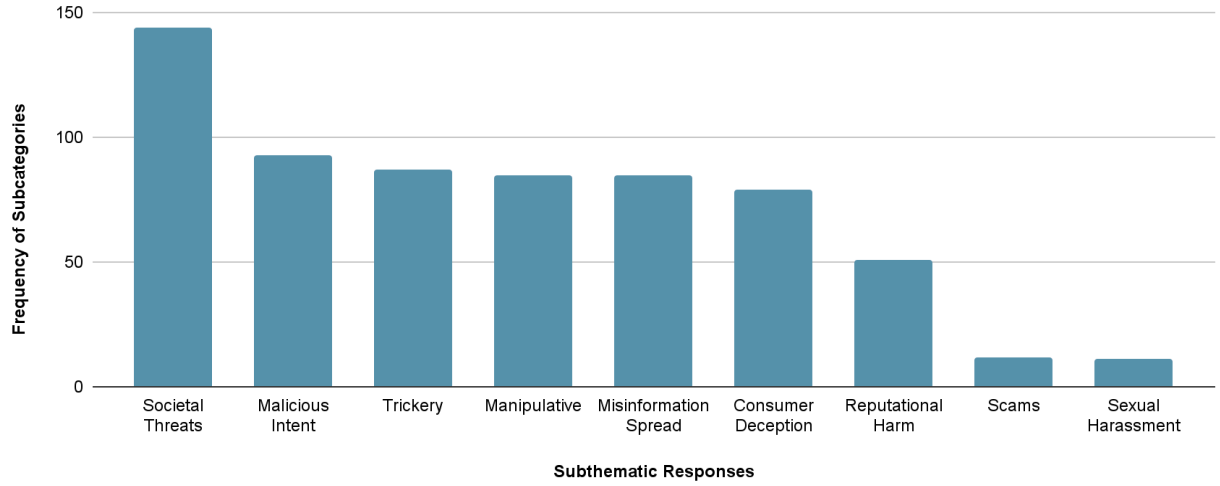
Frequency of Overarching Thematic Responses by Category



The ethical and malicious use theme had the highest frequency of mentions, indicating that deepfake-related risks, such as misinformation, deception, and reputational harm, are major concerns for participants (see Figure 4.18). Many responses described deepfakes as a societal threat (144 mentions) and a tool for malicious intent (93 mentions), emphasizing their potential for exploitation. Participants frequently associated deepfakes with deception and manipulation (e.g., consumer deception, trickery, and manipulative intent) as well as misinformation and harm (e.g., misinformation spread and reputational damage). Additionally, some responses pointed to exploitation and criminal use, with mentions of scams (12) and sexual harassment (11), indicating concerns that deepfakes facilitate harm beyond misinformation alone. Notably, some participants acknowledged that had the study not disclosed the existence of deepfakes, they might have been more vulnerable to manipulation, further reinforcing the perception of deepfakes as a deceptive and potentially dangerous technology.

Figure 4.18

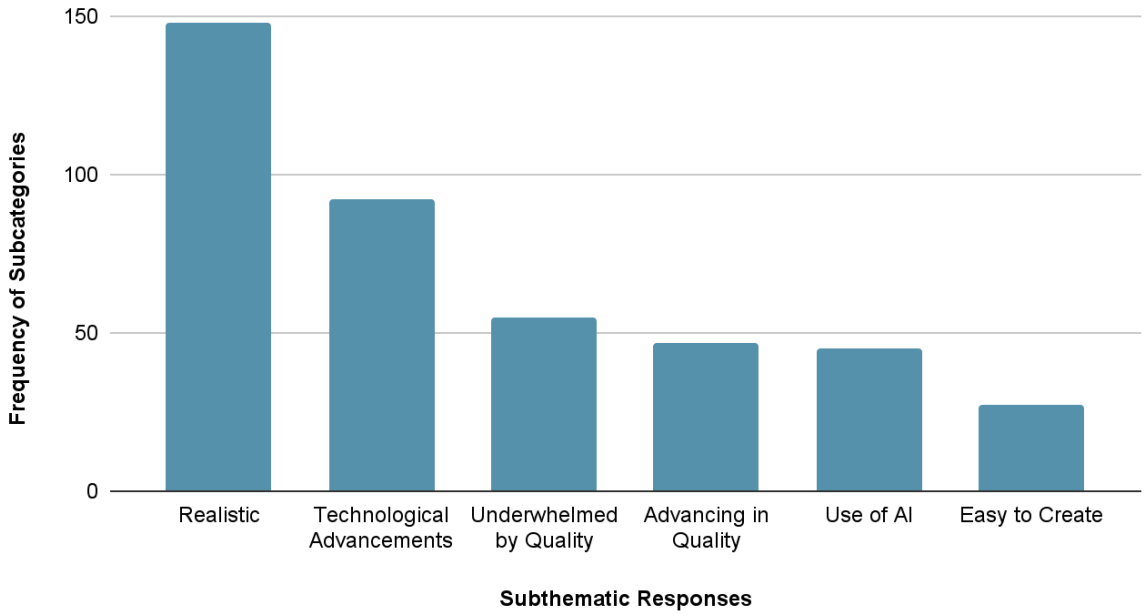
Frequency of Subthematic Responses for Ethical and Malicious Use



The frequency analysis of responses within the perceived deepfake quality & technology theme highlights mixed perceptions regarding the evolution and effectiveness of deepfake technology (see Figure 4.19). Many participants (148 mentions) described deepfakes as realistic, with technological advancements (92 mentions) and advancing in quality (47 mentions) reinforcing the perception that deepfakes are becoming increasingly convincing. However, some respondents (55 mentions) expressed skepticism, stating they were underwhelmed by the quality, suggesting that while some deepfakes appear highly realistic, other deepfakes still fail to meet expectations. Other discussed aspects included AI use (45 mentions) and the ease of creating deepfakes (27 mentions), highlighting a growing awareness of how AI-driven tools have made deepfake generation more accessible. The strong emphasis on AI suggests that participants view it as both a key driver of technological advancements and a major factor in increasing deepfake realism.

Figure 4.19

Frequency of Subthematic Responses for Perceived Deepfake Quality and Technology

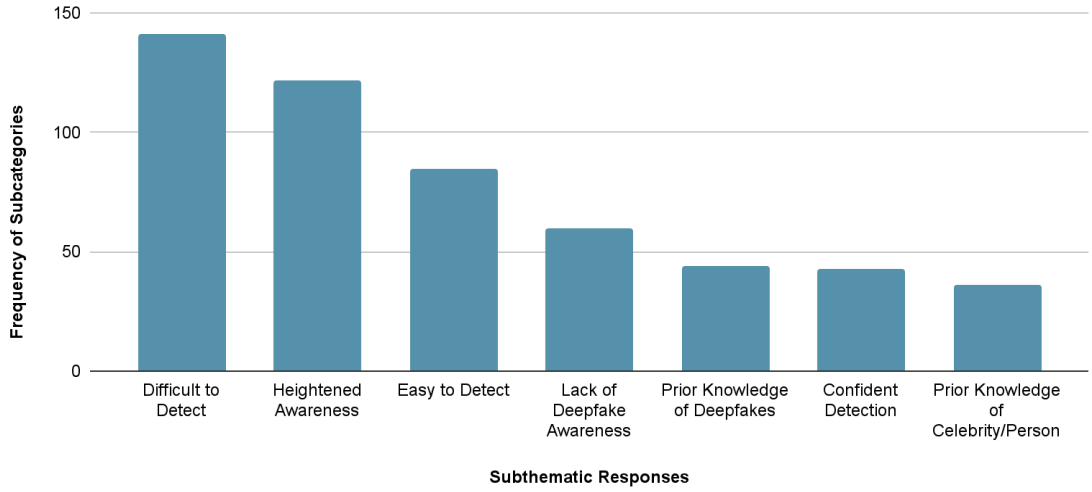


The awareness and detection theme included participants expressing a variety of perspectives on their ability to identify deepfakes (see Figure 4.20). Many found deepfakes difficult to detect (141 mentions), highlighting concerns about their potential for misuse, while others reported heightened awareness (122 mentions), often because the study itself made them conscious of deepfakes by disclosing their presence. Some participants noted that had they not been informed that fake videos were included, they might never have realized their existence, suggesting that heightened awareness influenced detection ability and may have led to lower false memory scores than if deepfakes had been undisclosed. Others felt deepfakes were easy to detect (85 mentions), reflecting a divide in digital literacy and detection skills. Additionally, some participants demonstrated detection confidence (43 mentions), whereas others lacked awareness of deepfakes altogether (60 mentions), reinforcing discrepancies in familiarity with the technology. Prior knowledge of deepfakes (44 mentions) and prior knowledge of the

celebrity/person (36 mentions) also appeared to influence how individuals assessed authenticity, indicating that exposure to deepfake-related content and familiarity with public figures shape perceptions. The contrast between confident and struggling participants suggests that deepfake literacy varies widely.

Figure 4.20

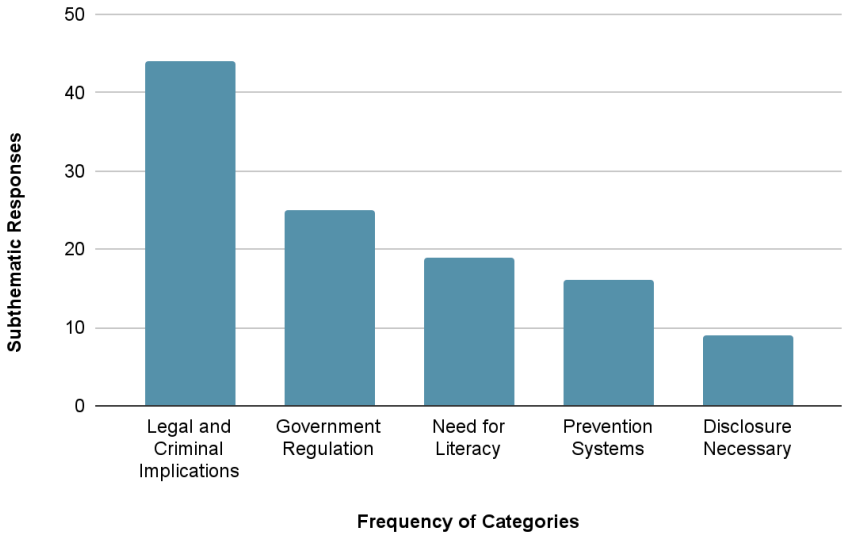
Frequency of Subthematic Responses for Awareness and Detection



The regulation and mitigation theme (see Figure 4.21) was mentioned less frequently than other concerns, suggesting that while participants recognize the need for oversight, it is not their primary focus. Mentions of legal and criminal implications (44) and government regulation (25) indicate that some see policy intervention as necessary. The need for literacy (19) and prevention systems (16) suggests that some participants view education and proactive measures as critical for mitigating deepfake-related risks. Additionally, a small number of participants emphasized the importance of transparency and disclosure (9), highlighting a belief that clear labeling and awareness initiatives could help address concerns. The relatively low discussion of regulation may suggest that participants either lack familiarity with legal frameworks or do not see regulation as an immediate or sufficient solution to deepfake-related challenges.

Figure 4.21

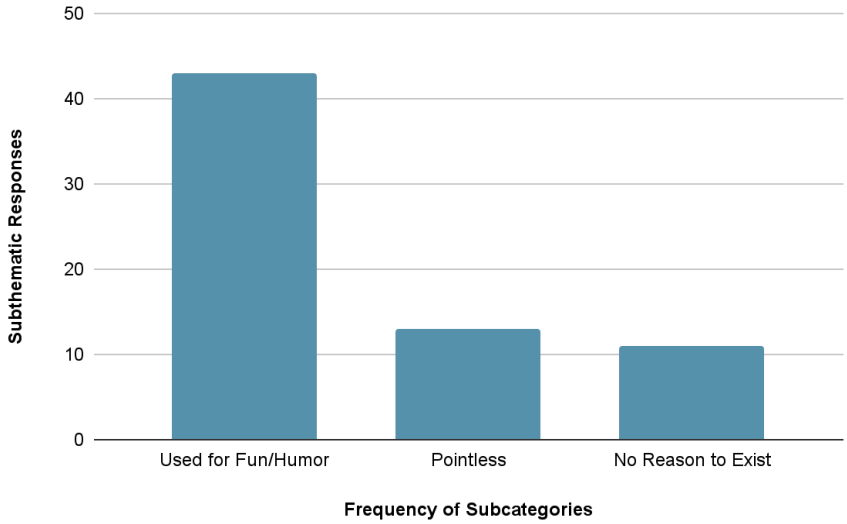
Frequency of Subthematic Responses for Regulation and Mitigation



The perceived value and purpose theme revealed a divide in how participants view deepfakes (see Figure 4.22). While some expressed a negative perception of their usefulness, stating they are pointless (13 mentions) or have no reason to exist (11 mentions), others acknowledged their role in entertainment and creativity, with fun/humor (43 mentions) being the most frequently mentioned subtheme. This suggests that while many participants see deepfakes as unnecessary or even problematic, a smaller group recognizes their potential for creative expression.

Figure 4.22

Frequency of Subthematic Responses for Perceived Values and Purpose



4.10.2 Summary of Frequency Analysis Thematic Findings

Results show that perceptions of deepfakes are seemingly shaped by concerns about misinformation, deception, and ethical risks, with many participants highlighting their potential for manipulation and societal harm. While some individuals might feel confident in detecting deepfakes, many struggle, suggesting that awareness may not always translate into effective discernment. Deepfake quality is widely perceived as improving, with advancements in AI making them more realistic and easier to create. However, some participants remain skeptical, finding them underwhelming or inconsistent in quality.

Ethical and malicious use is the most frequently discussed concern, with participants emphasizing deception, misinformation, and reputational harm. Regulation and mitigation efforts receive less attention, though there is broad support for mandatory disclosure and legal enforcement. The perceived value of deepfakes remains divided—while a small group sees them

as tools for entertainment or parody, most view them as unnecessary or harmful. Overall, concerns about misuse and deception could outweigh any perceived benefits, reinforcing the idea that deepfakes may be more associated with risks than with creative or practical applications.

4.10.3 RQ 11 Emotion and Sentiment Analysis Results

Participants' emotional responses to deepfakes were predominantly negative (80.4%) (see Table 4.15 and Figure 4.23). The frequency analysis indicated that the most common negative reaction was "Scary" (310 mentions), followed by perceptions of deepfakes as "Dangerous" (134 mentions), "Bad" (88 mentions), and explicit expressions of "Dislike" (82 mentions). Participants also frequently described deepfakes as "Weird" (56 mentions), "Creepy" (58 mentions), "Uncomfortable" (48 mentions), and expressed feeling "Worried" (44 mentions) about deepfakes. Positive emotional responses occurred less frequently but included labeling deepfakes as "Interesting" (30 mentions), feeling "Impressed" (28 mentions), considering them "Cool" (26 mentions), or explicitly describing them as "Good" (14 mentions). Neutral or ambivalent responses were also present, with "Unchanged Perception" (114 mentions) representing the majority of neutral reactions, alongside less frequent occurrences of being "Surprised" (24 mentions) or "Curious" (10 mentions).

Table 4.15

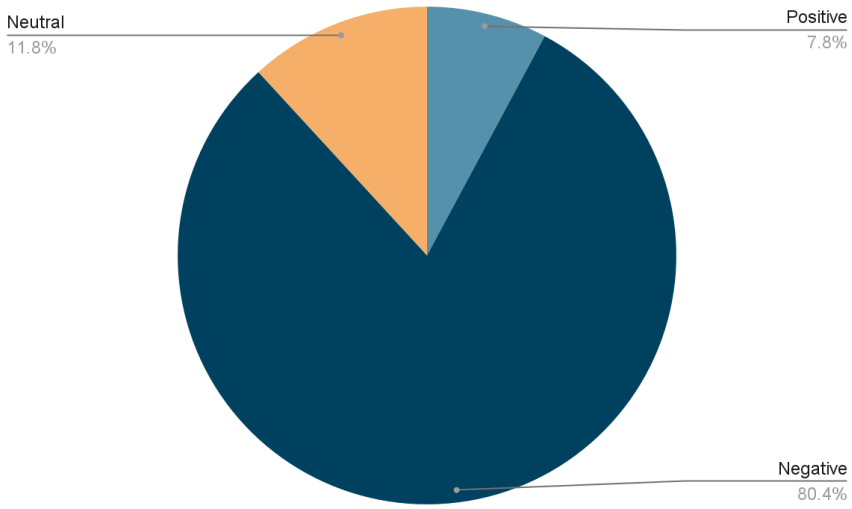
Frequency of Specific Emotional Reactions Grouped by Valence

Emotions and Sentiments	Positive	Negative	Neutral
Alarming	0	16	0
Annoying	0	6	0
Bad	0	88	0
Concerned	0	62	0
Cool	26	0	0

Creepy	0	58	0
Curious	0	0	10
Dangerous	0	134	0
Disgust	0	10	0
Dislike	0	82	0
Disturbing	0	22	0
Good	14	0	0
Impressed	28	0	0
Interesting	30	0	0
Sad	0	4	0
Scary	0	310	0
Shocking	0	30	0
Surprised	0	0	24
Unchanged Perception	0	0	114
Uncomfortable	0	48	0
Unsettling	0	36	0
Weird	0	56	0
Worried	0	44	0
Total	98	1,006	148

Figure 4.23

Emotional Reactions to Deepfakes (*n*=1252)



Note. Pie chart showing the distribution of valences (Positive: 98, Negative: 1006, Neutral: 148) among participants exposed to deepfake content.

4.10.4 Summary of Emotion and Sentiment Analysis

The analysis results demonstrate that exposure to deepfake content may generate predominantly strong, negative emotional reactions, suggesting widespread apprehension, unease, and perceived threat among participants. The significant frequency of negative emotions potentially indicates a clear societal concern toward the implications of deepfake technology, the future of deepfake distribution, and the possible threat of deepfake videos. Neutral responses suggest that many individuals remain indifferent or unaffected, potentially indicating familiarity, desensitization, or a perception that deepfakes do not pose a meaningful threat, or are easily identifiable. Although positive reactions such as intrigue or admiration exist, their limited presence suggests that positive engagement with deepfakes is comparatively rare and is largely overshadowed by societal discomfort and anxiety about their potential impacts.

4.11 Summary of Experimental Outcomes

This study examined whether exposure to deepfake videos of Kim Kardashian or Elon Musk could implant false memories, how well participants could distinguish real from fake videos, and their emotional reactions to deepfakes. Logistic regression results showed that deepfake exposure may significantly increase false memory formation, while active debriefing afterward effectively reduced it. However, neither prior knowledge of celebrities, deepfakes, or social media use mitigated false memory formation in this study.

When assessing participants' abilities to detect deepfakes, results indicated people were generally good at recognizing deepfake videos yet struggled to identify authentic videos correctly more frequently. This finding implies a possible skepticism toward digital media (Li, 2025), where genuine content is mistakenly assumed to be manipulated when told fake content may be present. Additionally, higher confidence levels improved accuracy in detecting certain deepfakes but occasionally led participants to incorrectly classify real videos, showcasing a complicated relationship between confidence and discernment accuracy. Participants primarily relied on visual and psychological cues, such as facial expressions, body language, and gut feelings of unnaturalness, to judge authenticity. However, reliance on audio and video quality was less effective and sometimes misled participants.

Finally, the frequency analysis of emotional responses revealed overwhelmingly adverse reactions toward deepfakes. Participants most commonly described deepfakes as "scary," "dangerous," and "disturbing," indicating widespread unease and anxiety about the potential misuse of generative AI technology (Omeje et al., 2024). Although some expressed curiosity or admiration, positive reactions were uncommon. Neutral reactions suggested that some participants felt unaffected or viewed deepfakes as relatively harmless or easy to detect

(Kietzmann et al., 2020). Overall, the study highlights individuals' vulnerability and skepticism regarding deepfake technology, emphasizing the critical importance of education, awareness, and explicit disclosure in mitigating its harmful impacts.

Chapter 5

Discussion of Results

This study aimed to examine whether or not deepfakes contribute to the formation of false memories, specifically among the Gen Z population of adults immersed in digitally saturated ecosystems. The existing literature suggests that while deepfakes are known to be deceptive and potentially harmful (Romero Moreno, 2024), their psychological effects have remained largely uninvestigated. This study consisted of controlled exposure to deepfake videos featuring Kim Kardashian and Elon Musk, followed by instruments requiring participants to complete false memory and discernment tasks while assessing the effects of debriefing and confidence in accurately detecting true and false video content. Within this chapter, the findings are discussed through sociotechnical lenses of media psychology (Reeves et al., 2015), cognitive science (Uddin, 2019), and digital visual culture (Bentkowska-Kafel, 2009) to demonstrate how artificial intelligence-generated content can affect memory systems and the critical faculties of the end-user (Ramos Vallecillo & Murillo-Ligorred, 2025).

5.1 Deepfake Exposure and the Formation of False Memories

The participants in this study demonstrated that exposure to deepfakes can significantly increase the effects of false memory formation, especially in the absence of active debriefing. Reinforcing increasing concerns that deepfakes may present reputational injury, misrepresent people, and can be morally wrong (De Ruiter, 2021), they can also actively reshape memories or

create new ones that did not exist before. Notable literature has already demonstrated how the Source Monitoring Framework and gist-based reasoning outlined in Fuzzy-Trace Theory explain how end-users can mistakenly view falsified video content as real when the content or narrative aligns with their preexisting beliefs (Reyna & Lloyd, 1997). Based on the results, deepfakes may exploit key cognitive vulnerabilities, precisely source monitoring error (American Association for Research into Nervous and Mental Diseases & Johnson, 1997) and plausibility heuristics (Urueña, 2019).

Deepfake viewers on social media may not know the origin or source of deepfake content, confusing it with real experiences due to its realistic cues (Johnson, Hashtroudi, & Lindsay, 1993). If the content feels believable or matches preconceived notions, people could accept it as true without questioning it (Tversky & Kahneman, 1974), especially if the content is represented as real. Previous research has indicated that emotions have a significant role in forming false memories. Emotionally charged responses to content increased memory distortions and moods shaped memory accuracy (Bookbiner & Brainerd, 2016). Thus, deepfakes might be incredibly manipulative in fast-paced or emotionally charged environments (Lewandowsky et al., 2017).

However, in this study, false memory formation was not uniform, as the Elon Musk deepfake was more effective at false memory implantation than the Kim Kardashian version. The findings suggest that context, preconceived notions, cognitive bias, celebrity personas, or perceived credibility of content (Ahmed, 2023) could affect how believable or memorable a deepfake video's content is. In broader societal landscapes, Elon Musk's association with technology, science, or leadership may be seen as more relevant to current events, thus potentially increasing susceptibility to false information. In contrast, Kim Kardashian's brand

association with fashion companies, being a reality television personality, and business entrepreneur may not be as well known. Thus, this dissertation highlights how public perceptions of celebrity personas and cultural significance might shape our susceptibility to manipulated media (Cashmore, 2023).

Active debriefing substantially combatted false memory formation, with results reporting significantly lower rates of false memories when participants were told about the presence of deception and deepfakes. This suggests that successful intervention, transparent disclosure, and basic knowledge of deepfakes can protect vulnerable viewers from implanted false memories and misinformation. Additionally, these findings highlight the importance of intervention strategies, deepfake detection, and media literacy in reducing the harmful or negative effects of deepfakes, an outcome aligned with previous work on cognitive inoculation and the misinformation effect (Lewandowsky et al., 2012; van der Linden et al., 2017).

These findings carry important implications for both social media platforms and educators. In turn, private and public sector stakeholders need to implement proactive disclosure strategies, such as visible watermarks, disclaimers, deepfake detection technologies, or clear labeling of videos (Akinyemi et al., 2024). In the future, these measures may be essential in mitigating the psychological impact of deepfakes and the spread of misinformation. By alerting viewers to the false nature of video content, these interventions could possibly help viewers build resilience in identifying source materials (Nenovski et al., 2023) and employ skepticism when watching videos (Hameleers et al., 2024). For misinformation and false memory scholars, this study provides empirical support for the argument that deepfakes are not merely deceptive, but that exposure to deepfakes might implant false memories. Even after a deepfake is revealed as

false, the potential for lasting belief change poses significant risks to democratic processes, public health communication, and overall trust in institutions.

5.2 How Prior Knowledge and Social Media Use Fall Short in Countering Deepfakes

Contrary to the study's hypotheses, the findings demonstrated that prior knowledge of celebrities, prior knowledge of deepfake technology, prior exposure to deepfake videos, and social media use frequency were not able to reliably reduce susceptibility or mitigate the implantation of false memories. Instead, participants with some or high prior knowledge of celebrities, especially when viewing the Elon Musk deepfake, were associated with an increased outcome of false memory formation. Although the impact was minimal, the research implies that individuals more familiar with a public figure are more inclined to accept deepfake content as plausible or in line with that individual's persona. Such findings may reflect a bias vulnerability (Ross et al., 2021) rooted in the plausibility of the content (Dias et al., 2020), where video content that "sounds or looks right" becomes more readily accepted as truth, especially when the familiarity of the subject gives users a false sense of discernment (Pennycook & Rand, 2021).

Similarly, prior knowledge of deepfake content and prior exposure to deepfake videos were not able to demonstrate any effect on false memory formation. Self-reported expertise and awareness appear to be poor indicators of critical thinking and resistance to misinformation, potentially due to overconfidence and cognitive biases. Through self-reported expertise, one could assume that awareness of deepfake technology, videos, or creation is sufficient to guard against the formation of false memories (Lewandowsky & Van Der Linden, 2021). Instead, prior knowledge did not foster critical awareness and may not be reliable in the future. Instead, the research infers that repeated exposure to deepfakes and falsified media may normalize such types

of content, desensitize viewers (LeBlanc III, 2019), and diminish the perceived threat when viewing deepfakes.

Finally, the research started with the assumption that the frequency of social media use correlated with digital nativity and higher media literacy among Gen Z participants (Li et al., 2021). However, despite their technical literacy, the results did not conclusively illustrate greater resistance to misinformation. While Gen Z is frequently categorized as "digital natives" from a tech-savvy generation (Soysal & Calli, 2019), this dissertation's results challenge the literature's assumptions and demonstrates the limitations in what Hargittai and Hinnant (2008) calls the Digital Natives Fallacy, where technical fluency is mistaken for critical literacy and overconfidence in the population's abilities to deploy critical evaluation skills online. The concern is further heightened by Gen Z's large workforce and role in shaping public discourse, social movements, and the voting population (Gomez, 2019).

These findings challenge the idea that technologically literate users or familiarity with AI-generated content and deepfakes protect users against misinformation (Frau-Meigs & Corbu, 2024). Indeed, familiarity with the content, the medium, and the social media platform may not enhance cognitive defenses against misinformation and false memories. The finding has broader implications for trust, governance, and legal ecosystems while highlighting the need to move beyond surface-level media literacy efforts and a deeper understanding of how people view, interpret, and analyze information. Left unanswered, the normalization and desensitization of malicious use of generative AI technologies could weaken public trust in authentic content and undermine societal institutions (Troshani, 2021). As the boundaries for audiovisual content blur between what is real and what is fabricated, the stakes of false memory formation extend beyond individuals and may reach the larger collective society.

5.3 Uncertain Judgments: The Role of Confidence, Accuracy, and Realism in Detecting Deepfakes

Surprisingly, participants were better at detecting the deepfakes than identifying the real videos, often misclassifying real ones as fake, while confidence levels were inconsistently tied to accuracy in discernment. This central finding suggests that users may deploy more skepticism, also described as a "reverse false positive effect," a term used to describe the phenomenon where participants anticipate deception (Forstmeier et al., 2017), which makes them doubt the authenticity of genuine content (Ismail & Latif, 2013). It also highlights how awareness of deepfakes and false media content may reshape how people engage with digital content (Alnaser, 2024).

The study suggests that when participants were informed about the possible use of deception, they did not simply become more discerning. Instead, they scrutinized every video during the discernment instrument process and became even more distrustful of the video content presented (Ternovski et al., 2021), regardless of its authenticity. This distrust demonstrates that increased awareness might not always lead to better outcomes; it can sometimes backfire. The results are significant because, in this case, there was a broader erosion of trust in visual media, which could raise serious concerns about how society responds to real content in high-stakes environments, such as consuming news (Allen et al., 2020), participating in global politics (Dobbler et al., 2021), or legal proceedings with video evidence (Maras & Alexandrou, 2019).

Participants who were overconfident in their ability to detect deepfakes added another layer of complexity, as they often felt highly confident about their incorrect discernments. Participants demonstrated high confidence levels in their judgments when discerning true or false videos, but their confidence did not correspond with accuracy, particularly with fake videos.

Users relied heavily on visual and psychological cues, such as irregular facial movement, lip sync, unnatural eye movement, or intuitive feelings that the video content was "off." As one participant noted, "Lips basically don't match up with the words he's speaking at all," while another remarked, "Only her head and face are moving." These kinds of audiovisual mismatches were repeatedly cited, especially when lips did not align with spoken audio or eyes seemed too fixed. One participant wrote, "Hilton did not move her eyes during the video," and another said, "Her voice kind of sounded fake for a split second… like she glitched."

Additional participants cited cues and clues that were associated with an uncanny valley perspective, and those cues and clues were more influential than technical cues, such as audio quality or production quality, which were not as effective and were sometimes misleading. One participant stated that "the [celebrity's] face looks more like a plastic doll's than a real person's," underscoring how the judgment of realism was tied to subtle expressions and bodily nuance. Whereas, when movements appeared natural or matched known behavior, participants were more likely to believe the video was real, making claims such as "the way she moves and talks seems really natural," and "her movements seemed natural."

This finding suggests a significant shift in how users perceive digital media online, from initially relying on technical assessments to now placing greater emphasis on instinctual judgment. Therefore, the results highlight the immediate need for media literacy tools and programs to move beyond simply teaching society what deepfakes are and what they look like; they must also address the dangers of overconfidence and false intuition. The concept that participants tended to rely more on gut instinct rather than more concrete, measurable cues and clues highlights the need to improve and address the gap in education on digital manipulation (DePaulo & Morris, 2024). One response plainly said, "Her mouth was almost 100% matching

the words that were being stated," illustrating how perceptual alignment, rather than metadata, context, or media analysis, drove perceived authenticity.

Training tools and efforts could prioritize subtle learning about behavioral cues (Cheever & Rokkum, 2015), such as facial expressions or unnatural movement, rather than technical artifacts (Brasel & Gips, 2017), which may be less reliable in helping humans detect manipulation. At the same time, there is always a risk of overcorrection. As social media users and society become more skeptical of digital media, the danger of dismissing trustworthy content increases (Bushey, 2015), further eroding the public's confidence in authentic media (Barclay, 2018), including media content generated from verified sources.

5.4 Emotional and Ethical Responses to Deepfakes: Mistrust, Anxiety, and Perceived Harm

Participants' reactions to deepfakes were incredibly emotional and anxious, with a significant moral discomfort about the implications of deepfake videos. The concerns were not abstract, as many described feeling unsettled and disturbed by the idea that reality can be so easily distorted by generative AI technology. One participant shared, "I really really really don't like them. I don't like the idea of people being able to put words in your mouth, just by having a picture of your face and recordings of your voice. That's incredibly creepy." This answer exemplifies a broader emotional hardship rooted in concerns about identity, manipulation, and a loss of control over one's physical likeness and reputation (Du, 2025).

The reactions from the study are not solely reflections of confusion; they represent deep-seated technological anxiety (Mokyr et al., 2015) and emerging media trauma (Kircaburun et al., 2020). As another participant noted, "It makes me afraid of the content I consume online… I worry about deepfakes that could be created of me and my identity being stolen." Reactions similar to this response illustrate how deepfakes may chip away at one's ability to trust what is

seen and heard, making people question the truth of even the most credible digital content. "Now, even real videos make me question if they're real," one participant admitted, showcasing how deepfakes challenge the basis of belief and truth in visual communication. The absence of trust may affect how people engage with digital media, social platforms, news, and even private messaging (Shahbazi & Bunker, 2024). This points to a broader cultural anxiety that users can no longer rely on their senses to validate what they see online (Villena, 2022).

The ethical concerns stood out when looking at ways deepfakes cause reputational harm, disseminate misinformation, and be used for sexual harassment, areas where deepfakes have already caused real-world harm. "They're creepy and have a super high potential for harm," one participant remarked. Others echoed the theme of harm: "It's very scary that [somebody] can be mimicked so easily… it could be used for downright criminal things," while another added, "It's scary how real they're starting to look. I'm aware they've been used for malicious intent."

By comparison, only a few participants raised concerns about regulatory standards (Van der Sloot, 2022), preventative measures (Seng et al., 2024), and current concrete solutions by government institutions and policymakers (Ball, 2024) to mitigate these issues. While one participant suggested, "There needs to be a warning or label on AI-generated videos," most participants focused more heavily on emotional and ethical concerns than on proposed remedies. Revealing a significant gap between awareness of harm and actionable solutions, another participant reported that, "There should be more awareness on deepfake videos to educate people on what to believe and what not to believe," further demonstrating a need for private sector companies and governing bodies to enable media literacy and policy intervention.

The disconnect between public perception and institutions suggests that fear of deepfakes, without clear guidance or reasonable regulation, may undermine trust in legitimate

sources of information (Etienne, 2021). "I think deepfake videos are a serious threat to the credibility of news and information sources," one participant claimed. The decline of trust in the news affects more than individual perception (Fletcher & Park, 2017), as it has the potential to destabilize democratic institutions and contribute to environments where misinformation thrives. As affective computing theory (Picard, 2000) suggests, emotional responses are not simply side effects of "doom-scrolling" on social media; they shape belief, judgment, and behavior. If the emotional landscape remains saturated with fear and betrayal, democratic institutions and journalistic credibility could suffer lasting damage.

While generative AI, particularly deepfakes, is seemingly often associated with deception and harm, the technology also holds meaningful potential for innovation, creativity, and entertainment (Babanikos et al., 2021). However, public fear and anxiety often overshadow these possibilities, shaping a narrative that focuses more on risk than opportunity. One noted, "It's a blessing and a curse. As widespread as information is, so is misinformation." At the same time, another shared, "They're fascinating but shouldn't be abused." For instance, the participants who focused their attitudes and perceptions on the potential for scams, trickery, and harm may cloud the technology's potential for constructive uses, such as applications for entertainment and humor (Smith & Hutson, 2024).

These nuanced perspectives reveal a polarizing mix of awe and apprehension. While the technology is impressive, the perceived risks often feel overwhelming. "Cool concept, but when it comes to stealing someone's likeness without consent, it's deeply disturbing," a participant concluded. One example of positive deepfake use includes generating AI-generated characters utilizing realistic, consensual renderings of faces, voices, and human-like behavior to integrate into different industries such as privacy, telecommunication, education, art, film, and therapy

(Danry, 2022). However, the dominance of fear-based narratives limits broader public engagement with these positive applications. If emotional responses continue to be driven primarily by mistrust, society risks missing out on cultural and creative innovation while stifling the application of responsible deepfake use.

Deepfakes have the ability to instill more than deception, fear, and manipulation. They may shape people's emotional and behaviorial responses (Wang & Kim, 2022) with public trust and engagement with traditional media networks, social media platforms, and the fabric of online communication. Fear could overshadow the creative use of deepfakes, leading even humorous or artistic uses to be viewed with high skepticism. This highlights a critical challenge: private-sector stakeholders must realize that any prevention and intervention methods should examine deploying technical detection of ill-intended deepfakes and prepare for the emotional and ethical consequences of deepfake proliferation on social media (Pawelec, 2022).

Many participants echoed concerns about the potential for misuse, expressing support for clearer safeguards. As one participant succinctly stated, “All deepfake videos should be watermarked to prevent misuse.” Therefore, regulators, policymakers, and private internet platforms should consider acting swiftly, adding clear labels and transparent disclosures, and enforcing rules to protect people and rebuild trust. It also calls for creators, artists, and entertainers to think about implementing ethical and responsible AI standards (Widder et al., 2022). By emphasizing the importance of ethical considerations, institutions can encourage a more responsible approach to deepfake technology, media literacy, and instilling trust (Hwang et al., 2021). Ultimately, the proliferation of deepfakes may have implications far beyond deception (Westerland, 2019); they shape public emotion, media engagement, and societal

cohesion. "They could disrupt societal trust in information," one participant concluded. Deepfake anxiety is not just about detecting a fake; it is about losing faith in what is real.

5.5 Future Research and Limitations

While the study results provided new insights into the role of deepfakes and their effects on false memory formation amongst Gen Z participants, there were several limitations that could be addressed in future research opportunities. One limitation included the temporal scope and drop-off rates. Participants were only assessed within a short timeframe of one to four days after exposure to deepfakes, which can raise new questions about their durability. Future studies can investigate whether memories persist over more extended periods of time, weeks, or months, to understand their potential long-term effects. Additionally, there was a significant participant abandonment rate, underscoring the need for inclusivity in research and the collection of an even larger sample size to ensure more participant responses.

The survey incorporated deception and focused on passive deepfake exposure with brief informational prompts. In turn, it is unclear if more active interventions, such as critical reflection exercises or guided literacy training, could equip individuals to resist false information. The potential impact of this research on media literacy is significant. New research that can compare the effectiveness of passive awareness of deepfakes versus structured resilience-building techniques would be especially valuable. With these new findings, private and public sector institutions can be better prepared to create media literacy and misinformation detection curricula, tools, and training.

Future work should expand the scope of stimuli used beyond deepfakes featuring content that, while exaggerated, might be plausibly aligned with the public persona or brand of the celebrity. Introducing more absurd or completely implausible deepfakes, such as public figures

behaving wildly inconsistent with their public image, can illustrate how users discern credibility and at what point deepfakes become too unbelievable or persuasive. For example, a video of Kim Kardashian robbing a 7-11 or becoming an astronaut can test believability limits and thresholds for false memory formation and content dismissal.

The current instrument design also highlighted limitations in how participants self-reported using cues and clues when discerning if a video was real or a deepfake. While open-text responses provided rich results, they may not have accurately reflected all the processes involved in the participants' judgment, considering that participants may have used cues and clues unconsciously or struggled to articulate how they used them. New iterations of this study design can provide multiple-choice options and list all cues and clues categories to capture both implicit and explicit detection strategies. Subsequently, the development of such an instrument could also measure individual susceptibility to cognitive biases or preconceived notions about the content, such as confirmation bias or familiarity effects. A new instrument testing for additional biases can provide further assessment to analyze who is at risk of believing or misremembering deepfake content and what factors may contribute to falling for deepfakes. Lastly, this research was limited in demographic scope, focusing primarily on digitally native Gen Z users. New research efforts should examine how different populations, such as older adults, perceive, respond to, and discern deepfake videos and technologies. This is crucial as it could help to ensure that the development of media literacy, misinformation detection tools, and training is inclusive and effective across all age groups. Another option involves further exploring the emotional responses to deepfake exposure over time. While participants self-reported a range of impassioned reactions, the study did not track how those reactions evolved or developed. A longitudinal study can assess whether repeated exposure to deepfakes

can lead to increasing anxiety, desensitization, or shifts in attitudes toward reputable media sources.

5.6 The Lasting Impact of Deepfakes

This study shows that deepfakes may do more than trick people; they could actually create false memories, as demonstrated among the Gen Z adult participants in this research.. Even those who knew about deepfakes or were familiar with the celebrities shown were still influenced, sometimes even more so. While many participants thought they could spot fakes, their confidence did not always match their accuracy, and sometimes they doubted real videos just as much as fake ones. Active debriefing helped reduce the creation of false memories, showing that simple interventions like telling viewers what to look out for could make a big difference. Still, just being aware of deepfakes was not enough. People need more than just knowledge; they need tools to think critically and emotionally cope with this kind of media.

Participants often felt disturbed, anxious, and mistrustful after watching deepfakes, worrying about their own safety and the trustworthiness of digital content in general. This emotional toll, paired with how easy it is to spread convincing fake videos, raises serious concerns about how we process and trust information in the digital age. Going forward, tech companies, educators, and policymakers should consider work together to improve media literacy, enforce clearer labeling, and consider how deepfakes affect not just what we believe but how we feel. As deepfakes become more realistic, the challenge is not just how to spot what is fake, but how to preserve trust in what is real.

Chapter 6

Conclusion

The dissertation began with a deceptively simple question: Can a deepfake video implant false memories in young adults, even when they are told the content is fake? Throughout the chapters, this study has investigated how the effects of deepfakes on false memory formation, trust in digital media, and the discernment abilities of Gen Z. The research was situated amid the rapid proliferation of AI-generated technology and an ever-evolving social media information ecosystem.

Through a 2x2 factorial experimental design, participants were exposed to genuine and falsified TikTok-style videos of Kim Kardashian, Elon Musk, and Paris Hilton and later tested for false memories, accuracy in discernment, and confidence in judgment. The results illustrated that exposure to deepfakes significantly increases the potential for false memory formation, particularly in the absence of debriefing. These findings add to current literature that suggests even short-term exposure to realistic-looking deepfake video content may influence an individual's memories and beliefs.

The study adds to several ongoing academic conversations. First, it empirically confirms that realistic-looking deepfakes have the ability to create measurable cognitive effects, including fabricated memories and beliefs of events that never occurred or are not true. Second, it confirms active debriefing as a dependable intervention strategy for mitigating these effects. Third, it provides insights into how participants evaluate a video's realism by relying on visual cues and

clues over technical or contextual information presented. Participants also frequently cited "something feeling off" as justification for their discernment judgments, suggesting that human deepfake detection relies more on intuition than knowledge.

The conclusions drawn from this dissertation emphasize that the current deepfake phenomenon extends beyond research questions of authenticity, technology, and future use. This emergent media form raises deeper concerns about how society remembers, believes, or observes digital media in an era dominated by audiovisual content generation and manipulation. In particular, my research calls for stronger safeguards around using generative AI tools, including content labeling, transparency, and accessible media literacy training. As technology progresses and improves, so must the research frameworks for understanding and addressing their societal, cultural, economic, and psychological effects.

Overall, the study has validated and extended a research framework for continued inquiry into memory, media trust, and human discernment. Deepfake detection technologies and content moderation policies will not be able to solely resolve the challenges of deepfake videos; rather, they will require an interdisciplinary response grounded in research, ethics, and education. In turn, this dissertation offers a foundation for future work to understand and mitigate the effects of deepfakes in everyday life.

References

Abadie, M., Guette, C., Troubat, A., & Camos, V. (2024). The influence of working memory mechanisms on false memories in immediate and delayed tests. *Cognition, 252*, 105901. https://doi.org/10.1016/j.cognition.2024.105901

Ahmed, N. (2019). Generation Z's smartphone and social media usage: A survey. *Journalism and Mass Communication*, *9*(3). https://doi.org/10.17265/2160-6579/2019.03.001

Ahmed, S. (2023). Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism. *New Media & Society*, *25*(5), 1108–1129. https://doi.org/10.1177/14614448211019198

Ahmed, S. (2023). Examining public perception and cognitive biases in the presumed influence of deepfakes threat: empirical evidence of third person perception from three studies. *Asian Journal of Communication*, *33*(3), 308–331. https://doi.org/10.1080/01292986.2023.2194886

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). The state of deepfakes: Landscape, threats, and impact. *Amsterdam: Deeptrace*, *27*.

Akinyemi, J. P., Chew, S. L., Geeling, S., Heuer, M., WANG, Q., Hassan, N., & Kude, T. (2024). Seeing Isn't Believing: AI Disclosure Labels and Sharing Behavior in the Era of Deepfakes.

Albahar, M.A., & Almalki, J. (2019). Deepfakes: Threats and Countermeasures Systematic Review.

Allen, J., Howland, B., Mobius, M., Rothschild, D., & Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. *Science Advances*, *6*(14), Eaay3539.

Allison, P. (2013, February 13). What's the best R-squared for logistic regression? *Statistical Horizons.* https://statisticalhorizons.com/r2logistic/

Almars, A. M. (2021). Deepfakes detection techniques using deep learning: A survey. *Journal of Computer and Communications*, *9*(5), Article 05. https://doi.org/10.4236/jcc.2021.95003

Alnaser, A. (2024). The Impact of AI-Augmented Image and Video Editing on Social Media Engagement, Perception, and Digital Culture. *Journal of Computational Intelligence, Machine Reasoning, and Decision-Making*, *9*(10), 15-32.

Amazeen, M. A., & Bucy, E. P. (2019). Conferring resistance to digital disinformation: The inoculating influence of procedural news knowledge. *Journal of Broadcasting and Electronic Media, 63*(3), 415-432. https://doi.org/10.1080/08838151.2019.1653101

American Association for Research into Nervous and Mental Diseases, & Johnson, M. K. (1997). Source monitoring and memory distortion. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *352*(1362), 1733-1745.

Amerini, I., Galteri, L., Caldelli, R & Del Bimbo, A. "Deepfake video detection through optical flow based CNN," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (ICCV), 2019.

Anthonysamy, L., & Sivakumar, P. (2022). A new digital literacy framework to mitigate misinformation in social media infodemic. *Global Knowledge Memory and Communication.* https://doi.org/10.1108/gkmc-06-2022-0142

Avram, M., Micallef, N., Patil, S., & Menczer, F. (2020). Exposure to social engagement metrics increases vulnerability to misinformation. *Harvard Kennedy School Misinformation Review, 1*. https://doi.org/10.37016/mr-2020-033.

Babanikos, A., & McAdams, M. (2021). AI Deepfakes in the Entertainment Industry.

Babarović, T., Blažev, M., & Serracant, P. (2021). Piloting longitudinal cohort surveys: Best practices and challenges. *European Survey Research Association (ESRA) 2021*.

Bago, B., Rand, D.G., Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology General*, *149*(8):1608-1613. https://doi.org/10.1037/xge0000729.

Ball, D. (2024). The Deepfake Challenge: Targeted AI Policy Solutions for States. *Mercatus Special Study*.

Barari, S., Lucas, C., & Munger, K. (2021). Political deepfake videos misinform the public, but no more than other fake media. https://doi.org/10.31219/osf.io/cdfh3

Barclay, Donald A. *Fake news, propaganda, and plain old lies: how to find trustworthy information in the digital age*. Rowman & Littlefield, 2018.

Bentkowska-Kafel, A., Cashen, T., & Gardiner, H. (Eds.). (2009). *Digital visual culture: Theory and practice*. Bristol: Intellect.

Bernstein, D., Pernat, N., & Loftus, E. (2011). The false memory diet: False memories alter food preferences. *Handbook of Behavior, Food, and Nutrition*. https://doi.org/10.1007/978-0-387-92271-3_107

Blandón-Gitlin, I., & Gerkens, D. (2010). The effects of photographs and event plausibility in creating false beliefs. *Acta Psychologica, 135*(3), 330–334. https://doi.org/10.1016/j.actpsy.2010.08.008

Bode, L. (2021). Deepfaking Keanu: YouTube deepfakes, platform visual effects, and the complexity of reception. *Convergence: The International Journal of Research into New Media Technologies, 27*, 919 - 934.

Bode, L., Lees, D., & Golding, D. (2021). The digital face and deepfakes on screen. *Convergence: The International Journal of Research into New Media Technologies, 27*, 849 - 854.

Bookbinder, S. H., & Brainerd, C. J. (2016). Emotion and false memory: The context–content paradox. *Psychological Bulletin, 142*(12), 1315–1351. https://doi.org/10.1037/bul0000077

Bourke, B. (2019). Connecting with generation z through social media. In H. Schnackenberg & C. Johnson (Eds.), *Preparing the Higher Education Space for Gen Z* (pp. 124-147). IGI Global. https://doi.org/10.4018/978-1-5225-7763-8.ch007

Brasel, S. A., & Gips, J. (2017). Media multitasking: How visual cues affect switching behavior. *Computers in Human Behavior*, *77*, 258-265.

Bronstein, M., Pennycook, G., Bear, A., Rand, D.G., & Cannon, T.D. (2019). Belief in fake news is associated with delusionality, dogmatism, religious fundamentalism, and reduced analytic thinking. *Journal of Applied Research in Memory and Cognition, 8*, 108-117.

Bushey, J. (2015, January). Trustworthy citizen-generated images and video on social media platforms. In *2015 48th Hawaii International Conference on System Sciences* (pp. 1553-1564). IEEE.

Caldera, E. (2019). “Reject the evidence of your eyes and ears”: Deepfakes and the law of virtual replicants. *Seton Hall Law Review*, *50*(1). https://scholarship.shu.edu/shlr/vol50/iss1/5

Caled, D., & Silva, M. J. (2022). Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation. *Journal of Computational Social Science*, *5*(1), 123–159. https://doi.org/10.1007/s42001-021-00118-8

Carvajal Rodriguez, L. C. (2021). Political deepfakes: Cultural discourses of synthetic audio-visual manipulations. (Order No. 28418067, Temple University). *ProQuest Dissertations and Theses, 71*. Retrieved June 7, 2024, from https://scholarshare.temple.edu/handle/20.500.12613/6583

Cashmore, E. (2023). *Celebrity culture*. Routledge.

Chai, L., Bau, D., Lim, S., & Isola, P. (2020). What makes fake images detectable? Understanding properties that generalize. In *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXVI 16* (pp. 103-120). https://doi.org/10.1007/978-3-030-58574-7_7.

Chatzoglou, P., Chatzoudes, D., Ioakeimidou, D., & Tokoutsi, A. (2020). Generation Z: Factors affecting the use of Social Networking Sites (SNSs). *2020 15th International Workshop on Semantic and Social Media Adaptation and Personalization.* 1–6. https://doi.org/10.1109/SMAP49528.2020.9248473

Chawla, R. (2019). Deepfakes: How a pervert shook the world. *International Journal for Advance Research and Development, 4*(6), 4-8.

Cheever, N. A., & Rokkum, J. (2015). Internet credibility and digital media literacy. *The Wiley Handbook of Psychology, Technology, and Society*, 56-73.

Chen, D.-T., Wu, J., & Wang, Y.-M. (2011). Unpacking new media literacy. *Journal of Systemics, Cybernetics and Informatics, 9*(2), 84-88.

Chi, M. T. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences*, *6*(3), 271-315.

Citron, D.K. & Chesney, R. (2019). Deepfakes and the new disinformation war. *Foreign Affairs*. Available at: https://scholarship.law.bu.edu/shorter_works/76

Collins, A. (Ed.). (2019). *Forged authenticity: Governing deepfake risks*. EPFL International Risk Governance Center (IRGC). https://doi.org/10.5075/epfl-irgc-273296

Crockett, M. (2017). Moral outrage in the digital age. *Nature: Human Behaviour, 1*, 769-771.

Dame Adjin-Tettey, T. (2022). Combating fake news, disinformation, and misinformation: Experimental evidence for media literacy education. *Cogent Arts & Humanities*, *9*(1), 2037229. https://doi.org/10.1080/23311983.2022.2037229

Danry, V., Leong, J., Pataranutaporn, P., Tandon, P., Liu, Y., Shilkrot, R., ... & Sra, M. (2022, April). AI-generated characters: putting deepfakes to good use. In *CHI conference on human factors in computing systems extended abstracts* (pp. 1-5).

De Ruiter, A. (2021). The distinct wrong of deepfakes. *Philosophy & Technology*, *34*(4), 1311-1332.

Deepfakes and Involuntary Pornography: Can Our Current Legal Framework Address This Technology? | Wayne Law Review. (n.d.). Retrieved August 10, 2023, from https://waynelawreview.org/deepfakes-and-involuntary-pornography-can-our-current-legal-framework-address-this-technology/

Deepfake videos easily fool face systems, researchers warn. (2019). *Biometric Technology Today*, *10*, 3-3. https://doi.org/10.1016/S0969-4765(19)30137-7

Deffenbacher, K. A., Bornstein, B. H., Penrod, S. D., & McGorty, E. K. (2004). A meta-analytic

review of the effects of high stress on eyewitness memory. *Law and Human Behavior*, *28*(6), 687-706. https://doi.org/10.1007/s10979-004-0565-x

Delort, J., Arunasalam, B., & Paris, C. (2011). Automatic moderation of online discussion sites. *International Journal of Electronic Commerce*, *15*(3), 9-30. https://doi:10.2753/jec1086-4415150302

Denham, H. (2020, August 3). *Another fake video of Pelosi goes viral on Facebook*. The Washington Post. Retrieved December 2, 2021 from https://www.washingtonpost.com/technology/2020/08/03/nancy-pelosi-fake-video-facebook/.

DePaulo, B. M., & Morris, W. L. (2004). Discerning lies from truths: Behavioral cues to deception and the indirect pathway of intuition. *BM DePaulo, WL Morris, ad. by PA Granhag. The detection of deception in forensic contexts//New York: Cambridge*, 15-40.

Desai, S. P., & Lele, V. (2017). Correlating internet, social networks and workplace – a case of generation z students. *Journal of Commerce and Management Thought*, *8*(4), 802. https://doi.org/10.5958/0976-478X.2017.00050.7

Diakopoulos, N. & Johnson, D. (2019). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.3474183.

Dias, N., Pennycook, G. and Rand, D.G., 2020. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & Vreese, C.H. (2020). Do (microtargeted) deepfakes have real effects on political attitudes. *The International Journal of Press/Politics*, *26*(1), 69-91. https://doi.org/10.1177/1940161220944364.

Doss, C., Mondschein, J., Shu, D., Wolfson, T., Kopecky, D., Fitton-Kane, V. A., Bush, L., & Tucker, C. (2023). Deepfakes and scientific knowledge dissemination. *Scientific Reports*, *13*(1), 13429. https://doi.org/10.1038/s41598-023-39944-3

Drummond, C., Siegrist, M., & Arvai, J. (2020). Limited effects of exposure to fake news about climate change. *Environmental Research Communications*, *2*, 081003. https://doi.org/10.1088/2515-7620/abae77

Du, F. (2025). The Dilemma of Applying Reputation Rights Norms in the Context of Information Authenticity and Its Institutional Optimization: An Empirical Study Based on Deep Synthesis and Text-to-Video Technologies. *Journal of Computing and Electronic Information Management*, *16*(1), 42-48.

Durall, R., Keuper, M., Pfreundt, F., & Keuper, J. (2019). Unmasking deepfakes with simple features. https://doi.org/10.48550/arXiv.1911.00686

Economic Times. (2024, November 23). Elon Musk poses in Iron Man suit; says will use 'the power of irony' to defeat villains. *The Economic Times*. https://economictimes.indiatimes.com/magazines/panache/elon-musk-poses-in-iron-man-suit-says-will-use-the-power-of-irony-to-defeat-villains/articleshow/115629044.cms?from=mdr

Effron, D. A., & Raj, M. (2019). Misinformation and morality: Encountering fake-news headlines makes them seem less unethical to publish and share. *Psychological Science, 31*(1), 75–87. https://doi.org/10.1177/0956797619887896

Etienne, H. (2021). The future of online trust (and why Deepfake is advancing it). *AI and Ethics*, *1*(4), 553-562.

Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2021). TweepFake: About

detecting deepfake tweets. *PloS one*, *16*(5), e0251415. https://doi.org/10.1371/journal.pone.0251415

Faragó, L., Kende, A., & Krekó, P. (2020). We only believe in news that we doctored ourselves: The connection between partisanship and political fake news. *Social Psychology, 51*(2), 77–90. https://doi.org/10.1027/1864-9335/a000391

Farkas, J., Schou, J. & Neumayer, C. (2018). Cloaked Facebook pages: Exploring fake Islamist propaganda in social media. *New Media & Society, 20*, 1850–1867. https://doi.org/10.1177/146144481770775

Fletcher, R., & Park, S. (2017). The impact of trust in the news media on online news consumption and participation. *Digital Journalism, 5*(10), 1281–1299. https://doi.org/10.1080/21670811.2017.1279979

Fogt, S. (2021). A rhetorical approach to assessing source credibility: Digital natives, lateral reading, and the need for media literacy curriculum. *Seton Hall University Dissertations and Theses (ETDs)*. https://scholarship.shu.edu/dissertations/2877

Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false‐positive findings–a practical guide. *Biological Reviews*, *92*(4), 1941–1968.

Frau-Meigs, D., & Corbu, N. (Eds.). (2024). *Disinformation debunked: Building resilience through media and information literacy*. Taylor & Francis.

García-Marín, D. (2021). Las fake news y los periodistas de la generación z. Soluciones post-millennial contra la desinformación. *Vivat Academia*, *154*, 37–63. https://doi.org/10.15178/va.2021.154.e1324

Gardiner, N. (2019). Facial re-enactment, speech synthesis and the rise of the deepfake. *Edith Cowan University.*

Garriga, M., Ruiz Incertis, R., & Magallón Rosa, R. (2024). Artificial intelligence, disinformation and media literacy proposals around deepfakes. *Observatorio (OBS*)*, *18*, 175–194. https://doi.org/10.15847/obsOBS18520242445

Garry, M., & Gerrie, M. P. (2005). When photographs create false memories. *Current Directions in Psychological Science*, *14*(6), 321–325. https://doi.org/10.1111/j.0963-7214.2005.00390.x

Garry, M. & Wade, K.A. (2005). Actually, a picture is worth less than 45 words: Narratives produce more false memories than photographs do. *Psychonomic Bulletin & Review, 12*, 359-366.

George, A. S., & George, A. S. (2023). Deepfakes: The evolution of hyper realistic media manipulation. *Zenodo*. https://doi.org/10.5281/zenodo.10148558

Gieseke, A. P. (2020). "The New weapon of choice": Law's current inability to properly address deepfake pornography, 73 *Vanderbilt Law Review* 1479 (2020) Available at: https://scholarship.law.vanderbilt.edu/vlr/vol73/iss5/4

Gil, R., Virgili-Gomà, J., López-Gil, J.-M., & García, R. (2023). Deepfakes: Evolution and trends. *Soft Computing, 27*(16), 11295–11318. https://doi.org/10.1007/s00500-023-08605-y

Godulla, A., Hoffmann, C., & Seibert, D. (2021). Dealing with deepfakes - an interdisciplinary examination of the state of research and implications for communication studies. *Studies in Communication and Media, 10*. https://doi.org/10.5771/2192-4007-2021-1-72.

Gomez, K., Mawhinney, T., & Betts, K. (2019). Welcome to generation Z. *Deloitte Network of Executive Women*, *24*.

Good Morning America. (2018, April 18). *Jordan Peele uses AI, President Obama in fake*

news PSA [Video]. YouTube. https://www.youtube.com/watch?v=bE1KWpoX9Hk

Gosse, C., & Burkell, J.A. (2020). Politics and porn: how news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication, 37*(5), 497 - 511.

Government of Canada, P. S. and P. C. (2002). *Deep fakes: What can be done about synthetic audio and video? / B.J. Siekierski, Economics, Resources and International Affairs Division, Parliamentary Information and Research Service.: YM32-5/2019-11E-PDF - Government of Canada Publications - Canada.ca*. https://publications.gc.ca/site/eng/9.878187/publication.html

Gow, K. (1998). The complex issues in researching "false memory syndrome." *Australasian Journal of Disaster and Trauma Studies, 2*(3). Retrieved March 5, 2024, from https://trauma.massey.ac.nz/issues/1998-3/gow1.htm

Graber, D. (2006). Seeing is remembering: How visuals contribute to learning from television news. *Journal of Communication*, *40*(3), 134–156. https://doi.org/10.1111/j.1460-2466.1990.tb02275.x

Greene, C. & Murphy, G. (2023). Debriefing works: Successful retraction of misinformation following a fake news study. PLOS ONE. 18. e0280295. 10.1371/journal.pone.0280295.

Greengard, S. (2019). Will deepfakes do deep damage? *Communications of the ACM*, *63*. 17-19. 10.1145/3371409.

Grigoreva, E. A., Garifova, L. F., & Polovkina, E. A. (2021). Consumer behavior in the information economy: Generation Z. *International Journal of Financial Research*, *12*(2), Article 2. https://doi.org/10.5430/ijfr.v12n2p164

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B. & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science, 363*(6245). 374-378. https://doi.org/10.1126/science.aau2706.

Groh, M. (n.d.). *Project overview ' detect deepfakes: How to counteract misinformation created by AI*. MIT Media Lab. Retrieved May 2, 2022, from https://www.media.mit.edu/projects/detect-fakes/overview/

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2021). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, *119*(1). https://doi.org/10.1073/pnas.2110013119

Guarnera, L., Giudice, O., & Battiato, S. (2020). Deepfake detection by analyzing convolutional traces. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2841-2850.

Güera, David & Delp, Edward. (2018). Deepfake video detection using recurrent neural networks. 1-6. https://doi.org/10.1109/AVSS.2018.8639163.

Guess, A. M., Lerner, M., Lyons, B., Montgomery, J. M., Nyhan, B., Reifler, J., & Sircar, N. (2020). A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, *117*(27), 15536–15545. https://doi.org/10.1073/pnas.1920498117

Guess, A. M., Nyhan, B., & Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 US election. *Nature: Human Behaviour, 4*(5), 472–480. https://doi.org/10.1038/s41562-020-0833-x

Guo, L. & Johnson, B. G. (2020). Third-person effect and hate speech censorship on Facebook. *Social Media + Society*. https://doi.org/10.1177/2056305120923003

Hameleers, M., van der Meer, T. G., & Dobber, T. (2024). They would never say anything like this! Reasons to doubt political deepfakes. *European Journal of Communication*, *39*(1), 56-70.

Hancock, J.T., & Bailenson, J.N. (2021). The social impact of deepfakes. *Cyberpsychology, behavior and social networking, 24*(3), 149-152.

Hargittai, E., & Hinnant, A. (2008). Digital inequality: Differences in young adults' use of the Internet. *Communication research*, *35*(5), 602-621.

He, A., Wesley, C., Briggs, E., Burns, M. J., Marlatt, J., & Tran, K.(2024). *What to know about Gen Z's engagement with social media, entertainment, and technology. Morning Consult.* https://pro.morningconsult.com/analyst-reports/gen-z-engagement-social-media-entertainment-tech

Heaps, C. M., & Nash, M. (2001). Comparing recollective experience in true and false autobiographical memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(4), 920–930. https://doi.org/10.1037/0278-7393.27.4.920

Hern, A. I. (2021). Don't want to upset people': Tom Cruise Deepfake Creator Speaks Out. *The Guardian*. https://www.theguardian.com/technology/2021/mar/05/how-started-tom-cruise-deepfake-tiktok-videos

Hertwig, R., & Ortmann, A. (2008). Deception in social psychological experiments: Two misconceptions and a research agenda. *Social Psychology Quarterly, 71*(3), 222–227. https://doi.org/10.1177/019027250807100304

Hilton, P. [@parishilton]. (2024, January 1). What's better than @hilton rewards for playing in #Slivingland, my Roblox world?! 🎮👾 If you complete the new Hilton scavenger hunt [Video]. TikTok. https://www.tiktok.com/@parishilton/video/7326968371042438443

Hwang, Y., Ryu, J. Y., & Jeong, S.H. (2021). Effects of disinformation using deepfake: The protective effect of media literacy education. *Cyberpsychology, Behavior and Social Networking*, *24*(3), 188–193. https://doi.org/10.1089/cyber.2020.0174

Ismail, S., & Latif, R. A. (2013). Authenticity issues of social media: Credibility, quality and reality.

Johansen, A. G. (2020, August 13). *How to spot deepfake videos - 15 signs to watch for*. How to spot deepfake videos — 15 signs to watch for. Retrieved May 2, 2022, from https://us.norton.com/internetsecurity-emerging-threats-how-to-spot-deepfakes.html

Johnson, D., & Diakopoulos, N. (2021). What to do about deepfakes. *Communications of the ACM*, *64*(3), 33–35. https://doi.org/10.1145/3447255

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin, 114*(1), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Jones-Jang, S. M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information literacy helps, but other literacies don't. *American Behavioral Scientist, 65*(2), 371–388. https://doi.org/10.1177/0002764219869406

Josephs, E., Fosco, C., & Oliva, A. (2024). Effects of browsing conditions and visual alert design on human susceptibility to deepfakes. *Journal of Online Trust and Safety*, *2*(2), Article 2. https://doi.org/10.54501/jots.v2i2.144

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology, 7*(4), 396–403. https://doi.org/10.9734/BJAST/2015/14975

Kahan, D. (2013). Ideology, motivated reasoning, and cognitive reflection. *Judgment and Decision Making, 8*(4). 407-424.

Karasavva, V., & Noorbhai, A. (2021). The real threat of deepfake pornography: a review of Canadian policy. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 203-209.

Kardaras, N. (2016). Generation Z: Online and at risk? *Scientific American Mind, 27*(5), 64-69. https://doi.org/10.1038/SCIENTIFICAMERICANMIND0916-64

Kardashian, K. [@kimkardashian]. (2023, April 27). Just dropped new @SKIMS shapewear [Video]. TikTok. https://www.tiktok.com/@kimkardashian/video/7226922259217272106?lang=en

Karmarkar, U. R., & Tormala, Z. L. (2010). Believe me, I have no idea what I'm talking about: The effects of source certainty on consumer involvement and persuasion. *Journal of Consumer Research, 36*(6), 1033–1049. https://doi.org/10.1086/648381

Karnouskos, S. (2020). Artificial intelligence in digital media: The Era of deepfakes. *IEEE Transactions on Technology and Society, 1*(3), 138-147.

Kasra, M., Shen, C., & O'Brien, J. (2016). Seeing is believing: How people fail to identify fake images on the web. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3051561

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons, 63*(2), 135–146. https://doi.org/10.1016/j.bushor.2019.11.006

Kikerpill, K. (2020). Choose your stars and studs: The rise of deepfake designer porn. *Porn*

Studies, *7*(4), 352-356.

Kircaburun, K., Demetrovics, Z., Király, O., & Griffiths, M. D. (2020). Childhood emotional trauma and cyberbullying perpetration among emerging adults: A multiple mediation model of the role of problematic social media use and psychopathology. *International Journal of Mental Health and Addiction*, *18*(6), 548-566.

Klomp, J. (2020, November 17). *Creating false memories by (misleading) visual narratives.* https://jeanineklomp.wordpress.com/2020/11/17/blog-2/

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience*, *24*(11), 103364. https://doi.org/10.1016/j.isci.2021.103364

Köbis, N.C. and Mossink, L. (2021). Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry. *Computers in Human Behavior, 114*(2), 106553.

Koç, B. (2023). *The role of user interactions in social media on recommendation algorithms: Evaluation of TikTok's personalization practices from user's perspective*. https://doi.org/10.13140/RG.2.2.34692.71040

Korona, M., & Hutchison, A. (2023). Integrating media literacy across the content areas. *Reading Research Quarterly*, *58*(4), 601–623. https://doi.org/10.1002/rrq.517

Korshunov, P., & Marcel, S. (2018). Deepfakes: A new threat to face recognition? Assessment and detection. https://doi.org/10.48550/arXiv.1812.08685.

Korshunov, P., & Marcel, S. (2020). Deepfake detection: humans vs. machines. https://doi.org/10.48550/arXiv.2009.03155.

Kozyreva, A., Lewandowsky, S., & Hertwig, R. (2020). Citizens versus the internet: Confronting digital challenges with cognitive tools. *Psychological Science in the Public Interest, 21*(3), 103–156. https://doi.org/10.1177/1529100620946707

Kruijt, J., Meppelink, C. S., & Vandeberg, L. (2022). Stop and think! Exploring the role of news truth discernment, information literacy, and impulsivity in the effect of critical thinking recommendations on trust in fake COVID-19 news. *European Journal of Health Communication, 3*(2), 40–63. https://doi.org/10.47368/ejhc.2022.203

Langa, J. (2021). Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *Boston University Law Review*, *101*, 761.

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P., & Schroeder, D. T. (2021). Don't trust your eyes: Image manipulation in the age of deepfakes. *Frontiers in Communication*, *6*. https://doi.org/10.3389/fcomm.2021.632317

Le, V. (2020). The deepfakes to come: A turing cop's nightmare. *Identities: Journal for Politics, Gender and Culture*, *17*(2-3), 8-18. https://doi.org/10.51151/identities.v17i2-3.468

LeBlanc III, H. P. (2019). Communication ethics and public deception: The effects of misinformation on desensitization to truth-telling. *Quarterly Review of Business Disciplines*, *6*(2), 113-132.

Levine, R. J. (1982). Consent to incomplete disclosure as an alternative to deception. *IRB: Ethics & Human Research, 4*(10), 9–11. https://doi.org/10.2307/3564068

Lewandowsky, S., Ecker, U. K. H., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition, 6*(4), 353–369. https://doi.org/10.1016/j.jarmac.2017.07.008

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European review of social psychology*, *32*(2), 348-384.

Li, J. (2025). Not all skepticism is "healthy" skepticism: Theorizing accuracy- and identity-motivated skepticism toward social media misinformation. *New Media & Society, 27*(1), 522–544.

Li, J., Qu, Y., Chao, F., Shum, H., Ho, E., & Yang, L. (2019). Machine learning algorithms for network intrusion detection. *Intelligent Systems Reference Library* (pp. 151–179). https://doi.org/10.1007/978-3-319-98842-9_6

Li, K. Y., Zahiri, M. A., & Jumaat, N. F. (2021). Eye on digital media literacy from the perspective of 'generation Z'. *European Proceedings of Social and Behavioural Sciences*.

Li, Y., & Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts.

Liamputtong, P. (2007). *Researching the vulnerable: A guide to sensitive research methods*. SAGE Publications.

Liao, K., Lin, C., & Zhao, Y. (2021). A deep ordinal distortion estimation approach for distortion rectification. *IEEE Transactions on Image Processing*, *30*, 3362–3375. https://doi.org/10.1109/TIP.2021.3061283

Lindsay, D.S., Hagen, L.H., Read, J., Wade, K.A., & Garry, M. (2004). True photographs and false memories. *Psychological Science, 15*(3), 149–154. https://doi.org/10.1111/j.0956-7976.2004.01503002

Liv, N., & Greenbaum, D. (2020). Deep fakes and memory malleability: False memories in the service of fake news. *AJOB Neuroscience*, *11*(2), 96–104. https://doi.org/10.1080/21507740.2020.1740351

Livingstone, S. (2004). Media literacy and the challenge of new information and communication technologies. *The Communication Review, 7*(1), 3–14. https://doi.org/10.1080/10714420490280152

Loftus, E.F. (1999). Lost in the mall: Misrepresentations and misunderstandings. *Ethics & Behavior*, *(9)*1, 51-60. https://doi.org/10.1207/s15327019eb0901_4

Loftus, E. F., & Bernstein, D. M. (2005). Rich false memories: The royal road to success. In A. F. Healy (Ed.), *Experimental cognitive psychology and its applications* (pp. 101–113). American Psychological Association. https://doi.org/10.1037/10895-008

Loftus, E.F., & Hoffman, H.G. (1989). Misinformation and memory: The creation of new memories. *Journal of experimental psychology. General, 118*, 100-4.

Loftus, E. F. and Pickrell, J. E. (1995). The formation of false memories. *Psychiatric Annals, 25*(12), 720–725. https://doi.org/10.3928/0048-5713-19951201-07

Lovato, J., St-Onge, J., Harp, R., Salazar Lopez, G., Rogers, S. P., Haq, I. U., Hébert-Dufresne, L., & Onaolapo, J. (2024). Diverse misinformation: Impacts of human biases on detection of deepfakes on networks. *Npj Complexity*, *1*(1), 1–13. https://doi.org/10.1038/s44260-024-00006-y

Luitse, D. M. R. (2019). *Does digital literacy have a role in mitigating the spread of misinformation? A critical analysis of deepfake videos and their comment sections*. https://studenttheses.uu.nl/handle/20.500.12932/823

Luoma-aho, H. R., Jaana T. K., &Vilma, L. (2020). Generation Z and organizational listening on social media. *Media and Communication, 8*(2). https://doi.org/10.17645/mac.v8i2.2772

Maddocks, S. (2020). 'A Deepfake porn plot intended to silence me': exploring continuities between pornographic and 'political' deep fakes. *Porn Studies, 7*(4). 1-9.

https://doi.org/10.1080/23268743.2020.1757499

Maksutov, A., Morozov, V., Lavrenov, A., & Smirnov, A. (2020). Methods of deepfake detection based on machine Learning. *2020 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering,* 408-411. https://doi.org/10.1109/EIConRus49466.2020.9039057

Mallinson, C., Janeja, V. P., Evered, C., Khanjani, Z., Davis, L., Bhalli, N. N., & Nwosu, K. (2024). A place for (socio)linguistics in audio deepfake detection and discernment: Opportunities for convergence and interdisciplinary collaboration. *Language and Linguistics Compass, 18*(5). https://doi.org/10.1111/lnc3.12527

Mara, C. A., Cribbie, R. A., Flora, D. B., LaBrish, C., Mills, L., & Fiksenbaum, L. (2012). An improved model for evaluating change in randomized pretest, posttest, follow-up designs. *Methodology, 8*(3), 97–103. https://doi.org/10.1027/1614-2241/a000041

Maras, M. H., & Alexandrou, A. (2019). Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. *The International Journal of Evidence & Proof*, *23*(3), 255-262.

Marra, F., Gragnaniello, D., Cozzolino, D., and Verdoliva, L. (2018). Detection of GAN-generated fake images over social networks. *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), 95*(c), 384-389. https://doi.org/10.1109/MIPR.2018.00084.

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications, 5*(1), 47. https://doi.org/10.1186/s41235-020-00252-3

Mather, M. (2009). When Emotion Intensifies Memory Interference. In B. H. Ross (Ed.), *The*

psychology of learning and motivation, *51*, 101–120. Elsevier Academic Press. https://doi.org/10.1016/S0079-7421(09)51003-1

McAlister, A. M., Lee, D. M., Ehlert, K. M., Kajfez, R. L., Faber, C. J., & Kennedy, M. S. (2017). Qualitative coding: An approach to assess inter-rater reliability. *ASEE Annual Conference & Exposition*.

McCosker, A. (2024). Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. *New Media & Society*, *26*(5), 2786–2803. https://doi.org/10.1177/14614448221093943

Mihailidis, P. & Viotty, S. (2017). Spreadable spectacle in digital culture: Civic expression, fake news, and the role of media literacies in "post-fact" society. *American Behavioral Scientist, 61*(4), 441–454. https://doi.org/10.1177/0002764217701217

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR), 54*(1), 1–41. https://doi.org/10.1145/3425780

Mokadem, S. S. E. (2023). The effect of media literacy on misinformation and deep fake video detection. *Arab Media & Society, 35,* 115-138.

Mokyr, J., Vickers, C., & Ziebarth, N. L. (2015). The history of technological anxiety and the future of economic growth: Is this time different? *Journal of Economic Perspectives*, *29*(3), 31-50.

Moravec, P., Minas, R., & Dennis, A. (2018). Fake news on social media: People believe what they want to believe when it makes no dense at all. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3269541

MRQUICK. (2022, October 10). Real men drink white claw – whiteclawMAN [Video]. YouTube. https://www.youtube.com/watch?v=fbYSMZdb03s

Müller, N. M., Pizzi, K., & Williams, J. (2022). Human Perception of Audio Deepfakes. *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia*, 85–91. https://doi.org/10.1145/3552466.3556531

Murphy, G., Ching, D., Twomey, J., & Linehan, C. (2023). Face/Off: Changing the face of movies with deepfakes. *PloS one*, *18*(7), e0287503.https://doi.org/10.1371/journal.pone.0287503

Murphy, G., & Flynn, E. (2021). Deepfake false memories. *Memory (Hove, England)*, 1–13. Advance online publication. https://doi.org/10.1080/09658211.2021.1919715

Muschalla, B., & Schönborn, F. (2021). Induction of false beliefs and false memories in laboratory studies-A systematic review. *Clinical Psychology & Psychotherapy*, *28*(5), 1194–1209. https://doi.org/10.1002/cpp.2567

Nardi, P. M. (2018). *Doing survey research: A guide to quantitative methods* (4th ed.). Routledge.

Nash, R. A., Wade, K. A., & Brewer, R. J. (2009). Why do doctored images distort memory? *Consciousness and Cognition*, *18*(3), 773–780. https://doi.org/10.1016/j.concog.2009.04.011

Nast, C. (2022, July 7). Kim Kardashian answers beauty questions from the internet. *Allure*. https://www.allure.com/video/watch/kim-kardashian-answers-beauty-questions-from-the-Internet

Nenovski, B., Ilijevski, I., & Stanojoska, A. (2023). Strengthening Resilience Against Deepfakes as Disinformation Threats.

Neuendorf, K. A. (2018). Content analysis and thematic analysis. In P. Brough (Ed.), *Advanced research methods for applied psychology: Design, analysis and reporting* (1st ed., pp.

211–223). Routledge. https://doi.org/10.4324/9781315517971-21

Newman, E. J., & Lindsay, D. S. (2009). False memories: What the hell are they for? *Applied Cognitive Psychology, 23*(8), 1105–1121. https://doi.org/10.1002/acp.1613

Nightingale, S. J., Wade, K. A., & Watson, D. G. (2017). Can people identify original and manipulated photos of real-world scenes? *Cognitive Research: Principles and Implications, 2*(1), 30. https://doi.org/10.1186/s41235-017-0067-2

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep learning for deepfakes creation and detection: A survey. *Computer Vision and Image Understanding*, *223*, 103525. https://doi.org/10.1016/j.cviu.2022.103525

O'Connor, C., & Weatherall, J. O. (2019). *The Misinformation Age: How False Beliefs Spread*. Yale University Press. https://doi.org/10.2307/j.ctv8jp0hk

Omeje, S. O., Oparaugo, B., & Chukwuka, M. O. (2024). Artificial intelligence on social media: Use, misuse, and impacts. In *Emergence of social media: Shaping the digital discourse of the next generation* (p. 38).

Ozkan, M., & Solmaz, B. (2015). Mobile addiction of generation z and its effects on their social lifes: (An application among university students in the 18-23 age group). *Procedia - Social and Behavioral Sciences, 205*(9), 92–98. https://doi.org/10.1016/j.sbspro.2015.09.027

Panagıotou, N., Lazou, C., & Baliou, A. (2022). Generation Z: Media consumption and mil. *İmgelem, 6*(11), 455-476. https://doi.org/10.53791/imgelem.1187245

Paris, B., & Donovan, J. "Deepfakes and cheap fakes." Data & Society, September 18, 2019. Challenges of AI-Altered Video, *Open Information Science*, *3*(1), 32-46.

https://doi.org/10.1515/opis-2019-0003'

Pataranutaporn, P., Archiwaranguprok, C., Chan, S., Loftus, E. F., & Maes, P. (2024). Synthetic human memories: AI-edited images and videos can implant false memories and distort recollection. *ArXiv preprint*. https://doi.org/10.48550/arXiv.2409.08895

Pawelec, M. (2022). Deepfakes and democracy (theory): How synthetic audio-visual media for disinformation and hate speech threaten core democratic functions. *Digital Society*, *1*(2), 19.

Pehlivanoglu, D., Lin, T., Deceus, F., Heemskerk, A., Ebner, N. C., & Cahill, B. S. (2021). The role of analytical reasoning and source credibility on the evaluation of real and fake full-length news articles. *Cognitive Research: Principles and Implications, 6*(1), 24. https://doi.org/10.1186/s41235-021-00292-3

Peng, C. Y. J., Lee, K. L., & Ingersoll, G. M. (2002). An introduction to logistic regression analysis and reporting. *The Journal of Educational Research, 96*(1), 3–14. https://doi.org/10.1080/00220670209598786

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological science*, *31*(7), 770–780. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019b). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality*, *88*(2), 185–200. https://doi.org/10.1111/jopy.12476

Pennycook, G., & Rand, D. G. (2021). The Psychology of fake news. *Trends in cognitive sciences*, *25*(5), 388–402. https://doi.org/10.1016/j.tics.2021.02.007

Pérez Dasilva, J.Á., Meso Ayerdi, K., & Mendiguren Galdospin, T. (2021). Deepfakes on Twitter: Which actors control their spread? *Media and Communication, 9*(1), 301-312.

Pérez-Escoda, A., Pedrero Esteban, L., Romero, J., & Jimenez-Narros, C. (2021). Fake news reaching young people on social networks: Distrust challenging media literacy. *Publications*, *9*(2). https://doi.org/10.3390/publications9020024

Peter, K., Izsák, T., & Vasa, L. (2020). Attitudes of Z generations to job searching through social media. *Economics and Sociology*, *13*(4), 227–240. https://doi.org/10.14254/2071-789X.2020/13-4/14

Peters, M. A. (2017). The information wars, fake news and the end of globalisation. *Educational Philosophy and Theory*, *50*(13), 1161–1164. https://doi.org/10.1080/00131857.2017.1417200

Picard, R. W. (2000). *Affective computing*. MIT press.

Polage, D. C. (2012). Making up history: False memories of fake news stories. *Europe's Journal of Psychology*, *8*(2), 245-250. https://doi.org/10.5964/ejop.v8i2.456

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Potter, W. (2014, December 15). *Guidelines for Media Literacy Interventions in the Digital Age*. https://www.semanticscholar.org/paper/Guidelines-for-Media-Literacy-Interventions-in-the-Potter/c02f04cc3b0eb996f83259b4a3e0287da5a231d9

Prelog, L., & Bakic-Tomic, L. (2020). The perception of the fake news phenomenon on the internet by members of Generation Z. *2020 43rd International Convention on Information, Communication and Electronic Technology (MIPRO)*, 452–455.

https://doi.org/10.23919/MIPRO48935.2020.9245169

Qiu, C. (2024). The Impact of Media Literacy and Internet Information Trustworthiness on Fact-Checking Behavior among Elderly Internet Users: The Moderating Role of Critical Thinking. *International Journal of Social Sciences and Public Administration*, *3*(3), 352–358. https://doi.org/10.62051/ijsspa.v3n3.44

Răduț, F. (2021). Generation "Z" and social networks. *Journal of Educational Studies*, *3*(1), Article 1.

Ramos Vallecillo, N., & Murillo-Ligorred, V. (2025). *The phenomenon of artificial intelligence-generated images in university teacher training and its impact on developing critical thinking* (No. ART-2025-141309).

Rao, A. (2022). Deceptive claims using fake news advertising: The impact on consumers. *Journal of Marketing Research, 59*(3), 534–554. https://doi.org/10.1177/00222437211072896

Real Time with Bill Maher. (2023, April 29). Elon Musk (Full Interview) | Real Time with Bill Maher (HBO) [Video]. YouTube. https://www.youtube.com/watch?v=oO8w6XcXJUs

Reeves, B., Yeykelis, L., & Cummings, J. J. (2015). The Use of Media in Media Psychology. *Media Psychology*, *19*(1), 49–71. https://doi.org/10.1080/15213269.2015.1030083

Reyna, V. F., & Lloyd, F. (1997). Theories of false memory in children and adults. *Learning and individual differences*, *9*(2), 95-123.

Riesthuis, P., Mangiulli, I., Broers, N., & Otgaar, H. (2022). Expert opinions on the smallest effect size of interest in false memory research. *Applied Cognitive Psychology, 36*(1), 203–215. https://doi.org/10.1002/acp.3902

Romero Moreno, F. (2024). Generative AI and deepfakes: a human rights approach to tackling

harmful content. *International Review of Law, Computers & Technology*, *38*(3), 297–326. https://doi.org/10.1080/13600869.2024.2324540

Ross, A. R., Chadwick, A., & Vaccari, C. (2021). Digital media and the proliferation of public opinion cues online: Biases and vulnerabilities in the new attention economy. In *The Routledge Companion to Political Journalism* (pp. 241-251). Routledge.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1-11.

Sánchez-Castillo, S., & López-Olano, C. (2021). Vetting and verifying multimodal false information. A challenge for democratic societies. *Anàlisi: Quaderns de Comunicació i Cultura*, 1-7.

Salovich, N. A., Donovan, A. M., Hinze, S. R., & Rapp, D. N. (2021). Can confidence help account for and redress the effects of reading inaccurate information? *Memory & Cognition*, *49*(2), 293–310. https://doi.org/10.3758/s13421-020-01096-4

Sawilowsky, S., Kelley, D. L., Blair, R. C., & Markman, B. S. (1994). Meta-analysis and the Solomon four-group design. *The Journal of Experimental Education, 62*(4), 361–376. https://doi.org/10.1080/00220973.1994.9944140

Schetinger, V., Oliveira, M., Silva, R., & Carvalho, T. (2015). Humans are easily fooled by digital images. *Computers & Graphics, 68*. 142-151. https://doi.org/10.1016/j.cag.2017.08.010

Schwarz, G. (2005). Overview: What is media literacy, who cares, and why? *Teachers College Record*, *107*(13), 5–17. https://doi.org/10.1177/016146810810701301

Security firm: Deepfakes are ‘fraud’s next frontier.’ (2021). *Biometric Technology Today*,

2021(6), 2–3. https://doi.org/10.1016/S0969-4765(21)00064-3

Seng, L. K., Mamat, N., Abas, H., & Ali, W. N. H. W. (2024). AI Integrity Solutions for Deepfake Identification and Prevention. *Open International Journal of Informatics*, *12*(1), 35-46.

Shahbazi, M., & Bunker, D. (2024). Social media trust: Fighting misinformation in the time of crisis. *International Journal of Information Management*, *77*, 102780.

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. *2023 International Conference on Computer and Applications (ICCA)*, 1–7. https://doi.org/10.1109/ICCA59364.2023.10401723

Silveira, P. (2020). Fake news consumption through social media platforms and the need for media literacy skills: a real challenge for Z Generation. *INTED2020 Proceedings*, (pp. 3830–3838). https://doi.org/10.21125/inted.2020.1063

Singh, V. K., Ghosh, I., & Sonagara, D. (2021). Detecting fake news stories via multimodal analysis. *Journal of the Association for Information Science and Technology*, *72*(1), 3–17. https://doi.org/10.1002/asi.24359

Singhal, R., & Rana, R. (2015). Chi-square test and its application in hypothesis testing. *Journal of the Practice of Cardiovascular Sciences, 1*(1), 69–71. https://doi.org/10.4103/2395-5414.157577

Skibba, R. (2020). Media enhanced by artificial intelligence: Can we believe anything anymore? *Engineering*, *6*. https://doi.org/10.1016/j.eng.2020.05.011

Smith, A., & Hutson, J. (2024). Satirical Deepfakes, Surreal Dreamscapes & Nostalgic Pixels: The Rapid Evolution and Cultural Commentary of AI-Aesthetics. *ISAR Journal of Arts, Humanities and Social Sciences*, *2*(9).

Soysal, F., & Calli, B. A. (2019). Intra and intergenerational digital divide through ICT literacy, information acquisition skills, and internet utilization purposes: An analysis of Gen Z.

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., Zaffar, M.A., & Zaffar, M.F. (2021). Seeing is believing: Exploring perceptual differences in deepfake videos. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445699

Ternovski, J., Kalla, J., & Aronow, P. M. (2021). Deepfake warnings for political videos increase disbelief but do not improve discernment: Evidence from two experiments. *OSF Preprints*, *10*. https://doi.org/10.31219/osf.io/dta97

Thomas, A. K., & Loftus, E. F. (2002). Creating bizarre false memories through imagination. *Memory & Cognition, 30*(3), 423–431. https://doi.org/10.3758/BF03194942

Tortora, J. (2013). Reconsidering the standards of admission for prior bad acts evidence in light of research on false memories and witness preparation. *Fordham Urban Law Journal*, *40*(4). Retrieved June 12, 2024, from https://law-journals-books.vlex.com/vid/reconsidering-bad-evidence-witness-464173986

Troshani, I., Rao Hill, S., Sherman, C., & Arthur, D. (2021). Do we trust in AI? Role of anthropomorphism and intelligence. *Journal of Computer Information Systems*, *61*(5), 481-491.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Uddin, Mohd. (2019). Cognitive Science and Artificial Intelligence: Simulating the Human Mind and Its Complexity. Cognitive Computation and Systems. 1. 10.1049/ccs.2019.0022.

Urueña, S. (2019). Understanding "plausibility": A relational approach to the anticipatory heuristics of future scenarios. *Futures*.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*. https://doi.org/10.1177/2056305120903408

Van Bavel, J. J., & Pereira, A. (2018). The Partisan Brain: An Identity-Based Model of Political Belief. *Trends in Cognitive Sciences*, *22*(3), 213–224. https://doi.org/10.1016/j.tics.2018.01.004

Van der Sloot, B., & Wagensveld, Y. (2022). Deepfakes: regulatory challenges for the synthetic society. *Computer Law & Security Review*, *46*, 105716.

Van Duyn, E., & Collier, J. (2018). Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication and Society, 22*(1), 29–48. https://doi.org/10.1080/15205436.2018.1511807

Veeriah, J. (2021). Young adults' ability to detect fake news and their new media literacy level in the wake of the COVID-19 pandemic. *Journal of Content, Community, & Communication, 13*(7), 372-383.

Vegetti, F. & Mancosu, M. (2020). The impact of political sophistication and motivated reasoning on misinformation, *Political Communication, 37*(5), 678-695. https://doi.org/10.1080/10584609.2020.1744778

Verma, N., Fleischmann, K.R., & Koltai, K.S. (2017). Human values and trust in scientific

journals, the mainstream media and fake news. *Proceedings of the Association for Information Science and Technology, 54*, 426 - 435.

Vidani, J. (2024). *Exploring the influence of reels and short videos on the reading and listening habits of Generation Z: A comprehensive study*. SSRN. https://ssrn.com/abstract=4889495

Villena, D. Deepfakes, deception, and distrust Epistemic and social concerns.

Vițelar, A. (2019). Like me: Generation Z and the use of social media for personal branding. *Management Dynamics in the Knowledge Economy*, *7*(2), 257–268.

Volti, R., & Croissant, J. (2024). *Society and Technological Change* (Ninth). Worth Publishers.

Vraga, E. K., Bode, L., & Tully, M. (2021). The effects of a news literacy video and real-time corrections to video misinformation related to sunscreen and skin cancer. *Health Communication, 37*(13), 1622–1630. https://doi.org/10.1080/10410236.2021.1910165

Wade, K.A., Garry, M., & Pezdek, K. (2018). Deconstructing Rich False Memories of Committing Crime: Commentary on Shaw and Porter (2015). *Psychological Science, 29*(3), 471-476. https://doi.org/10.1177/095679761770366

Wade, K. A., Garry, M., Read, J. D., & Lindsay, D. S. (2002). A picture is worth a thousand lies: using false photographs to create false childhood memories. *Psychonomic Bulletin & Review*, *29*(3), 597–603. https://doi.org/10.3758/bf03196318

Wang, S., & Kim, S. (2022). Users' emotional and behavioral responses to deepfake videos of K-pop idols. *Computers in Human Behavior*, *134*, 107305.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, *9*(11). https://doi.org/10.22215/timreview%2F1282

Wickens, T. (2002). Elementary Signal Detection Theory. *Oxford University Press.* https://doi.org/10.1093/acprof:oso/9780195092509.001.0001

Widder, D. G., Nafus, D., Dabbish, L., & Herbsleb, J. (2022, June). Limits and possibilities for "Ethical AI" in open source: A study of deepfakes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2035-2046).

Williams, L. J., & Abdi, H. (2010). Post-hoc comparisons. *Encyclopedia of Research Design*, 1060–1067.

Winter, R., & Salter, A. (2019). DeepFakes: uncovering hardcore open source on GitHub. *Porn Studies*, *7*(4), 382–397. https://doi.org/10.1080/23268743.2019.1642794

Yong, E. (2021, May 3). *9/11 memories reveal how flashbulb memories are made in the brain*. (2008, October 14). *Science*. https://www.nationalgeographic.com/science/article/911-memories-reveal-how-flashbulb-memories-are-made-in-the-brain

Zaller, J. R. (1992). *The Nature and Origins of Mass Opinion*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511818691

Zhang, B., Zhou, J. P., Shumailov, I., & Papernot, N. (2020). On attribution of deepfakes. *arXiv.* https://doi.org/10.48550/arXiv.2008.09194

Zmigrod, L., Burnell, R., & Hameleers, M. (2023). The misinformation receptivity framework. *European Psychologist*. https://doi.org/10.1027/1016-9040/a000519

Appendix A. Recruitment Materials

SUBJECT LINE: UNDERSTANDING GENERATION Z'S PERCEPTION OF CELEBRITY INFLUENCERS PROMOTING PRODUCTS ON SOCIAL MEDIA

Dear Prospective Participant,

Lead Researcher, Nika Nour, and researchers from the Informatics Department at the University of California, Irvine are recruiting participants for a research study about the influence of celebrities promoting products on social media. This study may help us to better understand Generation Z's perception of marketing, promotional material, and brand engagements through video clips.

You are eligible to participate in this study if you were born between 1997-2006 and at least 18+years of age or older.

The study will take place online and over email. Your participation will include an initial survey (10-15 minutes) and a post-test survey (5-10 minutes) over the course of four days.

If you fully complete the study, you will receive extra credit for your participation in this research. If you do not wish to participate, the class will have alternative extra credit opportunities made available.

If you participate in the study, you may receive course credit for full completion. pending faculty approvals. Additionally, this study may explain consumer insights in developing future products, advertising effectiveness on social media, and the social impact of promotion on the internet.

If you are interested in participating in this study, please complete this pre-screen survey. You will then receive an email to begin the study over email within a week. Please check your spam, junk, and filters in case you do not receive this email. Otherwise you can contact Nika Nour at npnourmo@uci.edu for the link to begin the study or email if you have any questions.

Thank you very much for your time.

Appendix B. Screening Instrument

1. Please enter your email address. [Email Address]

2. Please enter your date of birth. [Month/Day/Year]

3. How many hours, on average, do you spend on social media easy day (not including messaging apps)? [Numeric Input]

4. Which gender do you identify with the most? [Multiple Choice]
    a. Male
    b. Female
    c. Nonbinary
    d. Prefer to Describe [Open Text Box]
    e. Prefer Not to Answer

5. What is the highest level of education you have completed?
    a. Doctorate or Professional Degree
    b. Master's Degree
    c. Bachelor's Degree
    d. Associate Degree
    e. Some College (No Degree)
    f. High School Diploma or Equivalent
    g. No Formal Education Credential
    h. Other [Open Text Box]

6. Do you understand and speak English?
    a. Yes
    b. No

7. Click below if you're interested in participating in the study. In doing so, you agree to participated in the study and to the use of your survey responses as described in the Study Information Sheet available here

    If you do not wish to participate in the study, please close your browser.
    a. Yes, I consent to participating the study and have reviewed the Study Information Sheet

Appendix C. Survey Protocol 1

Condition 1: No Active Debriefing

Survey Instrument 1

You're going to participate in a study to determine if Generation Z users are influenced by celebrity endorsements on social media. You will watch a series of videos of celebrities promoting different products and be asked questions about them.

1. First, we need to check that your speaker/headphones are working. Please click the play button and write the numbers that you hear. If necessary, adjust the volume until you can hear the audio.

   Do not input any spaces in your answer. [Numeric Input]

Next, you will be asked some questions about three social media video clips of celebrities endorsing products, then we will ask you some further questions about the videos. The social media clips will be accompanied by text. The social media clips are less than 10 seconds, and we will ask you about them at the end of the study so please watch each clip before answering the questions.

Please close all your windows and browsers.

2. How knowledgeable are you about the celebrity Paris Hilton?
   a. 1- Not Knowledgeable at All
   b. 2- Slightly Knowledgeable
   c. 3- Moderately Knowledgeable
   d. 4- Very Knowledgeable
   e. 5- Extremely Knowledgeable

3. How knowledgeable are you about the celebrity Elon Musk?
   a. 1- Not Knowledgeable at All
   b. 2- Slightly Knowledgeable
   c. 3- Moderately Knowledgeable
   d. 4- Very Knowledgeable
   e. 5- Extremely Knowledgeable

4. How knowledgeable are you about the celebrity Kim Kardashian?
   a. 1- Not Knowledgeable at All
   b. 2- Slightly Knowledgeable

c. 3- Moderately Knowledgeable
d. 4- Very Knowledgeable
e. 5- Extremely Knowledgeable

Group A: Videos with Deepfake Kim Kardashian

Paris Hilton announces Hilton Honors Program coming to Roblox

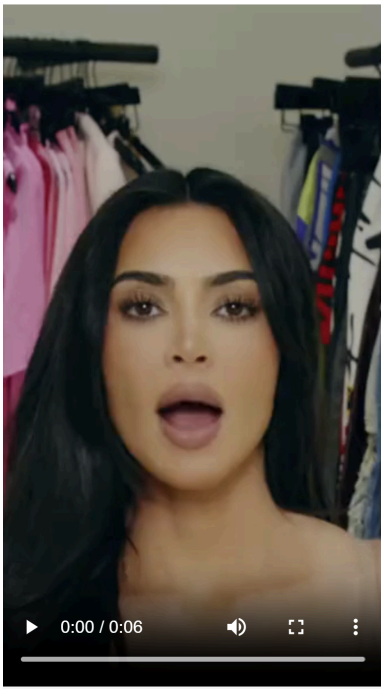
Elon Musk advertises White Claw: An alcoholic beverage

Kim Kardashian launches Eco Edibles: Wearable, edible clothing

Group B: Videos with Deepfake Elon Musk

Paris Hilton announces Hilton Honors Program coming to Roblox

Elon Musk introduces OmniSphere: a venture building Ironman suits



Kim Kardashian launches catsuits for SKIMS shapewear

5. Did you skip any of the videos or watch them without sound?
    a. Yes
    b. No

6. How interested are you in learning more about one or more of the items or services advertised in the videos?
    a. 1- Not Interested at All
    b. 2- Somewhat Interested
    c. 3- Neither Interesteed nor Uninterested
    d. 4- Somewhat Interested
    e. 5- Very Interested

7. How likely are you to purchase one or more of the items advertised in the videos?
    a. 1- Very Unlikely
    b. 2- Somewhat Unlikely
    c. 3- Neither Likely nor Unlikely
    d. 4- Somewhat Likely
    e. 5- Very Likely

8. Have you ever bought a product after a celebrity promoted it on social media?
    a. Yes
    b. No
    c. Maybe
    d. I Don’t Remember
    e. Other [Open Text Box]

9. Which social media platforms do you use the most? Select all that apply.
    a. Facebook
    b. X (formerly known as Twitter)
    c. TikTok
    d. YouTube
    e. Instagram
    f. SnapChat
    g. LinkedIn
    h. Pinterest
    i. Reddit
    j. Discord
    k. I don’t use social media
    l. Other [Open Text Box]

Before you begin this survey please close all other browsers.

10. To the best of your knowledge, mark all of the claims that are true about Paris Hilton. [Check all that apply]
    a. Paris Hilton launches products in Roblox.
    b. Paris Hilton starred on "The Simple Life."
    c. Paris Hilton is launching her own airline company.
    d. Paris Hilton performed on Dancing with the Stars.
    e. Paris Hilton has released music albums and singles.
    f. Paris Hilton invented the selfie.
    g. Paris Hilton's family is known for the Hilton Hotel chain.

11. To the best of your knowledge, mark all of the claims that are true about Elon Musk. [Check all that apply]
    h. Elon Musk recently bought Twitter.
    i. Elon Musk is one of PayPal's early co-founders.
    j. Elon Musk is building a real-life Ironman suit.
    k. Elon Musk is known for being the CEO of Tesla Motors.

l. Elon Musk enjoys drinking White Claws.
m. Elon Musk created a robot companion called "Anne Droid."
n. Elon Musk is on a quest to become a superhero.

12. To the best of your knowledge, mark all of the claims that are true about Kim Kardashian. [Check all that apply]

o. Kim Kardashian is a famous reality TV star.
p. Kim Kardashian is launching EcoEdibles, a new clothing line.
q. Kim Kardashian produces shapewear and undergarments.
r. Kim Kardashian is studying to become a lawyer.
s. Kim Kardashian is creating wearable, edible apparel.
t. Kim Kardashian has her own cryptocurrency called "Kardashian Kash."
u. Kim Kardashian has multiple sisters with various business ventures.

Thank you for completing the survey. We will reach out to you in 24 hours to follow up with a second survey. Please check your email, spam, and junk mail if you have not received an email in 24 hours.

If you have any questions regarding this study, please email npnourmo@uci.edu.

Survey Instrument 2: Post-Test

The real purpose of this study is to examine the correlation between deepfakes and false memories. Some people who participated in this study may have been shown deepfakes (videos that were entirely fabricated and may have involved doctored images or videos). Below, please rate whether you think the social media clips you saw were true or false.

1. We need to check that your speaker/headphones are working. Please click the play button and write the numbers that you hear. If necessary, adjust the volume until you can hear the audio.

Please do not input any spaces in your answer. [Numeric Input]

Group A: Videos with Kim Kardashian Deepfake

2. Is this video real or fake?
    a. Real
    b. Fake

Paris Hilton announces Hilton Honors Program coming to Roblox

3. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
   Paris Hilton announces Hilton Honors Program coming to Roblox

4. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

5. Is this video real or fake?
    a. Real
    b. Fake

Elon Musk advertises White Claw: An alcoholic beverage

6. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
   Elon Musk advertises White Claw: An alcoholic beverage

7. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

8. Is this video real or fake?
    a. Real
    b. Fake

Kim Kardashian launches Eco Edibles: Wearable, edible clothing

9. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
   Kim Kardashian launches Eco Edibles: A wearable, edible clothing line

10. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

Group B: Videos with Elon Musk Deepfake

11. Is this video real or fake?
    a. Real
    b. Fake

Paris Hilton announces Hilton Honors Program coming to Roblox

12. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
    Paris Hilton announces Hilton Honors Program coming to Roblox

13. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

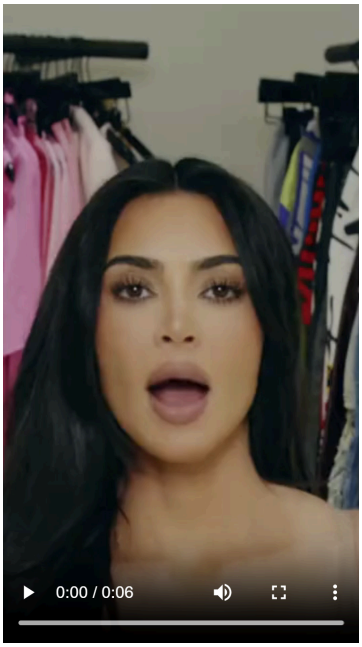
14. Is this video real or fake?
    a. Real
    b. Fake

Elon Musk introduces OmniSphere: a venture building Ironman suits

15. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
    Elon Musk introduces OmniSphere: a venture building Ironman suits

16. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

17. Is this video real or fake?
    a. Real
    b. Fake



Kim Kardashian launches catsuits for SKIMS shapewear

18. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
    Kim Kardashian launches catsuits for SKIMS shapewear

19. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

You will now view the video clips that were doctored and entirely fabricated, also known as deepfakes. Deepfakes are highly-realistic forms of doctored videos where the audio, visual, or both audio and visual components of a video are altered to demonstrate an event as true when in fact it's completely false.

Group A: Videos with Kim Kardashian Deepfake

THIS VIDEO IS A DEEPFAKE - completely falsified and not true

Kim Kardashian launches Eco Edibles: Wearable, edible clothing

Group B: Videos with Elon Musk Deepfake

THIS VIDEO IS A DEEPFAKE - completely falsified and not true

Elon Musk introduces OmniSphere: a venture building Ironman suits

20. Prior to this study, did you consider yourself knowledgeable about deepfake technology? (0-Not Knowledgeable At All to 10-Extremely Knowledgeable) [Sliding Scale]

21. Have you ever looked at deepfake videos online before?
    a. Yes
    b. No
    c. I'm not sure
    d. I didn't know what a deepfake was

22. After completing this survey, how do you feel about deepfake videos? [Open Text Box]

23. At any point during the study, did you search or Google information about the celebrities, claims, or products mentioned?
    a. Yes
    b. No
    c. I can't remember

Thank you so much for participating in this study. Your participation was very valuable to us. We know you are very busy and very much appreciate the time you devoted to participating in this study.

There was some information about the study that we were not able to discuss with you prior to the study, because doing so probably would have impacted your actions and thus skewed the study results. I would like to explain these things to you now.

In this study, we were interested in understanding the correlation between deepfakes and false memories. Based on prior research, we expect to find that exposure to deepfakes could lead to the creation of false memories, especially if participants are not actively debriefed.

You were told that the purpose of the study was Understanding Generation Z's Perception of Celebrity Endorsed Products on Social Media; however, in reality, some of the videos were deepfakes created for the purpose of this experiment and they are not real. This deception was necessary because informing you beforehand could have altered your responses and the accuracy of our findings.

It is very important that you do not discuss this study with anyone else until the study is complete. Our efforts will be greatly compromised if participants come into this study knowing what it is about and how the ideas are being tested. If you have any questions or concerns, you may contact the lead researcher at npnourmo@uci.edu. Thank you again for your participation!

Now that you know the true reason of the research study, if you do not wish to include your data please email npnourmo@uci.edu.

Appendix D. Survey Protocol 2

Condition 2: With Active Debriefing

Survey Instrument 1

You're going to participate in a study to determine if Generation Z users are influenced by celebrity endorsements on social media. You will watch a series of videos of celebrities promoting different products and be asked questions about them.

1. First, we need to check that your speaker/headphones are working. Please click the play button and write the numbers that you hear. If necessary, adjust the volume until you can hear the audio.

   Do not input any spaces in your answer. [Numeric Input]

Next, you will be asked some questions about three social media video clips of celebrities endorsing products, then we will ask you some further questions about the videos. The social media clips will be accompanied by text. The social media clips are less than 10 seconds, and we will ask you about them at the end of the study so please watch each clip before answering the questions.

Please close all your windows and browsers.

2. How knowledgeable are you about the celebrity Paris Hilton?
   a. 1- Not Knowledgeable at All
   b. 2- Slightly Knowledgeable
   c. 3- Moderately Knowledgeable
   d. 4- Very Knowledgeable
   e. 5- Extremely Knowledgeable

3. How knowledgeable are you about the celebrity Elon Musk?
   a. 1- Not Knowledgeable at All
   b. 2- Slightly Knowledgeable
   c. 3- Moderately Knowledgeable
   d. 4- Very Knowledgeable
   e. 5- Extremely Knowledgeable

4. How knowledgeable are you about the celebrity Kim Kardashian?
   a. 1- Not Knowledgeable at All

b. 2- Slightly Knowledgeable
c. 3- Moderately Knowledgeable
d. 4- Very Knowledgeable
e. 5- Extremely Knowledgeable

Group A: Videos with Deepfake Kim Kardashian

Paris Hilton announces Hilton Honors Program coming to Roblox

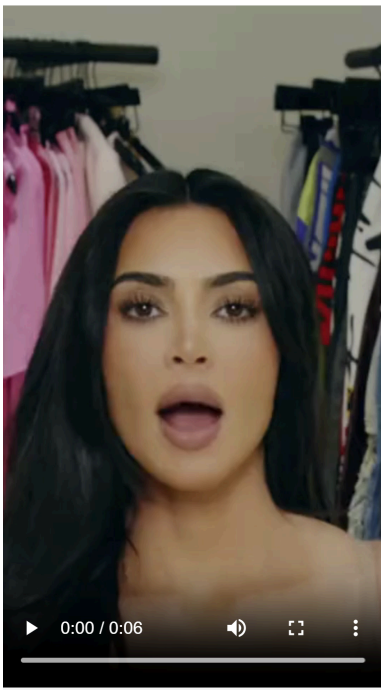
Elon Musk advertises White Claw: An alcoholic beverage

Kim Kardashian launches Eco Edibles: Wearable, edible clothing

Group B: Videos with Deepfake Elon Musk

Paris Hilton announces Hilton Honors Program coming to Roblox

Elon Musk introduces OmniSphere: a venture building Ironman suits



Kim Kardashian launches catsuits for SKIMS shapewear

5. Did you skip any of the videos or watch them without sound?
    a. Yes
    b. No

6. How interested are you in learning more about one or more of the items or services advertised in the videos?
    a. 1- Not Interested at All
    b. 2- Somewhat Interested
    c. 3- Neither Interesteed nor Uninterested
    d. 4- Somewhat Interested
    e. 5- Very Interested

7. How likely are you to purchase one or more of the items advertised in the videos?
    a. 1- Very Unlikely
    b. 2- Somewhat Unlikely
    c. 3- Neither Likely nor Unlikely
    d. 4- Somewhat Likely
    e. 5- Very Likely

8. Have you ever bought a product after a celebrity promoted it on social media?
    a. Yes
    b. No
    c. Maybe
    d. I Don't Remember
    e. Other [Open Text Box]

9. Which social media platforms do you use the most? Select all that apply.
    a. Facebook
    b. X (formerly known as Twitter)
    c. TikTok
    d. YouTube
    e. Instagram
    f. SnapChat
    g. LinkedIn
    h. Pinterest
    i. Reddit
    j. Discord
    k. I don't use social media
    l. Other [Open Text Box]

The real purpose of this study is to examine the correlation between deepfakes and false memories. Some people who participated in this study may have been shown deepfakes (videos that were entirely fabricated and may have involved doctored images or videos). Below, please rate whether you think the social media clips you saw were true or false.

10. We need to check that your speaker/headphones are working. Please click the play button and write the numbers that you hear. If necessary, adjust the volume until you can hear the audio.

Please do not input any spaces in your answer. [Numeric Input]

Group A: Videos with Kim Kardashian Deepfake

11. Is this video real or fake?
    a. Real
    b. Fake

Paris Hilton announces Hilton Honors Program coming to Roblox

12. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
Paris Hilton announces Hilton Honors Program coming to Roblox

13. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

14. Is this video real or fake?
    a. Real
    b. Fake

Elon Musk advertises White Claw: An alcoholic beverage

15. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
Elon Musk advertises White Claw: An alcoholic beverage

16. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

17. Is this video real or fake?
    a. Real
    b. Fake

Kim Kardashian launches Eco Edibles: Wearable, edible clothing

18. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
Kim Kardashian launches Eco Edibles: A wearable, edible clothing line

19. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

Group B: Videos with Elon Musk Deepfake

20. Is this video real or fake?
    a. Real
    b. Fake

Paris Hilton announces Hilton Honors Program coming to Roblox

21. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
Paris Hilton announces Hilton Honors Program coming to Roblox

22. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

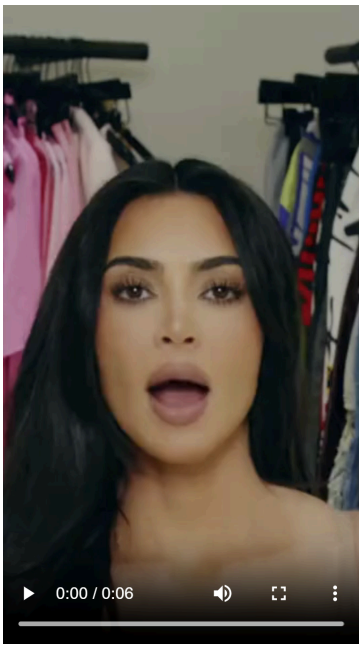
23. Is this video real or fake?
    a. Real
    b. Fake

Elon Musk introduces OmniSphere: a venture building Ironman suits

24. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
    Elon Musk introduces OmniSphere: a venture building Ironman suits

25. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

26. Is this video real or fake?
    a. Real
    b. Fake



Kim Kardashian launches catsuits for SKIMS shapewear

27. How confident are you in your answer? (0- Not Confident At All to 10- Extremely Confident) [0-10 Sliding Scale]
    Kim Kardashian launches catsuits for SKIMS shapewear

28. What cues and clues (i.e. features) lead you to believe this video is real or fake? Please be as specific as possible. [Open Text Box]

You will now view the video clips that were doctored and entirely fabricated, also known as deepfakes. Deepfakes are highly-realistic forms of doctored videos where the audio, visual, or both audio and visual components of a video are altered to demonstrate an event as true when in fact it's completely false.

Group A: Videos with Kim Kardashian Deepfake

THIS VIDEO IS A DEEPFAKE - completely falsified and not true

Kim Kardashian launches Eco Edibles: Wearable, edible clothing

Group B: Videos with Elon Musk Deepfake

THIS VIDEO IS A DEEPFAKE - completely falsified and not true

Elon Musk introduces OmniSphere: a venture building Ironman suits

29. Prior to this study, did you consider yourself knowledgeable about deepfake technology? (0-Not Knowledgeable At All to 10-Extremely Knowledgeable) [Sliding Scale]

30. Have you ever looked at deepfake videos online before?
    a. Yes
    b. No
    c. I'm not sure
    d. I didn't know what a deepfake was

31. After completing this survey, how do you feel about deepfake videos? [Open Text Box]

Thank you for completing the survey. We will reach out to you in 24 hours to follow up with a second survey. Please check your email, spam, and junk mail if you have not received an email in 24 hours.

If you have any questions regarding this study, please email npnourmo@uci.edu.

Survey Instrument 2: Post-Test

The real purpose of this study is to examine the correlation between deepfakes and false memories. Some people who participated in this study may have been shown deepfakes (videos that were entirely fabricated and may have involved doctored images or videos). Below, please rate whether you think the social media clips you saw were true or false.

1. To the best of your knowledge, mark all of the claims that are true about Paris Hilton. [Check all that apply]
    a. Paris Hilton launches products in Roblox.
    b. Paris Hilton starred on "The Simple Life."
    c. Paris Hilton is launching her own airline company.
    d. Paris Hilton performed on Dancing with the Stars.
    e. Paris Hilton has released music albums and singles.
    f. Paris Hilton invented the selfie.
    g. Paris Hilton's family is known for the Hilton Hotel chain.

2. To the best of your knowledge, mark all of the claims that are true about Elon Musk. [Check all that apply]
    a. Elon Musk recently bought Twitter.
    b. Elon Musk is one of PayPal's early co-founders.
    c. Elon Musk is building a real-life Ironman suit.
    d. Elon Musk is known for being the CEO of Tesla Motors.
    e. Elon Musk enjoys drinking White Claws.
    f. Elon Musk created a robot companion called "Anne Droid."
    g. Elon Musk is on a quest to become a superhero.

3. To the best of your knowledge, mark all of the claims that are true about Kim Kardashian. [Check all that apply]
    a. Kim Kardashian is a famous reality TV star.
    b. Kim Kardashian is launching EcoEdibles, a new clothing line.
    c. Kim Kardashian produces shapewear and undergarments.
    d. Kim Kardashian is studying to become a lawyer.
    e. Kim Kardashian is creating wearable, edible apparel.
    f. Kim Kardashian has her own cryptocurrency called "Kardashian Kash."
    g. Kim Kardashian has multiple sisters with various business ventures.

4. At any point during the study, did you search or Google information about the celebrities, claims, or products mentioned?
    a. Yes

b. No
c. I can't remember

Thank you so much for participating in this study. Your participation was very valuable to us. We know you are very busy and very much appreciate the time you devoted to participating in this study.

There was some information about the study that we were not able to discuss with you prior to the study, because doing so probably would have impacted your actions and thus skewed the study results. I would like to explain these things to you now.

In this study, we were interested in understanding the correlation between deepfakes and false memories. Based on prior research, we expect to find that exposure to deepfakes could lead to the creation of false memories, especially if participants are not actively debriefed.

You were told that the purpose of the study was Understanding Generation Z's Perception of Celebrity Endorsed Products on Social Media; however, in reality, some of the videos were deepfakes created for the purpose of this experiment and they are not real. This deception was necessary because informing you beforehand could have altered your responses and the accuracy of our findings.

It is very important that you do not discuss this study with anyone else until the study is complete. Our efforts will be greatly compromised if participants come into this study knowing what it is about and how the ideas are being tested. If you have any questions or concerns, you may contact the lead researcher at npnourmo@uci.edu. Thank you again for your participation!

Now that you know the true reason of the research study, if you do not wish to include your data please email npnourmo@uci.edu.

Appendix E. Codebook: Feelings About Deepfakes

Theme	Code	Definitions	Example
Innovation in Technology	Advancing in quality	The progressive enhancement of deepfake technology in terms of quality, speed, and realism, making them increasingly convincing over time.	" I think that as AI grows in popularity and quality the creation of these videos will become even more realistic" "they are gradually becoming more convincing and may reach a point where a person familiar with deepfakes would not question its authenticity."
Innovation in Technology	Easy to create	The increasing accessibility and simplicity of tools and software that allow users to generate deepfake content	"they are easy to create" "it seems like video editing is so good that it can be easy to create deepfake videos." "anyone can create them about anyone"
Innovation in Technology	Realistic	The deepfake appearing convincing to viewers, closely mimicing authentic content in appearance, sound, or behavior.	"things could even look more realistic" "they can be created to seem even more real if given a realistic-sounding script." "Very convincing, i knew they exist but some of these with celebrities is very convincing" "they could easily be convincing"
Innovation in Technology	Technological advancements	Continuous development and improvement of deepfake technology	"given where the technology is currently, maybe not so much a few years from now" "especially as technology gets better" "I think the technology has a long way to go but I know will convince some"
Innovation in Technology	Use of AI	The application of artificial intelligence and machine learning techniques	"as AI gets more advanced these videos become more realistic " " I realized that AI was stepping up its game with AI generated images and music"

Innovation in Technology	Underwhelmed by quality	Reaction to deepfakes that fail to convincingly replicate realistic features, resulting in disappointment or diminished expectations of the technology, often due to noticeable flaws, glitches, or a lack of realism.	"there are a lot of details that indicate videos that are deepfake." "while most deepfake programs don't work well" " but also not that great in capturing like actual facial movements."
Awareness	Prior knowledge of deepfakes	Pre-existing awareness or understanding of what deepfakes are	"I have been aware of the dangers" "I've heard of deepfake videos being made for revenge and to frame people "
Awareness	Prior knowledge of celebrity/person	Awareness or familiarity with the individual featured in a deepfake	" I have to really think about if a celebrity of their status would be advertising " " it does sound like Elon Musk."
Awareness	Heightened awareness	An increased level of vigilance or sensitivity to the possibility of manipulated media, prompting closer examination and critical thinking when evaluating the authenticity of content.	" if I wasn't paying closer attention, I don't think I would've fully realized it was a deepfake." "I won't believe too fast anymore" "I should be more aware of it" " I could probably be fooled if i wasn't looking for deepfakes specifically"
Awareness	Lack of deepfake awareness	Limited or no prior understanding of deepfake technology and its capabilities	" I see how they may become issues in the future especially now to those unaware" "they are easily mistaken for an actual video if one isn't paying attention" "That someone who is not as educated as me on deepfakes may believe these videos."
Detection	Easy to detect	The ability to identify a deepfake quickly and with minimal effort	"unless the video is obviously fake" "I feel like there are small signs that indicate that a video is a deepfake"

Detection	Difficult to detect	Challenges in identifying a deepfake due to its high quality and realism, making it hard to distinguish from authentic content.	"it is a bit tricky to spot them as i had to really think about it" " I'm not surprised that I wasn't able to realize the video above was a deepfake the first time around."
Detection	Confident detection	A strong belief in one's ability to recognize deepfakes accurately	"I was and am able to spot them" "but confident in my own ability to identify them"
	Valence: Positive	An overall positive emotional response to deepfakes, such as amusement, fascination, or appreciation for the technology's capabilities.	"It's pretty cool technology." "Very interesting"
	Valence: Neutral	A lack of strong emotional reaction to deepfakes, indicating indifference or a balanced perspective that neither leans positive nor negative.	"They're still identifiable and don't look fully real." "fine" "not very convincing"
	Valence: Negative	An overall negative emotional response to deepfakes, such as fear, discomfort, or disapproval, often tied to perceived risks or ethical concerns.	"There is a feeling I have when I watch this like fear and anxiety." "I think it's scary how accurate they've become and how much more believable they might get in the future"
Emotional Reactions	Alarming	Concern or urgency due to perceived risks or dangers.	"I feel as if they can be a bit alarming to a degree"
Emotional Reactions	Annoying	Irritation or frustration	"They're annoying, as you have to pay even more attention"
Emotional Reactions	Bad	Perceived as harmful or undesirable,, evoking disapproval or dissatisfaction.	"That it is bad" "Very scary as bad things may happen"
Emotional Reactions	Concerned	Expresses or apprehension about potential impacts.	"I'm both concerned and afraid " "because it is concerning how easily a fake message can be produced"
Emotional Reactions	Cool	Perceived as novel or appealing in a positive way.	"They are cool but at the same time .."

Emotional Reactions	Creepy	Elicits discomfort or fear due to something unnatural or eerie.	"it is kinda cool how you can deepfake certain things"
Emotional Reactions	Curious	A desire to learn more.	"make you wonder how many things are being doctored and altered"
Emotional Reactions	Dangerous	Perceived as a threat or source of harm.	"I think deepfake videos are extremely dangerous"
Emotional Reactions	Dislike	(Includes hate) Evokes strong disapproval or aversion.	" I strongly dislike them"
Emotional Reactions	Disgust	Feelings of revulsion or intense disapproval.	"quite frankly a bit disgusting"
Emotional Reactions	Disturbing	Evokes discomfort or distress due to irksome content.	" I find them disturbing"
Emotional Reactions	Good	Viewed positively or favorably.	"deepfakes are becoming more and more real every day, which can be both good and bad"
Emotional Reactions	Interesting	User's attention is peaked without strong emotional valence.	" It is pretty interesting tech"
Emotional Reactions	Impressed	Admiration or appreciation.	"Very interesting"
Emotional Reactions	Sad	Feelings of sorrow or disappointment.	"I still feel sad that I have to pay attention"
Emotional Reactions	Scared	(Includes scary) Fear or apprehension due to perceived risks or threats.	"I feel a little scared" "It is kind of scary to see how realistic deepfake videos can look"
Emotional Reactions	Shocking	Elicits strong surprise or disbelief, especially upon realizing a video thought to be real is actually a deepfake or how realistic the content looks.	"it is pretty shocking how far they have come"
Emotional Reactions	Surprised	Reacting to something unexpected.	"I'm very surprised"
Emotional Reactions	Unchanged perception	No significant emotional, new opinion, or cognitive shift.	"Just about the same I already did"

Emotional Reactions	Uncomfortable	Elicits a sense of unease, awkwardness, or mild apprehension, often caused inappropriate elements.	"they are also dangerous and make people very uncomfortable"
Emotional Reactions	Unsettling	(Includes uncanny) Creates a profound sense of unease or discomfort due to unnatural or incongruous elements.	"I worry how much longer we have until we can no longer properly distinguish falsified/doctored videos from real humans speaking."
Emotional Reactions	Weird	(Includes strange) Perceived as odd, unusual, or out of place.	"It's very weird"
Emotional Reactions	Worried	Expresses concern or anxiety about potential risks, the future, or outcomes.	"then I would be very worried about the future of information and disinformation"
Negative Impacts	Misinformation spread	Focuses on the dissemination of false or misleading information.	"spreading false campaigns or promotions" "a serious threat to the credibility of news and information sources" "being able to spread false information by just placing your face over someone else's"
Negative Impacts	Consumer deception	The intentional act of misleading individuals specifically in consumer-related contexts, such as encouraging purchases, altering opinions about products or services, or spreading false endorsements, often for personal or financial gain.	"It's easy to be used to deceive people" "They can be highly deceptive" "celebrities can influence some people's decision making"
Negative Impacts	Scams	Covers financially or socially exploitative activities, such as scams or phishing schemes	" can lead to misguidance and scams" "no upside other than to scam people for the most part"
Negative Impacts	Reputational harm	The deliberate or unintentional damage to an individual’s or entity’s reputation through the misuse of deepfake content.	"I also feel like it can be damaging to the person's image"

Negative Impacts	Sexual harassment	Non-consensual or harmful uses involving sexualized or explicit content, especially when targeting specific individuals.	"Especially when it comes to women, where an unfortunate reality is that real women's faces can be used in porn without their consent and distributed"
Negative Impacts	Societal Threats	A belief that deepfakes pose significant harm and risks to societal trust, stability, ethics, and democratic processes, particularly by exploiting vulnerable groups, outweighing potential benefits.	"I know there have been incidents in Korea " "I fear every generation—young students, elderly on Facebook, parents, etc.—falling victim to false ads or misinformation online" "the potential to cause issues in the future"
Negative Impacts	Manipulative	The deliberate alteration or fabrication of content designed to influence perceptions, opinions, or behaviors, often by exploiting trust. (includes subtle tactics that may not have overly harmful intentions but can distort reality)	"anyone can make videos of anyone saying anything" " I wouldn't want someone using a video of me for a product I'm not being paid to endorse"
Negative Impacts	Trickery	The ease with which some individuals or groups might believe or fall for falsified or manipulated content.	"It is likely easy to get tricked by a deepfake video from not be knowledge of the context with it" "a fool a bunch of people."
Negative Impacts	Malicious intent	The intentional use of deepfakes for unethical or harmful purposes and behaviours, including coercion, blackmail, or other forms of exploitation.	"if used in a harmful way." "it could be used to spread fake news in certain intention" " can be used for evil"
Legal and Regulatory Frameworks	Disclosure necessary	Belief that the use of deepfake technology should be transparently disclosed to audiences	"Unless it is explicitly stated that a video is a deepfake"
Legal and Regulatory Frameworks	Need for literacy	The necessity of educating individuals about deepfake technology and its capabilities,	"people should be learning more about deepfakes"

Legal and Regulatory Frameworks	Legal and criminal considerations	The ways deepfake technology can be misused to violate laws, infringe on rights, exploit individuals without their consent, or enable illegal activities	"my identity being stolen for any purpose whatsoever" "They should be illegal."
Legal and Regulatory Frameworks	Government regulation	The role of government policies and legislation in managing and mitigating the risks associated with deepfake technology	"if not controlled by some way or regulated"
Legal and Regulatory Frameworks	Prevention systems/of distribution	Measures and technologies designed to limit the dissemination of harmful deepfakes	" it needs to be facilitated to reduce fake information" "but must not be misused"
Perception of Usefulness	Pointless	The perception that deepfakes lack meaningful or constructive purpose, seen as unnecessary or irrelevant	"I think they're stupid" "Creating deepfakes aren't usually used for anything positive"
Perception of Usefulness	No reason to exist	A stronger perception that deepfakes should not exist at all due to lack of utility or inherent risks.	"I don't really see any benefit in deepfake videos"
Perception of Usefulness	Used for fun/humor	A belief that deepfakes can be used in harmless, entertaining, or humorous ways, often for creative or lighthearted purposes.	"i could see them being manipulative in serious contexts and also pretty funny otherwise"

Appendix F. Codebook: Cues and Clues

Theme	Physical Realism	Definition	Example
Physical Realism	Body Movements	Observations about the naturalness of body, head, or hand movements.	"The way she moves her neck is unnatural." "Her body stays in the same position, only her head moves." "Her head bobs naturally when she talks."
Physical Realism	Eye Movements	Observations about eye tracking, focus, or blinking patterns.	"Her eyes stare directly into the camera." "Her eye movements look too fake to be real." "Her eyes blink very unnaturally, like a doll."
Physical Realism	Mouth Movements	Observations about the way the mouth moves, including unnatural or exaggerated articulation.	"Her mouth movements look like they match up."
Physical Realism	Lip Sync	Observations about whether the lip movements align with the audio.	"The lips and the talking are not in sync."
Physical Realism	Facial Features	Observations about the shape, proportions, or specific elements of the face, such as skin texture or symmetry.	"Her face feels fake but it’s probably due to plastic surgery." "Her skin seems rubbery and unnatural."
Physical Realism	Facials Expressions	Observations about the naturalness and variability of facial expressions during the video.	"Her facial expressions and lip movements match her voice." "Her expressions seem genuine and human-like.
Physical Realism	Fine Details	Observations about the movement or appearance of small details. (hair, fingers, etc)	"Her hair moves a bit weird with the background."
Physical Realism	Mannerism Cues	Observations about naturalness in behavior, gestures, or tone, or demeanor.	"Her mannerisms seem natural." "She talks in a robotic and monotone way." "Her gestures are being repeated."

Physical Realism	Verbal Expression	Observations about voice intonation, speech rhythm, patterns, or delivery.	"The speech seems slurred in a way." "Her voice matches her facial expressions." "The voice tone doesn't sound like her usual self."
Theme	**Technical Quality**	**Definition**	**Example**
Technical Quality	Background Factors	Observations about the background or lighting.	"The background doesn't warp, which usually happens in AI videos."
Technical Quality	Audio artifacts	Observations about the quality, clarity, or naturalness of the audio, including distortions, synthetic tones, or other anomalies	The audio sounds very digital in a way.
Technical Quality	Visual Artifacts	Glitches, distortions, or technical imperfections in the video.	"There was a distorted area in her neck in the video."
Technical Quality	Production Quality	Observations about the editing or camera behaviour.	"The camera movement seems strange and indicates editing."
Technical Quality	Video Quality	Observations about the overall resolution or clarity of the video.	"The video resolution is pretty high." "The image quality is too smooth, which makes it feel fake."
Theme	**Contextual Plausibility**	**Definition**	**Example**
Contextual Plausibility	Confirmation Bias	Judgments influenced by expectations or context. Further: The tendency to interpret new information in a way that confirms pre-existing beliefs, expectations, or attitudes, often disregarding contradictory evidence.	"Knowing her persona, this is something she would do." "It seems real because it's a well-known product." "Knowing her persona, this is something she would do." "This fits with everything I know about Elon Musk, so it must be real."

Contextual Plausibility	Prior Knowledge Bias	Influence of prior knowledge or assumptions on judgment. Further: The influence of factual or perceived prior knowledge on the evaluation of new information, which may lead to acceptance or skepticism of the content.	"I know Kim Kardashian owns SKIMS, so this seems real." "I've seen Roblox collaborations with celebrities before."
Contextual Plausibility	Plausibility of Claim	Observations about the reasonableness of the product, message, or scenario.	"A hotel program being on Roblox makes no sense." "This product doesn't align with her character or demographic."
Contextual Plausibility	Fit with the Celebrity	Observations about the plausibility of the celebrity endorsing the content	"Doesn't seem like something she would do." "I've never seen Paris Hilton participate in an ad for a video game."
Theme	**Anomalies**	**Definition**	**Example**
Anomalies	Overly-Perfect	Features that seem overly smooth, polished, or idealized.	"Her face looks too perfect, almost cartoonish." "The audio is too clean, which makes it seem fake."
Anomalies	Uncanny Valley	Intuitive feelings of unnaturalness without specific glitches.	"It feels off but I can't pinpoint why." "The video has that weird deepfake feeling."
Theme	**Overall Perception**	**Definition**	**Example**
Overall Perception	Doubt	Statements expressing skepticism, uncertainty, or indecision about authenticity.	"It looks fake but I think it's real." "I'm not sure if this is real or fake."
Overall Perception	Realism	Statements indicating the video feels authentic.	"Her movements looked very real." "I would be shocked if this was a deepfake."
Overall Perception	Unnatural	Observations that describe elements of the video or subject that evoke a sense of being artificial, fake out of place, or inconsistent	"It looks robotic and there are cuts in between."
Theme	**Other**		