CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	The impact of deepfake exposure on beliefs, memory, and behaviour: investigation and intervention
Authors	Ching, Didier
Publication date	2025-12-09
Original Citation	Ching, D. 2025. The impact of deepfake exposure on beliefs, memory, and behaviour: investigation and intervention. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2025, Didier Ching. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-17 04:28:06
Item downloaded from	https://hdl.handle.net/10468/19275

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

The Impact of Deepfake Exposure on Beliefs, Memory, and Behaviour: Investigation and Intervention

Didier Ching

Submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Applied Psychology

School of Applied Psychology

College of Arts, Celtic Studies, and Social Sciences

University College Cork

9th December 2025

Supervisors:

Dr. Gillian Murphy

Dr. Conor Linehan

Head of School: Professor Chris McCusker

Abstract

This thesis seeks to understand how deepfake exposure affects beliefs, memories, and behaviours, and how best to counteract the negative consequences of misleading deepfakes. A comprehensive multi-study approach, comprising a scoping review, a thematic analysis of real-world discourse, and five rigorous quantitative experiments (N=7,666) have established that despite prevailing alarmist narratives, deepfakes are not a uniquely potent form of misinformation compared to simple text and audio. Rigorous comparative experiments, conducted during an active U.S. election period, consistently demonstrated that deepfakes were not especially effective at influencing political beliefs, memories, or voting intentions, nor did they produce lasting longitudinal effects. Crucially, this thesis demonstrates that beyond the direct manipulative harms of deepfake exposure, the surrounding alarmist discourse may generate significant unintended consequences, specifically the erosion of epistemic trust in authentic video-based information. To address this, the thesis culminated in the design and evaluation of an evidence-based intervention to educate the public about deepfakes without undermining epistemic trust. This novel strategy successfully reduced deepfake anxiety and improved discernment of authentic video content, providing a critical roadmap for the field to transition from fear-based rhetoric to informed, balanced media literacy.

Table of Contents

Chapter 1 Introduction **1**

1.1 What are Deepfakes? 1

1.2 Why Beliefs, Memories, and Behaviours Matter in Deepfake Research 6

1.3 Epistemic Harms of Deepfakes 10

1.4 Conceptual and Methodological Challenges in Deepfake Research 12

1.5 Methodological Challenges 14

1.6 Research Questions 16

1.7 Thesis Outline 17

1.8 References 20

Chapter 2 Can deepfakes manipulate us? Assessing the evidence via a critical scoping review **28**

2.1 Abstract 29

2.2 Introduction 29

2.3 Literature Review 31

2.4 Materials and Methods 37

2.5 Results 40

2.6 Discussion 54

2.7 References 63

Chapter 3 Encountering Deepfakes: A Thematic Analysis of Comments on a Music Video featuring Deepfake Content **73**

3.1 Abstract 75

3.2 Introduction 76

3.3 Background 78

3.4 Methods 85

3.5 Results 90

3.6 Discussion 100

3.7 Conclusion 108

3.8 References 109

Chapter 4 Deepfake Misinformation Effects on Voters in the 2024 US Presidential Election **120**

4.1 Abstract 122

4.2 Introduction 125

4.3 Experiment 1 128

4.4 Results 134
4.5 Experiment 2 136
4.6 Results 138
4.7 Experiment 3 140
4.8 Results 142
4.9 Experiment 4 143
4.10 Results 145
4.11 Experiment 5 146
4.12 Results 148
4.13 Discussion 149
4.14 References 152
Chapter 5 Evaluating Methods to Educate the Public about Deepfake Technology 158
5.1 Abstract 160
5.2 Introduction 161
5.3 Experiment 1 167
5.4 Results 173
5.5 Experiment 2 176
5.6 Results 185
5.7 Experiment 3 190
5.8 Results 193
5.9 Discussion 200
5.10 References 205
Chapter 6 Discussion 215
6.1 Thesis Overview 215
6.2 Contributions 218
6.3 Strengths 226
6.4 Limitations 227
6.5 Future Directions 230
6.6 Conclusion 232
6.7 References 233

List of Figures

Figure 2-1: PRISMA Flow Diagram 39

Figure 4-1. The screenshots of deepfakes of Kamala Harris and Donald Trump shown to participants in the deepfake condition under the "competency" theme. The stories were presented as text, audio, or as a deepfake. The text below each deepfake represents what was said in the video and was not shown to participants in the deepfake or audio conditions.....131

Figure 4-2. Bar charts showing the distribution of (a) feeling thermometer scores towards a politician and (b) voting intention scores across experimental conditions. Both graphs indicate that there are no significant differences across misinformation formats. 135

Figure 4-3. Sample screenshots of deepfakes from the former President endorsement and celebrity endorsement themes of Barack Obama and Elon Musk. The stories were presented as text, audio, or deepfake. The text below each deepfake represents what was said in the video and was not shown in the deepfake and audio conditions. 138

Figure 4-4. Bar charts in Experiment 2 showing no significant differences between (a) attitudes towards a politician and (b) voting intentions among misinformation formats.....139

Figure 4-5. Bar charts in Experiment 3 showing no significant differences between (a) attitudes towards a politician and (b)voting intentions across misinformation formats and across a 3-day time gap between wave measurements. 143

Figure 4-6. Bar charts in Experiment 4 showing no significant differences between attitudes (a) towards a politician and (b) voting intentions across misinformation formats and across a 3-day time gap between wave measurements. 146

Figure 5-1. The screenshots of the high-definition Joe Biden video (top video) and low-definition Joe Biden video (bottom video) shown to participants. 170

Figure 5-2. Bar charts for Experiment 1 showing no significant differences in (a) video belief, (b) sharing intentions, or (c) trust in news and social media across video definition with the

error bars representing 95% confidence intervals, but the experimental condition significantly reduced mean belief in video scores. 175

Figure 5-3. The intervention infographics showing (from top to bottom) the evidence-based intervention, detection-based intervention, and alarmist intervention. 181

Figure 5-4. Bar charts showing the evidence-based intervention leading to (a) no effect on trust in news and social media and reduced (b) negative perceptions of dystopian technologies and (c) deepfake anxiety scores with the error bars representing 95% confidence intervals. 187

Figure 5-5. Screenshots showing the deepfakes used for the deepfake discernment outcome measure with the transcript of each video underneath (the text was not shown to participants). 192

Figure 5-6. Bar charts showing participants who had previously been in the evidence-based condition having (a) trust in news and social media, (b) negative perceptions of dystopian technologies, and (c) deepfake anxiety scores compared to other conditions with the error bars representing 95% confidence intervals. 197

List of Tables

Table 2.1. Descriptive data of all included studies 41

Acknowledgements

My PhD journey over the past four years has encompassed some of the most rewarding and most challenging moments of my life. Through it all, I am grateful to have been surrounded by supportive supervisors, colleagues, friends, and family.

First and foremost, I would like to express my deep gratitude to my supervisors, Gillian and Conor. Thank you, Gillian, for being a steady voice of reason throughout my PhD and for guiding me back on track whenever I felt lost. I have always appreciated the time and patience you invested in supporting me, and your belief in my abilities, especially when I doubted myself, meant more than I can say. It has truly been an honour to work under your supervision. I would also like to thank Conor for being a constant source of wisdom and reassurance over the past four years. I am especially grateful for your kindness and support at the beginning of my PhD, during Gillian's leave, when I had no idea what to expect. Your insights and knowledge have repeatedly given me new perspectives. I genuinely could not have asked for better supervisors than you both.

I am also thankful to have had Matthew and Mike as my secondary supervisors. Our meetings were always incredibly insightful, and I left every conversation with something new to reflect on. I am fortunate to have benefited from the guidance of such supportive and thoughtful supervisors throughout this journey.

I would like to sincerely thank my PhD colleague John. At the outset, I imagined a PhD to be a solitary path, but working alongside you showed me otherwise. You have motivated me to keep working hard from start to finish, and I often turned to you for reassurance during moments of doubt. Your support has meant a great deal, and my experience of the PhD would not have been the same without you.

My thanks also go to all of my fellow PhD students in the PhD room. Your kindness and companionship made this journey far less lonely and intimidating. I especially want to thank Dan and Tadgh for patiently showing me the ropes when I first began teaching.

To the 7,666 participants who took part in my studies, I am grateful for your time and effort. Without your contributions, this thesis would not have been possible.

I would also like to thank Lero, the Science Foundation Ireland Research Centre for Software, for funding my PhD. Without this support, I could not have undertaken this journey. Beyond financial assistance, Lero also provided invaluable opportunities and encouraged me to engage with the public in ways I found deeply fulfilling.

My thanks extend to Aaron and Eoghan for your help in creating the deepfakes used in my research, and to Ting Na and Clodagh for your contributions to my papers. It has been an honour to supervise your projects.

Finally, I want to thank my friends and family for the support they have given me throughout this long journey. Thank you, Alasdair, Samantha, Tim, Michal, and especially Luke, whose reassurance and comfort helped carry me through the toughest moments. I am grateful to my parents, Ben and May, for everything you have done to help me reach this point, and to my brother Eugene for giving me much-needed escapes from work.

Declaration of Originality

This thesis is my own work and has not been submitted for another degree, either at University College Cork or elsewhere.

Didier Ching

Publications presented as part of the thesis

Published

Ching, D., Twomey, J., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2025). Can deepfakes manipulate us? Assessing the evidence via a critical scoping review. *PLoS One, 20*(5), e0320124.

Ching, D., Twomey, J., Linehan, C., & Murphy, G. (2025). Encountering Deepfakes: A Thematic Analysis of Comments on a Music Video featuring Deepfake Content. *Proceedings of the ACM on Human-Computer Interaction, 9*(7), 1-23.

Under review

Ching, D., Twomey, J., Meehan, E., Bolger, A., Linehan, C., Murphy, G. (under review). Deepfake Misinformation Effects on Voters in the 2024 US Presidential Election. *Advances in Psychological Science Open.*

Ching, D., Twomey, J., Linehan., C., Murphy, G. (under review). Evaluating Methods to Educate the Public about Deepfake Technology. *Media Psychology*

Other co-authored publications related to the thesis

Murphy, G., **Ching, D.**, Twomey, J., & Linehan, C. (2023). Face/Off: Changing the face of movies with deepfakes. *Plos one*, *18*(7), e0287503.

Murphy, G., **Ching, D.**, Meehan, E., Twomey, J., Bolger, A., & Linehan, C. (2025). An Average Joe, a Laptop, and a Dream: Assessing the Potency of Homemade Political Deepfakes. *Applied Cognitive Psychology*, *39*(2), e70061.

Twomey, J., **Ching, D.**, Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *Plos one*, *18*(10), e0291668.

Twomey, J., **Ching, D.**, Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2024). What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology. *IEEE Transactions on Technology and Society*.

Twomey, J., **Ching, D.**, O'Mahony, C., Linehan, C., & Murphy, G. (2025, April). I Was Never in the AI. RA: an Experimental Study on the Use of Manipulated Audio for the Censorship of Political Dissidents. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-8)

Chapter 1 Introduction

The aim of this thesis is to investigate how exposure to deepfake videos influences peoples' beliefs, memories, and behaviours, and how educational interventions can mitigate the effects of misleading deepfakes. Media outlets and research papers have frequently raised concerns about the unique and manipulative effects of deepfake exposure (Chesney & Citron, 2019; Dan, 2025; Duboust, 2023). These concerns seem reasonable, as deepfake technology can facilitate the creation of hyper-realistic videos of people doing or saying something that they never did (Westerlund, 2019). When applied to high-stakes environments such as politics or democratic processes, it is easy to imagine how deepfakes could cause harm through manipulating people's beliefs about politicians and erode the public's trust in online video-based information. Despite the obvious potential for deepfakes to spread misinformation and manipulate individuals, at the outset of this PhD, there was surprisingly little empirical research quantifying the effects of deepfakes on peoples' beliefs, memories, or behaviours. Without such information, it is difficult to understand the impact of deepfakes on human cognition and behaviour. Without sufficient evidence, narratives surrounding deepfakes perpetuated by the media and research literature can only be considered premature. This thesis begins to address this gap in our understanding of deepfakes through a series of empirical studies exploring the impact of deepfake exposure on people's beliefs, memories, and behaviours. The project culminates in the exploration and design of an effective intervention aimed at educating the public about the effects of misleading deepfake exposure.

1.1 What are Deepfakes?

1.1.1 Origin of Deepfakes

Deepfakes are commonly defined as videos or audio clips generated using generative adversarial networks (GANs), where one person's face or voice is artificially overlaid onto another. This results in media showing someone saying or doing something they never did (Twomey et al., 2025). According to Mirsky and Lee (2021), GANs are a type of machine learning frameworks that generate new data samples that resemble data it is given. When creating deepfakes, a "generator" neural network creates a face or an audio clip which is then directed towards a "discriminator" network attempting to distinguish the generated material from the real source data. Any discrepancy between generated material and source data is used to teach the generator network to create better materials in an adversarial loop. The term "deepfake" originated as a portmanteau of "deep learning", which references the subfield of machine learning using algorithms inspired by artificial neural networks (LeCun et al., 2015), and "fake" referring to the deceptive nature of deepfakes (Horvitz, 2022). Deepfake videos first appeared on a forum on Reddit in 2017 uploaded by a user named "deepfake" (Vasist & Krishnan, 2022). These deepfakes depicted the faces of female celebrities on the faces of actors in pornographic films (Albahar & Almalki, 2019). The subreddit where these deepfakes were first uploaded was shut down and deepfakes were banned from Reddit entirely (Albahar & Almalki, 2019), but the propagation of deepfake technology was irreversible and deepfakes have proliferated rapidly in the years since (Altuncu et al., 2024). The technology behind deepfakes has also gradually become more accessible to the public over time (Murphy et al., 2025).

Collectively, deepfakes have been perceived by many as a novel way to pollute the public's trust in video-based information and manipulate the public's attitudes towards individuals through hyper-realistic videos, a concern which is only compounded by the accessible nature of this technology (Chesney & Citron, 2019; Vizoso et al., 2021). However, some scholars don't view deepfakes as an isolated and unprecedented form of media

manipulation, but rather, an evolution of pre-existing strategies (Broinowski, 2022; Vasist & Krishnan, 2022). Media such as photos and videos have been manipulated for many years before the creation of deepfakes. For example, photos and videos were frequently manipulated during World War 2 as propaganda in an attempt to influence attitudes and behaviours (Margry, 1992). Recent examples of media manipulation before deepfakes include photoshop (Caldwell, 2016) and innovative video-editing techniques morphing faces and resurrecting deceased celebrities (Ngak, 2012). Research has closely monitored how these manipulated media affect human cognition (Nash et al., 2009; Saachi et al., 2007) and situating deepfakes in a long line of media manipulation strategies is beneficial as we can adapt findings from previous research to understand how deepfake exposure may impact beliefs, memories, and behaviours. This is especially important in understanding deepfake influences due to the lack of empirical studies conducted with deepfake stimuli (Ching et al., 2025).

1.1.2 Narratives around Deepfakes

Both media sources and academic literature often characterize deepfakes as a uniquely dangerous threat to society (Poidevin, 2025; Sunvy et al., 2023; TBIJ, 2024), capable of manipulating election outcomes and threatening national security through inciting violence and undermining trust in government and military institutions (Chesney & Citron, 2019). These accounts frequently emphasise that deepfakes have the unique potential to shape people's beliefs, memories, and behaviours, specifically in political contexts (Wahl-Jorgensen & Carlson, 2021). However, these narratives are often not based on empirical evidence, as my scoping review of empirical deepfake studies revealed that the current body of literature is insufficient in determining precise effects of deepfake exposure on beliefs, memories, and behaviours (Ching et al., 2025). Considering the limited empirical evidence on deepfake exposure effects, contemporary narratives around deepfakes may only be considered as

alarmist and premature. Another narrative suggests that deepfakes are rapidly proliferating and may soon become ubiquitous across online platforms and polluting people's trust in authentic information (Abraham et al., 2025; Romanishyn et al., 2025). Romanishyn et al. (2025) reference projections estimating up to 8 million deepfakes could be shared by the end of 2025, a significant rise from approximately 500,000 shared in 2023. However, these claims are potentially misleading and contribute to alarmist discourse for several reasons. Firstly, between 96-98% of deepfakes online are pornographic in nature (Ajder et al., 2019, SecurityHero, 2023) and thus, it is improbable that the projected eight million deepfakes will be distributed across general online platforms but rather their circulation is likely confined to a relatively small subset of users. Secondly, estimates suggest that misinformation and disinformation constitute only about 5% of the average individual's information consumption in the United States (Acerbi et al., 2022). Considering this, the likelihood of encountering a non-pornographic, manipulative deepfakes within that 5% is extremely low, challenging the concerns implied by current widespread narratives.

Alarmist and unsupported narratives not only distort the public's understanding of deepfakes but can also lead to tangible harm. Some emerging research reveals that deepfake pornography disproportionately targets women, particularly celebrities, highlighting the emotional, reputational, and psychological toll it imposes on its victims (Twomey et al., 2025). McGlynn and Toparlak (2025) extends these harms towards non-celebrity female victims who have experienced similar traumatic experiences with non-consensual deepfakes describing them as "radically altering their everyday life experiences, relationships, and activities". The experiences of harassment and abuse through deepfake pornography is being overlooked, in favour of political harms, by the broader research community and media outlets as evidenced by the relatively smaller amount of attention it is receiving (Gosse & Burkell, 2020). Furthermore, unfounded alarmist narratives risk misinforming the public and generating

unwarranted fear and panic around deepfakes. This is problematic because misinformed citizens who fear the narratives around deepfakes may lobby for quick legislation and policies to restrict deepfakes which may be overly broad or poorly drafted, leading to legal overreach and censoring legitimate speech and satirical expression (Feeney, 2021).

Alarmist narratives around emerging technologies such as deepfakes are not new. Dystopian predictions have been made about many mundane technologies we take for granted today. For example, the telephone had been judged by many as a source of social erosion as it ruined the communal aspect of visiting friends and family in person (Thompson, 2016) and lightbulbs were predicted cause significant harm to the eyes and even blindness (Almond, 2019). Even the humble book and the written word were forecasted as a threat to tradition and human memory promoting "social passivity" (Yadlin-Segal & Oppenheim, 2021). Understanding the origin of alarmist predictions enables us to moderate our expectations and anxieties regarding emerging technologies such as deepfakes and allows us to engage with them through a critical lens rather than fear. Broinowski (2021) proposes that dystopian predictions about technologies are typically a projection of how the technology is experienced today rather than a measured attempt at predicting the future. This proposition is especially relevant with deepfakes due to the perceived threat of misinformation today (Altay et al., 2023). Since 2016, academic research investigating misinformation and fake news has been increasing (Murphy et al., 2023). The United Nations considers misinformation as one of the number one threats to humanity in the global risk report (UN, 2024) and yet there is considerable research suggesting that misinformation only makes up a small part of our information diet (Acerbi et al., 2022), online platforms are not inundated with misinformation (Altay et al., 2023), and one-off exposures to misinformation are far from guaranteed to lead to behavioural change (Greene et al., 2025). This fixation on misinformation may be a contributing factor to how people are

predicting deepfake threats. The predictions are based on perceived threats rather than established empirical research.

1.2 Why Beliefs, Memories, and Behaviours Matter in Deepfake Research

Conducting rigorous empirical research on the effects of deepfake exposure is critical to validating or disproving the narratives around deepfakes. Although empirical research on the cognitive and behavioural effects of deepfake exposure remains limited, insights can be drawn from existing theories and prior studies to inform our understanding. This thesis specifically focuses on examining how deepfakes affect beliefs, memories, and behaviours because these areas constitute the core of how people process and react to information (Ajzen, 1991; Kahneman, 2011; Loftus & Palmer, 1974). For example, in a political context, it is reasonable to believe that deepfakes defaming candidates could influence voters' attitudes towards the candidate, how they remember the reputation of the candidate, and ultimately how they intend to vote for a candidate.

1.2.1 How deepfakes could affect beliefs

Several cognitive theories offer insights as to how viewing deepfakes can potentially affect belief formation in ways that may be problematic. For example, Kahneman (2011) explains how Dual Processing Theory (DPT) distinguishes between two modes of thinking involved in information processing: System 1 and System 2. System 1 describes an automatic and quick way of processing information that requires little effort, relying on heuristics and intuition, while System 2 describes a more deliberate, slow, and effortful thinking relying on logical analysis. Kahneman (2011) suggested that individuals may form erroneous beliefs from

unreliable information if they apply System 1 thinking and process mis/information intuitively using heuristics. This theory can explain how individuals may form erroneous beliefs when viewing deepfakes as viewers may passively use System 1 thinking when processing deepfakes as it reduces cognitive load. Therefore, misinformation presented as deepfakes could be accepted more easily due to the increased fluency of the information. The increased fluency is exemplified by the Modality Effect which suggests that presenting information through multiple channels such as both audio and visual enhances comprehension, as it lessens the cognitive load on any single processing pathway (Tabbers et al., 2004). The focus on the fluent processing of audio-visual information across the cognitive mechanisms suggests that the audio-visual nature of deepfakes may engage a more influential cognitive process, potentially manipulating beliefs and attitudes in ways that differ from traditional media formats such as text or photographs.

Kunda (1990) describes another useful theory in understanding belief formation through “motivated reasoning”, which refers to the tendency of an individual to process information in a way that favours their goals or preferences. Essentially, individuals may be more likely to accept deepfake content as true if it aligns with their pre-existing beliefs. For example, Republican supporting individuals may be more likely to believe a statement made by a Republican politician. Similarly, deepfakes conveying views congruent with the viewer may be more likely to be believed.

Finally, Social Identity Theory (SIT) (Tajfel et al., 2001) suggests that people derive part of their self-concept from group memberships. This affects beliefs by creating in-group favouritism and out-group bias, leading individuals to adopt the norms, values, and opinions of their group. As a result, people may be more likely to form beliefs from deepfakes if the information conveyed supports their social identities.

1.2.2 How deepfakes could affect memories

Deepfakes may induce false memories in a manner comparable to traditional media, such as photographs or text narratives. Loftus & Palmer (1974) explained how post-event information such as altering the terminology used to describe a car accident could affect people's memories of how violent the accident was. This phenomenon was later termed as the "misinformation effect" which theorises how retroactive information can influence an individuals' memories of an event (Loftus, 2005). The Misinformation Effect has consistently been observed, with experiments showing how fictional text narratives and doctored photos can induce false memories in individuals (Garry & Wade, 2005; Wade et al., 2002). Mechanistically, Johnson et al. (1993) theorise that the Misinformation Effect occurs when individuals misattribute the source of their memories. When receiving post-event information, people may mistakenly assign the source of that information to something that they personally experienced, creating a false memory. Garry & Wade (2005) found that participants exposed to a false childhood event described as a narrative were more likely to form detailed false memories than participants who were shown a doctored photograph of the same event, suggesting that textual misinformation influenced memories more than visual. They proposed that narrative descriptions prompt more active cognitive engagement with the false information, increasing processing fluency and thereby enhancing the event's perceived familiarity and plausibility, which in turn facilitates the formation of false memories. However, this finding is contrasted by the Picture Superiority Effect, which describes the tendency for people to have better memory for pictures over words (Stenberg, 2006). This effect is explained via Dual Coding Theory (Paivio & Clark, 2006) where visual information is encoded through visual and verbal systems while words may only be encoded verbally, giving visual information an advantage. Visuals are also processed faster and provide more context. I propose that video-based misinformation may be encoded more robustly than information presented in other

modalities, as its multimodal nature offers a more vivid and seemingly tangible representation of reality, potentially surpassing the Misinformation Effect of both static images and narratives.

1.2.3 How deepfakes could affect behaviours

Research on how misinformation affects behavioural outcomes has remained limited (Murphy et al., 2023). Only 1% of studies investigating misinformation effects between 2016 and 2022 directly measured behavioural outcomes and only 10% measured behavioural intentions, leading to a significant gap in empirical research exploring how misinformation exposure could lead to subsequent real-world behaviours (Murphy et al., 2023). Despite the lack of empirical studies, some theories can connect misinformation exposure to behavioural outcomes. Ajzen's (1991) Theory of Planned Behaviour (TPB) proposes that people's intentions to engage in behaviours are the main predictors of whether they conduct them. Behavioural intentions are influenced by three factors: attitudes towards the behaviour, subjective norms, and perceived behavioural control. Notably, exposure to misinformation can shape individual attitudes via mechanisms such as DPT, SIT, and motivated reasoning. In the context of deepfakes, the aforementioned Modality Effect and Picture Superiority Effect of audio-visual deepfakes could uniquely and persuasively influence attitudes and memories. Consequently, changes in attitudes resulting from deepfake exposure may lead to corresponding behavioural outcomes. Indeed, some empirical research has demonstrated this relationship such as Alwreikat (2021) surveying 210 students assessing their intentions to verify COVID-19 information before sharing using TPB and found that attitudes towards verification was a significant predictor of intentions to verify information before sharing. It is foreseeable that the potential processing and cognitive load advantage provided by deepfakes

could have downstream effects on attitudes and behavioural intentions ultimately leading to distinct behavioural changes compared to text and photo-based misinformation.

1.3 Epistemic Harms of Deepfakes

1.3.1 The Liar's Dividend

The public do not need to be directly exposed to deepfakes to be harmed by them. Conceptually, deepfake technology offers a hyper-realistic means of disseminating misinformation, enabling the creation of convincing depictions of any individual engaging in actions or making statements that a malicious actor might wish to attribute to them. This concept inherently casts doubt on all video-based information involving a person as it can be difficult for the average person to prove that a video is a deepfake. Politicians and public figures can take advantage of the idea of deepfake technology by falsely claiming that damaging but true information about themselves is a deepfake in what is known as the Liar's Dividend (Chesney & Citron, 2019). The Liar's Dividend presents several significant harms. First, it undermines democratic processes by allowing political figures to dismiss accurate information as deepfakes, thereby misleading the public and hindering their ability to make informed decisions about candidates (Schiff et al., 2025). Second, it contributes to the erosion of trust in authentic media by promoting the belief among supporters that realistic and truthful videos may be fabricated, even when they are genuine (Schiff et al., 2025). This erosion of trust may have downstream effects on trust in journalism and news sources as supporters who are deceived by the Liar's Dividend may perceive frontline news reporting unfavourably if they think they are reporting deepfakes as fact (Gregory, 2022). The Liar's Dividend presents itself as a weapon to be used by malicious actors in a political environment that is already saturated with misinformation and deflection, without any individual being exposed to a deepfake.

1.3.2 Epistemic Pollution

Malicious actors weaponizing the concept of deepfakes is not the only way to cause epistemic harm. As awareness of deepfake technology gradually becomes commonplace, researchers fear that people may start losing trust in authentic news sources and online video-based information because they won't be able to distinguish real videos from deepfakes (Fallis, 2020). Raising awareness about deepfakes is necessary to equip individuals with the skills and knowledge to identify and prevent deepfake manipulation, but well-intentioned deepfake interventions may be inadvertently causing epistemic harm. Sippy et al. (2024) have surveyed citizens from the UK about their views of deepfake technology and found that a significant concern was increasing distrust of information. The fear of being deceived by deepfakes can lead individuals to distrust even authentic videos, which is just as detrimental as being deceived by a deepfake itself. In both cases, the individual is deprived of reliable information, hindering their ability to accurately update their understanding of the world (Fallis, 2020).

These epistemic harms can be further explained through the lens of cognitive theory. Humans understand their broader environment by learning accurate information about it (Illeris, 2018). Due to limitations in cognitive capacity, it is not feasible to personally verify every piece of information we encounter. As a result, we tend to accept information as truthful (Levine, 2014) or rely on cognitive heuristics to guide our judgments. An example of such heuristics is the reputation heuristic, where information originating from sources perceived as credible is more likely to be considered accurate (Metzger & Flanagin, 2013). If people believe in a hypothetical future characterized by widespread deepfake content, an environment often anticipated due to alarmist discourse, then they are unable to passively accept information they receive as true and may only be able to rely on cognitive heuristics that may not always be accurate. The heightened cognitive load required to verify all video-based information may

exceed the individuals' cognitive capacity, leading to cognitive fatigue (Belabbes et al., 2023). Cognitive fatigue when processing information can be detrimental in two ways. People who are at their cognitive capacity are less trusting of information and limits their capacity for critical evaluation (Samson & Kostyszyn, 2015), making them potentially more susceptible to misinformation. Secondly, cognitive fatigue may lead to epistemic indifference where people no longer have the capacity to care about the veracity of information and indiscriminately distrust all information instead (Satta, 2024). Both outcomes due to the awareness of deepfake technology and fears about its prevalence cause significant epistemic harms to truthful information, yet little empirical research has been conducted to quantify and explore these effects. One of the primary goals of this thesis is to quantify and reduce the epistemic harms that originate from awareness about deepfakes through deepfake interventions.

1.4 Conceptual and Methodological Challenges in Deepfake Research

1.4.1 Conceptual Challenges

In conducting empirical research on deepfakes, I encountered several conceptual challenges, particularly regarding their definition. The academic discourse surrounding deepfakes is marked by considerable confusion, which complicates efforts to establish a shared understanding of what is being studied. In work conducted with Twomey et al. (2024), we observed that definitions of deepfakes in the literature are often inconsistent. Some studies provide no definition at all, while others misclassify traditional media manipulation techniques such as CGI or video editing as deepfakes, when they may more accurately be described as shallowfakes or cheapfakes. This inconsistency can lead to non-deepfake examples being used to illustrate deepfake harms, thereby conflating the concept and muddying the research

landscape. We further identified the importance of identity mimicry to both the definition of deepfakes and the harms they produce. In this thesis, I addressed the definitional challenges of deepfakes by adopting a definition aligned with the perspective shared in Twomey et al. (2024), while also incorporating a technical description of deepfakes to accurately capture their origins and contemporary harms.

Another conceptual issue I address in this in this thesis relates to the growing technical realism of deepfakes, making it difficult for individuals to reliably distinguish them from authentic videos (Kobis et al., 2021; Somoraj & Miller, 2023). While producing the most convincing deepfakes still requires specialized expertise (Twomey et al., 2024), current technology already enables the creation of highly effective deepfakes with relatively accessible tools (Murphy et al., 2025). As deepfakes approach perceptual indistinguishability from real footage, empirical deepfake research faces conceptual challenges, particularly when designing studies that rely on deepfake stimuli intended to simulate real-world conditions. Current empirical research attempts to measure cognitive effects of deepfake exposure by directly showing deepfakes to participants and measuring responses (Dobber et al., ,2021; Hameleers et al., 2022; Hameleers et al., 2024; Murphy et al., 2025). However, the purpose of experimental research becomes unclear when deepfakes are indistinguishable from real videos, as it complicates the interpretation of which psychological constructs or effects are truly being assessed. If participants are effectively watching real videos, it becomes difficult to distinguish between research on deepfake exposure and research solely measuring peoples' prior beliefs, biases, perceived credibility of the researchers showing the videos. This conceptual challenge potentially suggests that future research on deepfake effects should focus on pre-existing susceptibility factors to misinformation such as motivated reasoning, social identity, and heuristic biases. Contemporary deepfake research is still necessary however, due to perceptual and qualitative differences between real videos and deepfakes such as visual artifacts when the

faces are captured at different angles and blurred transitions between the original and deepfaked faces. It remains necessary to assess how effective deepfakes created with current capabilities are at influencing beliefs, memories, and behaviours in order to monitor their ongoing impact and development.

1.5 Methodological Challenges

One of the main methodological challenges in deepfake research is ensuring that it is conducted ethically. This challenge is not unique to deepfake research as it is commonly encountered in misinformation research more generally (Greene et al., 2023). However, the realistic audio-visual nature of deepfakes and lack of empirical evidence establishing its cognitive and behavioural impacts underscores the need to be vigilant when exposing participants to deepfake-based misinformation. In this context, two primary ethical challenges arise: the potential for participants to develop false beliefs and memories which may influence their behaviours, and the reputational damage done to the individual portrayed as a deepfake. Addressing these challenges is essential to responsibly advancing research in this area.

Before evaluating how deepfakes might disrupt cognitive or behavioural processes, it is essential to first assess the inherent harms posed by general misinformation itself. There is research to suggest that it is unlikely for one-off misinformation exposures to significantly and adversely impact participants' beliefs and memories (Greene & Murphy, 2021; Greene et al., 2025). Greene et al. (2025) experimentally tested whether misinformation about climate change and specific foods would affect participant's attitudes and intentions to consume the targeted food and sign an online petition. Their results show minimal effects of misinformation on attitudes or behavioural intentions. Greene et al. (2025) reasoned that a single exposure to generic misinformation was likely insufficient in producing attitudinal or behavioural effects. Even direct persuasion such as in political contexts has often been shown to be ineffective at

swaying beliefs and behaviours (Coppock et al., 2020; Kalla & Broockman, 2018). Lazarsfeld et al. (1968) who describes how our beliefs and behaviours are influenced by a multitude of factors such as identity, values, and demographics which are entrenched in our development from a young age and are therefore extremely resilient to change. Considering how difficult it is to influence beliefs and behaviours in traditional misinformation formats, it is difficult to predict how significant of an impact deepfakes have on attitudes and cognition due to the added audio-visual element.

In light of the uncertain effects that exposure to deepfakes may have on participants' beliefs and behaviours, one important safeguard for researchers is the implementation of a thorough and carefully designed debriefing procedure to minimise potential harm. Recent advances in ethical meta science have outlined an effective debriefing strategy to remove lingering negative effects of misinformation on beliefs and behaviours (Greene & Murphy, 2023; Murphy & Greene, 2023). Greene & Murphy (2023) describe specific steps to debrief participants including a detailed and explicit correction of the source of misinformation in the study, full disclosure of the purpose of the study and the mechanisms behind the misinformation strategy, asking participants for consent to use their data after they understand the experiment, and conducting a follow-up study to ensure that participants did not continue to be affected by the misinformation. Following these strategies, research has shown that participants can effectively be protected from misinformation in experimental studies (Greene et al., 2024; Greene & Murphy, 2023), thereby significantly reducing the risk of lasting epistemic harm and demonstrating that it is possible to minimize harm to participants in deepfake experiments in controlled research settings.

Participants are not the only group that can be harmed by deepfakes in deepfake experiments. Deepfakes designed to spread misinformation inherently involve misrepresenting the individual depicted, which could lead to reputational harm to the person portrayed. We

consider the risk of reputational harm and defamation of the individual due to deepfakes created or used in deepfake experiments to fall under the concept of "minimal risk" whereby the potential harm that an individual could experience from an experiment is not greater than harm that the individual would reasonably experience in their everyday life (Kopelman, 2004). Public figures such as politicians and celebrities, who can be depicted in deepfakes that damage their reputations, often face misinformation about them in everyday life through memes, fake news, and tabloid stories (Tandoc et al., 2018). Deepfakes created for use in experiments that depict such public figures can reasonably be understood as similar to this type of content, but unlike in everyday life, those exposed to this misinformation in an experimental study are quickly and effectively debriefed. Therefore, these deepfakes fit within the minimal risk framework, suggesting that their use in research does not cause unnecessary reputational harm to the depicted individuals. Ultimately, while deepfake research presents unique ethical challenges, careful methodological design and thorough debriefing procedures can effectively minimize risks, enabling valuable insights into deepfakes without compromising participant welfare or individual reputations

1.6 Research Questions

The research questions in subsequent chapters revolve around the central question: how do deepfakes affect beliefs, memories, and behaviours and how can the effects of misleading deepfakes be mitigated? This PhD project comprises of four studies which serve to answer different elements of my overall research question.

Study 1 is a scoping review of existing empirical studies that used deepfakes as stimuli to test how deepfake exposure influences beliefs, memories, and/or behaviours. No review had been conducted on empirical deepfake research when the search was first conducted, and there were only 10 studies that fit the inclusion criteria. It was considered appropriate to conduct a

scoping review to map out the strengths, weaknesses, and direction of current empirical deepfake research before conducting my own research in this area.

Study 2 involves the use of thematic analysis to analyse the comments of a popular YouTube music video that featured deepfakes to qualitatively understand people's lived experience of deepfakes. This was driven by the findings of Study 1, where the scoping review did not find any qualitative research on deepfake perceptions and revealed a gap in knowledge in how the everyday individual experiences and responds to deepfakes in a naturalistic context.

In Study 3, an experimental deepfake study was conducted to directly measure how deepfakes affects beliefs, memories, and behaviours. This study was motivated by the intention to substantiate or challenge the academic literature surrounding the political harms of deepfake technology by empirically testing deepfake effects in an ecologically valid political environment.

Finally, I considered it important to take the findings of my previous studies and identify optimal methods to educate the public about deepfakes. In Study 4, I moved to designing an educational deepfake intervention which educates the public about deepfakes based on empirical evidence from the previous 3 studies and intervention best-practices.

1.7 Thesis Outline

The overarching research question for this thesis is: how do deepfakes affect beliefs, memories, and behaviours, and how can the effects of misleading deepfakes be mitigated? By addressing this question, the thesis contributes to the field by providing empirical evidence on the real-world impacts of deepfakes and developing guidance for interventions to reduce misinformation-related harms. Each chapter was designed with a clear objective that collectively builds a coherent narrative: to understand deepfake effects, examine their real-

world experience, evaluate claims of unique harms, and develop effective public-facing interventions.

Chapter 2 presents a scoping review of empirical deepfake studies, with the aim of understanding the current evidence regarding deepfake effects. This review synthesises and critiques existing research on the cognitive and behavioural consequences of deepfakes, situates deepfakes among other misinformation strategies, and identifies methodological limitations. Findings revealed that while deepfakes can influence beliefs, memories, and behaviours, there is little evidence that they are uniquely harmful, challenging alarmist narratives. The chapter also highlighted gaps in the literature, including the absence of qualitative research and investigations into real-world experiences, providing methodological guidance for the subsequent studies in this thesis.

Chapter 3 demonstrates the psychological experience of encountering deepfakes in social media content through a thematic analysis of YouTube comments on a music video featuring deepfakes. This study examined how people respond to deepfakes outside the laboratory, demonstrating that audiences often perceive deepfakes positively, appreciating their creative and meaningful use. At the same time, limited awareness can lead to a truth-bias, and dystopian perceptions persist despite the lack of empirical evidence for uniquely harmful effects. By providing ecological context, Chapter 3 complements the controlled investigations in later chapters and highlights the importance of understanding deepfakes in situ. Specifically, this chapter gives insight into the everyday experience of viewing deepfakes and highlights the urgency for deepfake literacy, which eventually informed the design of the educational deepfake interventions in Chapter 5.

Chapter 4 builds on the scoping review and challenges claims regarding the unique harms of deepfakes using quantitative methods. Conducted in a politically relevant context during an imminent election, this study tested whether deepfakes uniquely influenced beliefs,

memories, and behaviours compared to other misinformation formats such as text and audio. Experiments also considered delayed effects, such as the sleeper effect and illusory truth. Findings indicate that while deepfakes can impact cognition and behaviour, they are not inherently more persuasive than other forms of misinformation, providing rigorous evidence that further challenges alarmist narratives.

Chapter 5 addresses the final objective of the thesis: creating and evaluating public-facing interventions to mitigate deepfake harms. Drawing on insights from previous chapters, I designed an evidence-based educational intervention aimed at reducing anxiety, improving discernment, and preserving epistemic trust in authentic videos. This intervention was compared to existing approaches, such as alarmist messaging or detection tips. Results show that evidence-based strategies can effectively educate the public about deepfakes while minimizing unintended consequences, translating the empirical findings of this thesis into actionable guidance.

Chapter 6 concludes the thesis by integrating findings across all four studies. It discusses theoretical and practical implications, considers limitations, and outlines directions for future research. Collectively, the chapters advance understanding of deepfake effects, demonstrate their psychological impact in real-world contexts, evaluate claims of unique harms, and offer empirically grounded strategies for mitigating misinformation-related consequences.

1.8 References

Abraham, T. M., Wen, T., Wu, T., & Chen, Y. W. (2025). Leveraging data analytics for detection and impact evaluation of fake news and deepfakes in social networks. *Humanities and Social Sciences Communications, 12*(1), 1-13.

Acerbi, A., Altay, S., & Mercier, H. (2022). Research note: Fighting misinformation or fighting for information?.

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). The State of Deepfakes: Landscape, Threats, and Impact. Amsterdam: Deeptrace

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Almond, G. (2019, January 14). Why Victorians feared modern technology would make everyone blind [Review of Why Victorians feared modern technology would make everyone blind]. World Economic Forum. https://www.weforum.org/stories/2019/01/why-victorians-feared-modern-technology-would-make-everyone-blind/

Belabbes, M. A., Ruthven, I., Moshfeghi, Y., & Rasmussen Pennington, D. (2023). Information overload: a concept analysis. *Journal of Documentation, 79*(1), 144-159.

Broinowski, A. (2022). Deepfake nightmares, synthetic dreams: A review of dystopian and utopian discourses around deepfakes, and why the collapse of reality may not be imminent—yet. *Journal of Asia-Pacific Pop Culture, 7*(1), 109-139.

Broockman, D. E., & Kalla, J. L. (2023). When and why are campaigns' persuasive effects small? Evidence from the 2020 US presidential election. *American Journal of Political Science, 67*(4), 833-849.

Caldwell, S. (2016). Framing digital image credibility: image manipulation problems, perceptions and solutions.

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff., 98*, 147.

Ching, D., Twomey, J., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2025). Can deepfakes manipulate us? Assessing the evidence via a critical scoping review. *PLoS One, 20(*5), e0320124.

Coppock, A., Hill, S. J., & Vavreck, L. (2020). The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Science advances, 6*(36), eabc4046.

Dan, V. (2025). Deepfakes as a Democratic Threat: Experimental Evidence Shows Noxious Effects That Are Reducible Through Journalistic Fact Checks. *The International Journal of Press/Politics*, 19401612251317766.

Duboust, O. (2023, July 21). *Deepfake videos prompt false memories of films, study shows*. Euronews; euronews.com. https://www.euronews.com/next/2023/07/21/deepfake-videos-prompt-false-memories-of-films-study-shows

Fallis, D. (2021). The epistemic threat of deepfakes. *Philosophy & Technology, 34*(4), 623-643.

Feeney, M. (2021, October 26). Deepfake Laws Risk Creating More Problems Than They Solve - Regulatory Transparency Project. Regulatory Transparency Project. https://rtp.fedsoc.org/paper/deepfake-laws-risk-creating-more-problems-than-they-solve/

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*(2), 359-366.

Greene, C. M., & Murphy, G. (2021). Quantifying the effects of fake news on behavior: Evidence from a study of COVID-19 misinformation. *Journal of experimental psychology: Applied, 27*(4), 773.

Greene, C. M., & Murphy, G. (2023). Debriefing works: Successful retraction of misinformation following a fake news study. *PloS one, 18*(1), e0280295.

Greene, C. M., de Saint Laurent, C., Murphy, G., Prike, T., Hegarty, K., & Ecker, U. K. (2023). Best practices for ethical conduct of misinformation research: A scoping review and critical commentary. *European Psychologist, 28*(3), 139.

Greene, C. M., Ryan, K. M., Ballantyne, L., Barrett, E., Cowman, C. S., Dawson, C. A., ... & Murphy, G. (2024). Unringing the bell: Successful debriefing following a rich false memory study. *Memory & Cognition, 52*(5), 1079-1092.

Greene, C. M., Brassil, M., Bryan, E., Glass, S. B., O'Connor, R., O'Keeffe, V., ... & Murphy, G. (2025). Evaluating real-world effects of one-off fake news exposure. *Scientific Reports, 15(*1), 1-19.

Gregory, S. (2022). Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism, 23*(3), 708-729.

Hameleers, M., van der Meer, T. G., & Dobber, T. (2022). You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media+ Society, 8*(3), 20563051221116346.

Hameleers, M., Van Der Meer, T. G., & Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior, 152*, 108096.

Horvitz, E. (2022, November). On the horizon: Interactive and compositional deepfakes. In *proceedings of the 2022 international conference on multimodal interaction* (pp. 653-661).

Illeris, K. (2018). A comprehensive understanding of human learning. In *Contemporary theories of learning* (pp. 1-14). Routledge.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Khasawneh, O. Y. (2023). Technophobia: how students' technophobia impacts their technology acceptance in an online class. *International Journal of Human–Computer Interaction, 39*(13), 2714-2723.

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *Iscience, 24*(11).

Kopelman, L. M. (2004). Minimal risk as an international ethical standard in research. *The Journal of medicine and philosophy, 29*(3), 351-378.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological bulletin, 108*(3), 480.

Lazarsfeld, P., Berelson, B. & Gaudet, H. (1968). *The People'S Choice: How the Voter Makes Up His Mind in a Presidential Campaign.* Columbia University Press. https://doi.org/10.7312/laza93930

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature, 521*(7553), 436-444.

Levine, T. R. (2014). Truth-default theory (TDT) a theory of human deception and deception detection. *Journal of Language and Social Psychology, 33*(4), 378-392.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*(5), 585–589. https://doi.org/10.1016/S0022-5371(74)80011-3

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & memory, 12*(4), 361-366.

Margry, K. (1992). ‘Theresienstadt’(1944–1945): The Nazi propaganda film depicting the concentration camp as paradise. *Historical Journal of Film, Radio and Television, 12*(2), 145-162.

McGlynn, C., & Toparlak, R. T. (2025). The ‘new voyeurism’: criminalizing the creation of ‘deepfake porn’. *Journal of Law and Society, 52*(2), 204-228.

Murphy, G., & Greene, C. M. (2023). Conducting ethical misinformation research: Deception, dialogue, and debriefing. *Current Opinion in Psychology, 54*, 101713.

Murphy, G., de Saint Laurent, C., Reynolds, M., Aftab, O., Hegarty, K., Sun, Y., & Greene, C. M. (2023). What do we study when we study misinformation? A scoping review of experimental research (2016-2022). *Harvard Kennedy School Misinformation Review*.

Murphy, G., Ching, D., Meehan, E., Twomey, J., Bolger, A., & Linehan, C. (2025). An Average Joe, a Laptop, and a Dream: Assessing the Potency of Homemade Political Deepfakes. *Applied Cognitive Psychology, 39*(2), e70061.

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR), 54*(1), 1-41.

Nash, R. A., Wade, K. A., & Lindsay, D. S. (2009). Digitally manipulating memory: Effects of doctored videos and imagination in distorting beliefs and memories. *Memory & Cognition, 37*(4), 414-424.

Ngak, C. (2012, November 9). Tupac Coachella hologram: Behind the technology. Www.cbsnews.com. https://www.cbsnews.com/news/tupac-coachella-hologram-behind-the-technology/

Paivio, A., & Clark, J. M. (2006). Dual coding theory and education. *Pathways to literacy achievement for high poverty children, 1*, 149-210.

Poidevin, O. L. (2025, July 11). UN report urges stronger measures to detect AI-driven deepfakes. Reuters. https://www.reuters.com/business/un-report-urges-stronger-measures-detect-ai-driven-deepfakes-2025-07-11/

Romanishyn, A., Malytska, O., & Goncharuk, V. (2025). AI-driven disinformation: policy recommendations for democratic resilience. *Frontiers in Artificial Intelligence, 8*, 1569115.

Rousay, V. (2023). Sexual deepfakes and image-based sexual abuse: Victim-survivor experiences and embodied harms (Master's thesis, Harvard University).

Sacchi, D. L., Agnoli, F., & Loftus, E. F. (2007). Changing history: Doctored photographs affect memory for past public events. Applied Cognitive Psychology: The Official *Journal of the Society for Applied Research in Memory and Cognition, 21*(8), 1005-1022.

Satta, M. (2024). Epistemic exhaustion and the retention of power. Hypatia, 39(3), 510-529.

Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The liar's dividend: can politicians claim misinformation to evade accountability?. *American Political Science Review, 119*(1), 71-90.

Security Hero. (2023). 2023 State Of Deepfakes: Realities, Threats, And Impact. Www.securityhero.io. https://www.securityhero.io/state-of-deepfakes/

Sippy, T., Enock, F., Bright, J., & Margetts, H. Z. (2024). Behind the Deepfake: 8% Create; 90% Concerned. Surveying public exposure to and perceptions of deepfakes in the UK. *arXiv preprint arXiv:2407.05529*.

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior, 149*, 107917.

Stenberg, G. (2006). Conceptual and perceptual factors in the picture superiority effect. *European Journal of Cognitive Psychology, 18*(6), 813-847.

Sunvy, A. S., Reza, R. B., & Al Imran, A. (2023). Media coverage of DeepFake disinformation: An analysis of three South-Asian countries. *Informasi, 53*(2), 295-308.

Tabbers, H. K., Martens, R. L., & Van Merriënboer, J. J. (2004). Multimedia instructions and cognitive load theory: Effects of modality and cueing. *British journal of educational psychology, 74*(1), 71-81.

Tajfel, H., Turner, J., Austin, W. G., & Worchel, S. (2001). An integrative theory of intergroup conflict. *Intergroup relations: Essential readings*, 94-109.

TBIJ. Deepfakes and their threat to global democracy. (2024). TBIJ. https://www.thebureauinvestigates.com/stories/2024-06-05/deepfakes-and-the-threat-to-global-democracy

Thompson, C. (2016, February 17). Texting Isn't the First New Technology Thought to Impair Social Skills. Smithsonian Magazine. https://www.smithsonianmag.com/innovation/texting-isnt-first-new-technology-thought-impair-social-skills-180958091/

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2024). What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology. *IEEE Transactions on Technology and Society.*

Twomey, J., Foley, S., Robinson, S., Quayle, M., Aylett, M. P., Linehan, C., & Murphy, G. (2025). " What do you expect? You're part of the internet": Analyzing Celebrities' Experiences as Usees of Deepfake Technology. *arXiv preprint arXiv:2507.13065.*

Vizoso, Á., Vaz-Álvarez, M., & López-García, X. (2021). Fighting deepfakes: Media and internet giants' converging and diverging strategies against hi-tech misinformation. *Media and Communication, 9*(1), 291-300.

Wade, K. A., Garry, M., Don Read, J., & Lindsay, D. S. (2002). A picture is worth a thousand lies: Using false photographs to create false childhood memories. *Psychonomic bulletin & review, 9*(3), 597-603.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review, 9*(11).

Yadlin-Segal, A., & Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence, 27*(1), 36-51.

Chapter 2 Can deepfakes manipulate us? Assessing the evidence via a critical scoping review

This paper was published in *PLOS One.*

Ching, D., Twomey, J., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2025). Can deepfakes manipulate us? Assessing the evidence via a critical scoping review. *PLoS One, 20*(5), e0320124.

Study Motivation

This scoping review collates and analyses studies that have empirically tested the effects of deepfake exposure on beliefs, memories, and behaviours. It was necessary to conduct this scoping review to gain a better understanding of the current direction, quality, and strength of empirical deepfake research.

Author Contribution

Didier Ching, Conor Linehan, and Gillian Murphy conceptualized the study. Didier Ching conducted the searches and screened papers, and John Twomey assisted in the screening of the papers as the secondary reviewer. Didier Ching wrote the original draft and edited the document based on feedback from John Twomey, Matthew Aylett, Michael Quayle, Conor Linehan, and Gillian Murphy. The final version of the manuscript was approved by every author.

2.1 Abstract

Deepfakes are one of the most recent developments in misinformation technology and are capable of superimposing one person's face onto another in video format. The potential of this technology to defame and cause harm is clear. However, despite the grave concerns expressed about deepfakes, these concerns are rarely accompanied with empirical evidence. We present a scoping review of the existing empirical studies that aim to investigate the effects of viewing deepfakes on people's beliefs, memories, and behaviour. Five databases were searched, producing an initial sample of 2004 papers, from which 22 relevant papers were identified, varying in methodology and research methods used. Overall, we found that the early studies on this topic have often produced inconclusive findings regarding the existence of uniquely persuasive or convincing effects of deepfake exposure. Moreover, many experiments demonstrated poor methodology and did not include a non-deepfake comparator (e.g. text-based misinformation). We conclude that concerns about the potential dystopian uses of deepfake technologies has outpaced the available experimental research assessing these harms. We close by offering insights on how to conduct improved empirical work in this area.

Introduction

Misinformation can be defined as information that is inaccurate or untrue, and which is unintentionally disseminated (Wu et al., 2019), whereas intentionally disseminated inaccurate information is defined as disinformation (Fallis, 2015). We will be using the term "misinformation" as shorthand for inaccurate information disseminated regardless of intent in this paper. Research into misinformation in the form of fake news articles, misleading health articles, and doctored photos shows that attitudes, behaviours, and memories can be

manipulated through these media (Greene & Murphy, 2021; Murphy & Flynn, 2022; Pennycook & Rand, 2021; Pickrell et al., 2016). “Deepfakes” are one type of media that has been used to spread misinformation and has raised serious concerns among the public and academic community (Albahar & Almalki, 2019).

Deepfakes are synthetic media generated using AI neural network technology (Chawla, 2019). Commonly, deepfakes are presented as manipulated videos showing one person’s face superimposed onto another person effectively creating a video of a person doing something they did not do. Deepfakes can potentially have exciting and positive applications, such as translating educational materials or films (Coupe, 2020; Dunne, 2018), in films and music videos for entertainment purposes (Lamar, 2022; Qobrtay, 2022), or even in museums to resurrect historical figures (Mihailova, 2021). However, concern about the negative applications of this technology such as defamation, revenge pornography, and political sabotage (Albahar & Almalki, 2019) often overshadows other applications.

Many authors have speculated about the potential threats of deepfakes and the behavioural, political, and legal implications of this technology (Fallis, 2021; Hancock & Bailenson, 2021; Pantserev, 2020; Wilkerson, 2021). For example, Chesney and Citron (2019) have suggested that deepfakes could destabilise society by instilling mistrust in citizens towards all media and influence voting behaviours thus jeopardizing election integrity. Much of the academic literature on this topic is conceptual or essay-based rather than empirical, and there remains relatively few studies that directly test the harms often discussed in relation to deepfakes (Ajder et al., 2019; Albahar & Almalki, 2019; Fallis, 2021; Hancock & Bailenson, 2021). A review of existing evidence is necessary to consolidate what we currently know about the effects of deepfake exposure in a rapidly developing and largely concept-driven field. In this paper, we present a scoping review mapping the current state of deepfake literature regarding the effects of viewing deepfakes on people’s beliefs, memories, and behaviour. These

outcomes were chosen to reflect elements of human psychology which are vulnerable to deepfake manipulation according to non-empirical papers and discussions (Albahar & Almalki, 2019; Chesney & Citron, 2019; Coupe, 2020). We will also be analysing the types of deepfakes used in empirical studies to identify strengths and weaknesses of their study design and methodology. This is to generate recommendations for future empirical study design.

While traditionally, scoping reviews do not aim to critically appraise the included studies, we will be doing so for several reasons. Firstly, since preliminary searches indicate a small amount of relevant empirical deepfake literature, critically appraising the available literature will yield a more in-depth and accurate analysis of the direction of the literature. Our research questions are concerned with whether evidence exists to suggest that deepfake exposure has a measurable effect on an individual's beliefs, memories, and/or behaviours. Appraising the evidence of existing research studies is crucial because it helps us accurately identify high quality research which contribute as evidence to answering our research questions. Surface-level analysis of study findings in our sample will include papers which could possibly give us a false representation of the strength of evidence in particular research areas.

This scoping review aims to answer four questions: RQ1) Is there evidence to suggest that deepfake exposure has measurable impacts on beliefs, memories, or behaviours of viewers? RQ2) Is there evidence to suggest that the effects of viewing deepfakes on beliefs, memories, and behaviours are different from those seen with misinformation spread via other media? RQ3) Is there evidence of social, psychological, and technical factors that influence the effects deepfakes have on beliefs, memory, and behaviour? RQ4) What kind of deepfake technology is being used in modern empirical deepfake research?

2.2 Literature Review

2.2.1 The Threat of Deepfakes

There is a substantial literature discussing the possible societal harms associated with deepfakes. For example, Albahar and Almalki (2019) suggest that deepfakes could make fake news and misinformation more widely present in all forms of media, leading to citizens being unable to differentiate between real and misleading information. Indeed, some research has shown that individuals cannot discern real videos from deepfakes at a rate better than chance (Kobis et al., 2021). Just as worryingly, Kobis et al. (2021) also showed that individuals overestimate their ability to discern deepfakes from real videos, possibly further increasing their susceptibility to deepfake deception. In a legal context, the validity of images and videos for use as evidence in courtrooms may be in jeopardy because video evidence, which is often considered the gold standard for proof, can no longer be trusted. In a political context, deepfakes could be used as a form of sabotage and defame candidates from opposing parties as you could create a deepfake depicting the opposition candidate saying something derogatory. These concerns are repeatedly highlighted in published literature (Ajder et al., 2019; Fallis, 2021; Westerlund, 2019; Wilkerson, 2021). Chesney and Citron (2019) further discuss that deepfake technology could distort democratic discourse, manipulate elections, and undermine public safety. Concerns around deepfakes have also spread to military and national security risks where, according to Chesney and Citron (2019), deepfakes "have utility as a form of disinformation supporting strategic, operational, or even tactical deception" (p. 1783).

Deepfakes may also cause harm in other ways. Firstly, some authors argue that deepfakes could give rise to a "Liar's Dividend", where factual information from media and news in society is invalidated and incorrectly labelled as fake news (Chesney & Citron, 2019). As an example, a politician could deny the veracity of an incriminating video of themselves by declaring it as a deepfake. This could serve to undermine citizen's trust in any information they receive regardless of authenticity. Similarly, deepfakes may cause "epistemic pollution" where knowledge of the existence of deepfakes may cause individuals to accuse real videos of being

deepfakes (Fallis, 2021). For example, Twomey et al. (2023) explored discourses surrounding the Russo-Ukrainian war on Twitter. Surprisingly, the most frequent form of deepfake-related misinformation found in the dataset was real media being labelled as being deepfake. Secondly, deepfakes may cause reputational harm, even in situations where the videos have been identified as inauthentic, such as with the generation of deepfake revenge pornography (Karasavva & Noorbhai, 2021). The reputational damage comes from being depicted in realistic-looking sexually suggestive positions, meant to embarrass and humiliate the victims (Henry et al., 2017; Karasavva & Noorbhai, 2021). Importantly, the warnings seen so frequently in literature about deepfakes are based on the assumption that deepfakes can successfully influence the beliefs, memories, and behaviours of viewers. However, there is no existing review that assembles and synthesises the available evidence, to determine whether that view is supported empirically.

2.2.2 Misinformation Effects on Beliefs, Memories, and Behaviour

Decades of research have demonstrated that encountering misinformation can affect people's beliefs, with more recent research focused on fake news stories. Exposure to fake news is correlated with a range of adverse outcomes. For example, Ognyanova et al. (2020) demonstrated that false information is linked to reduced trust in mainstream media, while Balmas (2014) showed that subjection to political misinformation is associated with attitudes of inefficacy, alienation, and cynicism towards political candidates. Misinformation in a health context can also influence people's beliefs. For example, studies have shown that misinformation about COVID-19 during the pandemic has been associated with increased vaccine hesitancy (de Saint Laurent et al., 2022; Mahmud et al., 2023; Pierri et al., 2022). Beyond text-based fake news stories, research has shown that doctored photographs and other forms of manipulated media can affect beliefs as well. For example, Hameleers et al. (2020)

demonstrated that a combination of doctored photos and misinformation texts was more credible than text alone, suggesting that visual media could enhance the potency of misinformation. In agreement with this finding, Nash et al., (2009a) discovered a small but consistent effect of doctored photos on people's beliefs about prior events, as well as the potential for doctored videos to distort participants' beliefs about actions they had performed (Nash et al., 2009a; Nash et al., 2009b). Messaris & Abraham (2001) suggests that visual information provides an evidential quality that enhances its credibility over text. The multi-modal audio-visual aspect of deepfakes may suggest an enhanced potency at generating and propagating misinformation. Current literature has also supported this sentiment, with researchers speculating that the visual aspect of deepfakes makes them more likely to affect our beliefs (Liv & Greenbaum, 2020; Vaccari & Chadwick, 2020).

A large body of evidence has also demonstrated clear effects of misinformation on memory. Elizabeth Loftus first established "the misinformation effect" where post-event misleading information can alter one's memories of an event (Loftus, 2005). The misinformation effect can be seen in political settings where false memories were induced by fake news stories for the abortion referendum in Ireland (Murphy et al., 2019) and Brexit (Greene et al., 2021). Similarly, there is also evidence to support the distortion of memory due to exposure to doctored photos. Wade et al. (2002) induced false memories of a childhood hot air balloon ride by showing participants doctored photos. This effect was replicated by Sacchi et al. (2007) when they found that doctored photos caused participants to misremember elements of important historical events such as the 1989 Tiananmen Square protest in Beijing and the 2003 Rome anti-war protest. Many researchers have proposed that deepfakes will similarly induce vivid false memories (Liv & Greenbaum, 2020; Rini & Cohen,, 2022; Vaccari & Chadwick, 2020). However, Greene et al. (2022) and Garry and Wade (2005) found that visual-based misinformation in the form of photographs did not induce false memories better

than simple text narratives. They reasoned that the narratives encouraged deeper processing of misinformation, leading to more significant distortions in memory compared to photographs. Many previous studies, including Wade et al. (2002) and Saachi (2007) did not include text-based misinformation as a comparator, so though they reported substantial rates of false memories, these may not have been due to the doctored images. Thus, it is difficult to predict whether deepfakes might be more potent at manipulating memory than non-technical means of conveying misinformation.

There are significantly fewer studies on the effects of misinformation on behaviours. Research on the COVID-19 pandemic has shown that misinformation exposure from social media has induced a decline in intent to be vaccinated (Loomba et al., 2021). Similarly, a decline in childhood vaccination rates against measles, mumps, and rubella (MMR) was found to be related to misinformation about the MMR vaccine (Leask et al., 2010). Greene and Murphy (2021) experimentally quantified the effects of misinformation exposure on behavioural outcomes and found a small but significant effect of COVID-19 misinformation on behavioural intentions. Conversely, de Saint Laurent et al. (2022) found no effect of COVID-19 misinformation on behavioural intentions. The existing literature focuses on health contexts and misinformation delivered through text only with little agreement, thus it is unclear how misinformation affects behavioural intentions in different contexts and mediums. However, general claims have been made in the literature about the ability of misinformation to modify and manipulate behaviour, threaten individual autonomy, and jeopardize democracy (Bastick, 2021).

2.2.3 Factors Influencing the Effects of Misinformation on People's Beliefs, Memories, and Behaviour

There is a wealth of research investigating the individual factors that lead people to be influenced by inaccurate or false information. For example, people who engage in analytical reasoning and reflection are less likely to be deceived by misinformation (Bryanov & Vziatysheva, 2021; Van Bavel et al., 2021). Pennycook and Rand (2021) and Bronstein et al. (2019) also mirror these findings, showing the negative correlation between cognitive reasoning and misinformation susceptibility. This shows that the tendency to engage in careful analysis rather than the intuitive acceptance of information is important in discerning truth from misinformation. This idea has then been used to design interventions to help combat misinformation. For example, misinformation "inoculation" or "pre-bunking" is a strategy which forewarns individuals of misinformation by providing accurate information on the misinformation content they are about to view. This strategy works by provoking individuals to critically analyse the content they are about to view, bypassing simple intuitive acceptance of information in favour of a purposeful engagement with the information (Van der Linden, 2022). Despite the wealth of literature investigating these individual factors and their implications, very few of these studies have specifically investigated deepfakes. Due to the multi-modal format of deepfakes, it is not known whether deepfakes are more, less, or equally effective in delivering misinformation and manipulating people's beliefs, memories, and behaviours compared to misinformation presented as text or photos, despite many claiming this to be the case (Liv & Greenbaum, 2020; Vaccari & Chadwick, 2020). Also, it is evident that factors such as cognitive style can influence the effects of misinformation on beliefs, memory, and behaviour and it would be reasonable to assume that deepfake misinformation would follow closely with this literature. The current scoping review will provide a clearer picture regarding the impact of deepfakes. be the first step to establishing the effects of deepfakes in misinformation literature.

2.3 Materials and Methods

2.3.1 Eligibility Criteria

In order to be included in the review, research papers were required to either test the effect of viewing deepfakes on beliefs, memories, or behaviour, compare the effects of deepfakes with other pre-existing forms of misinformation, or identify factors influencing the effects of viewing deepfakes. Papers were required to be published in peer-reviewed journals and written in English. The period of literature included spanned January 2017 to April 2024 (when we conducted our search). Years 2017–2024 were chosen as the first recorded incident of deepfakes occurred in 2017 (Cole, 2017). Any form of empirical study meeting these criteria were included, such as quantitative, qualitative, or mixed methods approaches.

2.3.2 Information Sources

5 electronic databases were searched on April 13th, 2024: PsycINFO (through the platform of American Psychological Association), Web of Science Core Collection, ACM, DBLP, and Scopus.

2.3.3 Search Strategy

The five electronic databases were searched using the terms: [deepfake] AND [Belief OR Memory OR Behavior OR Attitude]. The keywords were purposefully broad with no limitations to keep a wide scope due to the expected low number of deepfake psychological studies. Other keywords were considered such as “trust” or “believe” but these terms were considered to be already covered within the parameters of our search strategy. At peer-review stage, we re-conducted the search with these terms included, but did not find any further relevant papers. The search terms “belief”, “memory” “behavior” and “attitude” were

considered sufficiently broad to cover relevant forms of manipulation. As well as this, due to the possible broad interpretation of deepfakes in research papers (Whittaker et al., 2023), we considered including other terms to describe AI-generated videos or deepfake software, such as “AI generated videos” or “Fake Videos”, but practically, it was difficult to make this list exhaustive without including largely irrelevant general search terms (such as “ AI generated”) which would yield a redundant amount of papers. The references of relevant papers were also hand-searched for studies that fit the eligibility criteria.

2.3.4 Screening Process

The research papers were screened by two reviewers, using the Rayyan software package (Ouzzani et al., 2016). Both reviewers independently screened papers based on paper titles and abstracts first. After individual screening, reviewers compared screened papers, and discrepancies were resolved by discussion. Subsequently, a full-text screening process was carried out on the remaining papers screening out papers that 1) did not directly measure the effects of viewing deepfakes on beliefs, memories, or behaviours, 2) compare the effects of deepfakes to other forms of misinformation, or 3) investigate potential psychological, social, or technical factors which influences the effect of deepfakes of an individual’s beliefs, memories, and/or behaviours. Inter-rater agreement was approximately 80%, and disagreements were resolved by discussion. Following the title and abstract screening and the full-text screening, a total of 22 papers met the inclusion criteria and were included in the review. The selection and screening process details is illustrated in the PRISMA flow diagram in Fig 1.

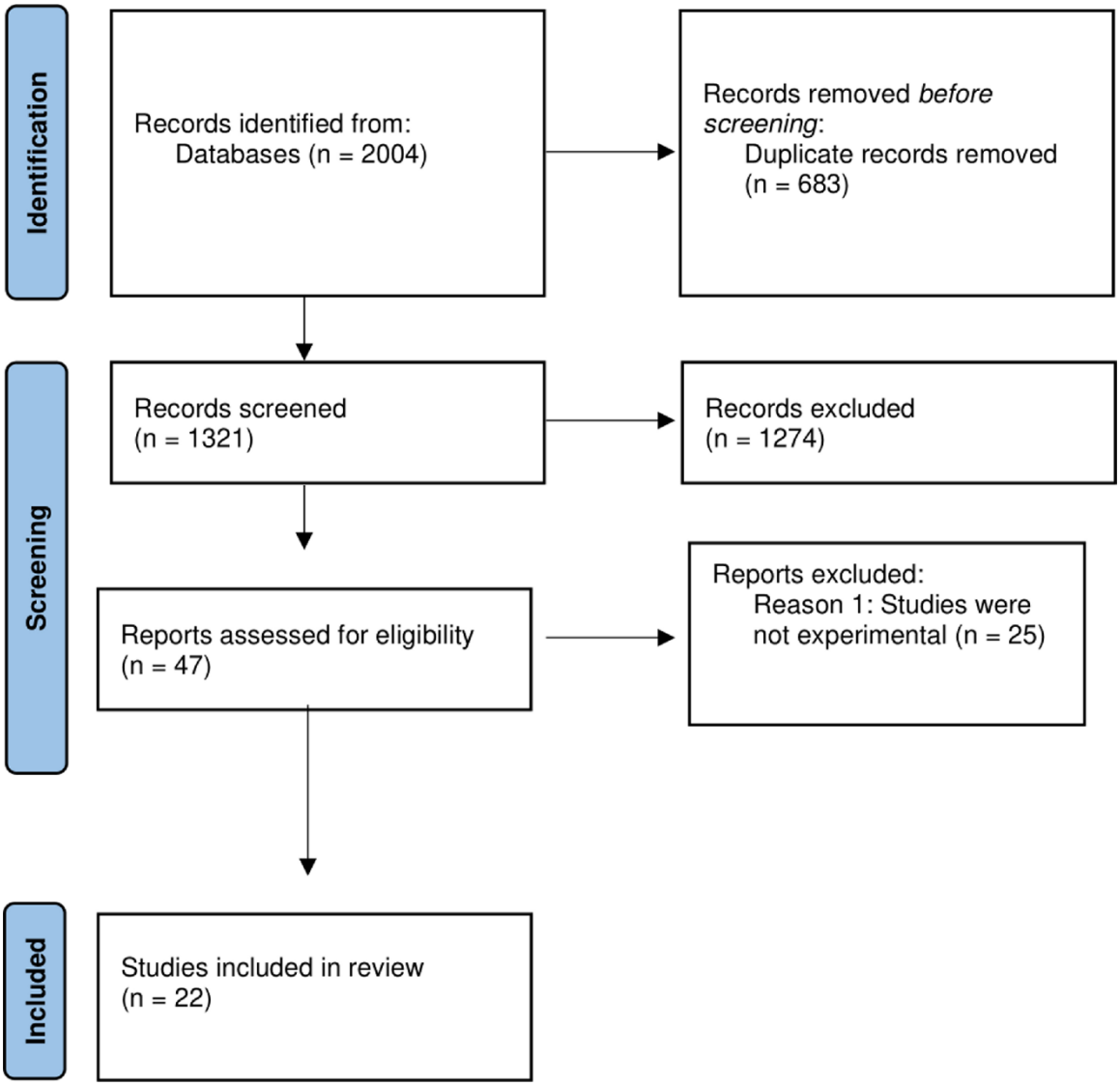


Figure 2-1: PRISMA Flow Diagram

2.3.5 Extraction and Analysis

Following guidance from Arksey and O'Malley's (2005) framework for conducting scoping reviews, data from the included articles for the review are charted below in Table 1. This data includes each paper's research question and main findings, as well as extracted data such as study design, source of deepfake, comparisons, and outcome measures. A narrative synthesis was carried out to collate all the data together to answer the four research questions.

2.4 Results

Out of the 22 included papers, 14 studies answered RQ1, six studies answered RQ2, 13 studies answered RQ3, and all 22 studies contributed to answering RQ4. The total studies added up to more than 22 papers because some papers had properties which helped answer multiple research questions. (see Table 2.1).

Table 2.1. Descriptive data of all included studies

Author	Descriptives						
	Year	Study Design	Source of deepfake	Comparisons	Methodology	Outcome Measures	Delay between exposure and outcome
Ahmed	2021a	Survey	N/A	None	Survey	Inadvertent deepfake sharing, political interest, cognitive ability, social network size, political trust, deepfake exposure and deepfake concern	None
Ahmed	2023	Survey	N/A	None	Survey	Social media news use, deepfake concern, deepfake exposure, inadvertent deepfake sharing, and cognitive ability	None
Ahmed	2021b	Experimental	Canny AI art installation	None	Participants were exposed to a deepfake with an educational caption or no captions and asked to answer questions	Perceived claim accuracy, sharing intention, and cognitive ability	None
Ahmed, Wang, & Bee	2023b	Experimental	Canny AI art installation	None	Participants asked to rate news article as real or fake twice. Experimental group sees a deepfake in-between ratings.	Perceived accuracy of stimuli and cognitive ability	None
Ahmed, Ting Ng, & Bee	2023a	Experimental Survey	Canny AI art installation	None	Survey	Perceived accuracy, Self-regulation, Fear of missing out, cognitive ability, and deepfake sharing	None

Ahmed & Chua	2023	Experimental Survey	RepresentUS	Deepfakes vs cheapfakes vs audio	Survey. Participants view misinformation in different modalities (deepfakes vs cheapfakes vs audio) and asked to answer questions.	Perceived accuracy, sharing intentions, and cognitive ability	None
							None
Dobber, Metoui, Trilling, Helberger, & de Vreese	2020	Experimental	Author-made deepfake using Tacotron 2 to synthesise voice and AI based lip synchronizing Technique	None	Participants exposed to political deepfake in experimental condition and asked to answer questions.	Attitudes towards politician, attitude towards politician's party, and ethics	None
Hameleers, van der Meer, & Dobber	2022	Experimental	VFX and Deepfake artist	Text-only to deepfake	Participants exposed to political deepfake in experimental condition and asked to answer questions	Issue agreement, credibility, and evaluations of the depicted politician	None
Hameleers, van der Meer, & Dobber	2024	Experimental	VFX and Deepfake artist	None	Participants exposed to political deepfake/text misinformation in experimental condition and asked to answer questions	Perceived credibility, delegitimizing effect on evaluations of political actor	None

Hameleers	2024	Experimental	Collaboration with computer vision and AI-Experts. Professional voice actor for speech	Deepfakes vs cheapfakes vs text	Participants exposed to political deepfake/cheapfake in experimental condition and asked to answer questions	Perceived credibility	None
Murphy, Ching, Twomey, & Linehan	2023	Experimental	YouTube	Deepfake vs text	Participants exposed to deepfakes in films and asked if they remembered seeing the film.	Presence of false memories	None
Lu & Chu	2023	Experimental	"Preface" mobile app	None	Participants exposed to deepfakes of people they were told to be deceased and asked to answer questions.	Perceived realism, identification, compassion, perceived desecration of the dead, surprise, policy support, and activism intentions	None
Sharma, Jain, Behl, Baabdullah, Giannakis, Dwivedi	2023	Experimental Survey	Custom created by deep learning technology professional	None	Participants exposed to political deepfakes and asked to answer questions as part of a survey.	Intentions to share, intention to verify political deepfake video, and moral consciousness	None
Yu-Leung Ng	2023	Experimental Survey	Developed by French charity Solidarite	None	Participants exposed to political deepfakes/Real political videos and asked to answer	Perceived video fakeness, perceived message fakeness, perceived dangerousness, and perceived trustworthiness	None

			Sida and computer scientists		questions as part of a survey. Before each video, there was either a description explaining what deepfakes are or no description.		
Jin, Zhang, Gao, Gao, Zhou, Yu, & Wang	2023	Experimental Survey	YouTube	None	Participants were exposed to deepfake videos and asked to answer questions as part of a survey. The videos were made to emulate real videos on platforms which included the number of followers, popularity, description as well as definition, duration, and editing.	Credibility, realism, source credibility, and popularity	None
Hwang, Ryu, & Jeong	2021	Experimental	Canny AI art installation	Text-only condition to deepfake + text	Participants were exposed to one of three interventions: deepfake specific literacy education, general disinformation education, or no literacy education. Participants were then exposed to deepfake videos and	Vividness, persuasiveness, credibility, and intent to share	None

					asked to answer questions afterwards.		
Iacobucci, De Cicco, Michetti, Palumbo, & Pagliaro	2021	Experimental Survey	YouTube	None	Participants were assigned to either read an excerpt on deepfakes or read nothing at all before viewing a deepfake video. Participants were then asked to answer questions afterwards.	Bullshit receptivity, attitude towards video, and intention to share video	None
Lee & Shin	2022a	Experimental	Faceswap Software	Text-only condition to text-photo to deepfakes	Participants were exposed to text, text-photo, or deepfake misinformation and asked to answer questions. Participants in a second experiment either saw a fake news correction or no correction before deepfakes and asked to answer questions afterwards.	Source vividness, fake news credibility, and fake news engagement intentions	None
Lee & Shin	2022b	Experimental	Faceswap Software	None	Participants were exposed to a deepfake and asked to answer questions afterwards.	News Credibility, Viral Behavioral Intention, Pre-existing Issue Attitude, Cost Perception, and Prior Source Familiarity	None
Murphy & Flynn	2021	Experimental	YouTube, Faceswap	Text-only condition to	Participants were exposed to	Self-reported false memory and false memory vividness	None

			software, and Canny AI art installation	text-photo to deepfakes	misinformation presented as text, text-photo, or deepfake and asked whether they remembered the contents of the misinformation happening.		
Vaccari & Chadwick	2020	Experimental	YouTube	None	Participants were exposed to a deepfake video and asked to answer questions afterwards.	Participant's belief in deepfake videos and trust in news and social media	None
Wu, Ma, & Zhang	2021	Experimental	ZAO deepfake app	None	Participants were exposed to a deepfake video and asked to answer questions afterwards.	Body image, state appearance self-esteem, state appearance comparison, and attractive possible self (APS) perception	None

2.4.1 RQ1. Is there evidence to suggest that deepfake exposure has a measurable impact on people's beliefs, memories, or behaviour?

Our review yielded eight studies that examined how deepfake exposure does have a measurable effect on people's beliefs across four research topics. Firstly, experiments by Hwang et al. (2021) and Lee and Shin (2022) found that participants rated misinformation messages presented as a deepfake as more vivid and credible compared to text and text-photo formats, suggesting deepfakes may be a uniquely persuasive form of misinformation. Secondly, Hameleers et al. (2024) and Dobber et al. (2021) document the effects of deepfakes on beliefs in a political context by exposing participants to deepfakes of political figures saying disrespectful (Dobber et al., 2021) and politically incongruent (Hameleers et al., 2024) statements. Their results show that deepfakes negatively impact people's attitudes towards the depicted politicians, even if the deepfake was not rated as credible (Hameleers et al., 2024). Thirdly, Vaccari and Chadwick (2020) and Ahmed et al. (2023) show that deepfakes can also be harmful by influencing our beliefs about authentic information, promoting mistrust in news and social media (Vaccari & Chadwick, 2020), as well as affecting our perceptions of real information about public figures by retroactively showing participants deepfakes of those public figures (Ahmed et al., 2023). Finally, studies by Wu et al. (2021) and Lu and Chu (2023) investigated how deepfakes influence our beliefs in a potentially positive way. Wu et al. (2021) demonstrated how viewing deepfakes of oneself on the body of a celebrity positively impacts young women's feelings of satisfaction with their own attractiveness and self-image while Lu and Chu (2023) explores the impact of so-called "digital resurrections" of deceased individuals and how they could be used advocate for issues related to their cause of death. However, they found that individuals considered these resurrections as disrespectful and demonstrated

reduced support for these issues. While there is evidence that deepfake exposure has a measurable effect on beliefs, the varied research questions and relatively low number of studies in each research topic makes it difficult to identify obvious trends in the literature.

Two studies in the review investigated how deepfakes affect memory and potentially create false memories. Murphy and Flynn (2022) found that deepfakes can indeed affect people's memories of events by creating false memories for events that did not occur, but this effect does not seem to be significantly different than other formats of misinformation such as text-only or text with a photo. Similarly, Murphy et al. (2023) found that deepfakes were capable of inducing false memories in participants in an entertainment context when this technology is used to replace the faces of actors in films with other famous actors. Interestingly, deepfakes were no more effective at inducing false memories than simple textual misinformation. This finding is tentative and based on just two studies, but it suggests that while deepfakes can manipulate memories, concerns about our memories being *uniquely* affected by the rise of deepfake misinformation (Liv & Greenbaum, 2020) may exceed the available empirical evidence.

Finally, we identified four studies that indirectly explored how attitudes and behavioural intentions were impacted by deepfakes. Lee and Shin (2022) and Hwang et al., (2021) found that people perceived deepfaked messages as more credible and vivid compared to other formats of misinformation, such as text-only and text-photos. They also found that this increase in vividness and credibility was associated with elevated levels of engagement and sharing intentions. Similarly, Ahmed (2021) found that individuals are more likely to share deepfakes online when they deem the deepfake to be credible, and Ahmed and Chua (2023) found that deepfakes, compared to cheapfakes or audio deepfakes, are especially likely to be shared. Noticeably absent from these few studies of deepfakes on behaviour are studies on a wider variety of behaviours beyond deepfake sharing (e.g. vaccine uptake, voting choice, etc.)

and any studies assessing real-life behaviours or behaviours over a longer time period. This echoes gaps in misinformation research more generally and is not unique to deepfake research (de Saint Laurent et al., 2022).

Evidently, the current empirical literature examining the effects of deepfakes on beliefs, memory, and behaviour is insufficient in number to draw strong conclusions. While the initial studies suggest a measurable effect of deepfakes on beliefs, memories, and behaviours, the relatively low number of studies, the varied research findings, inconsistent deepfake stimuli, and lack of behavioural research highlight many unanswered questions.

2.4.2 RQ2. Is there evidence to suggest that the effects of deepfakes on beliefs, memories, and behaviour are different from pre-existing forms of misinformation?

As shown in Table 1, we found just six studies that compared the deceptive potential of deepfakes to other misinformation formats such as text or text-photos (Hameleers et al., 2022; Hameleers, 2024; Hwang et al., 2021; Lee & Shin, 2022; Murphy & Flynn, 2022; Murphy et al., 2023). These six studies demonstrated mixed results. For example, Murphy and Flynn (2022) and Murphy et al. (2023) both compared misinformation presented as deepfakes and as text to participants and found that deepfakes did not produce significantly different effects than the text conditions in inducing false memories in participants. Hameleers et al. (2022) and Hameleers (2024) found that deepfakes were not perceived as more credible than misinformation presented as text. In direct contrast to this, Lee and Shin (2022) and Hwang et al. (2021) compared deepfakes to misinformation text and texts with photos and found that misinformation messages presented as deepfakes were rated as more credible and vivid. Thus, the very sparse evidence to date suggests that at least when we assess memory distortion or credibility, there is no clear advantage for deepfakes relative to existing forms of

misinformation. Despite much speculation as to the threats posed by deepfake technology and how much more potent deepfake misinformation is relative to other formats (Ajder et al., 2019; Albahar & Almalki, 2019; Liv & Greenbaum, 2020), the current comparative evidence remains limited and mixed.

2.4.3 RQ3. Is there evidence of social, psychological, or technical factors that influence the effects deepfakes have on beliefs, memories, and behaviour?

13 studies provided evidence of psychological, social, or technical factors which influence how deepfake exposure affects our beliefs, memories, and behaviours. Seven of these studies highlighted cognitive ability or a factor related to cognition as a factor that may influence the impact of deepfake exposure. Most of these studies considered cognitive ability as the ability to think analytically and be able to critically assess information or stimulus past simple intuition (Pennycook & Rand, 2021). Ahmed (2021a), Ahmed (2023), Ahmed (2023a), and Ahmed and Chua (2023c) reported that individuals with higher cognitive ability were less likely to self-report the inadvertent sharing of deepfakes. This was measured by directly asking participants whether they had recently shared a deepfake and later discovered that the video was a hoax. This is problematic as it relies on participants' own subjective interpretation of what deepfakes are and does not allow the reporting of deepfake sharing that went unnoticed. Individuals with higher levels of cognitive ability are less likely to consider deepfakes they see as credible or accurate (Ahmed et al., 2023b; Hameleers et al., 2024). Finally, Iacobucci et al., (2021) found that individuals with low levels of bullshit receptivity (the tendency of an individual to believe baseless claims) and primed with knowledge about deepfakes were more likely to recognise that a video was a deepfake. This experiment measured deepfake recognition

by asking participants whether a clip of a face-swapped actor in the movie "The Shining" was similar to the original scene due to the actor's acting ability, or digital video-editing technology.

Five studies found that an individual's political interests, views of politicians, and ideology can also affect how deepfakes are interpreted and consumed (Ahmed, 2021a; Hameleers et al., 2022; Ng, 2023; Sharma et al., 2023; Shin & Lee, 2022). Hameleers et al. (2022) found that deepfake disinformation that was ideologically congruent was perceived as more credible than disinformation that was ideologically incongruent – however this was equally true for text-based disinformation, so does not suggest anything unique to deepfakes. Sharma et al. (2023) found that political brand hate (an individual's hatred towards a political party or politician due to who or what they represent) was associated with sharing deepfakes which represent the ideologically incongruent party or "brand" in poor light. Similarly, Shin and Lee (2022) found that individuals were more likely to believe and intend to share deepfake news which advocated for positions congruent with their own beliefs than individuals with incongruent beliefs. Regardless of political stance or position agreement, Ahmed (2021a) reported that individuals who simply have an interest in politics are more likely to inadvertently share any deepfakes they encounter. However, this finding is unreliable due to the aforementioned issues of interpretation, self-reporting, and measurement of deepfake sharing behaviours. On a separate note, Ng (2023) found that an individuals' susceptibility to deepfake videos of a politician depends on the perceived trustworthiness and dangerousness of that politician. If an individual perceived a politician to be dangerous, then they are less likely to regard deepfake videos of them as authentic, and deepfakes of perceived trustworthy politicians are more likely to be regarded as authentic. Ng (2023) argues that people use an error management system to ascertain truthfulness, and therefore if a politician is perceived as dangerous, then trusting information from that source could be perceived as more costly and thus are less likely to be trusted.

Finally, our search yielded one study which identified social factors which influenced how individuals perceived the credibility and authenticity of deepfakes. Jin et al. (2023) presented participants with a YouTube video, where they manipulated the number of views the video had (representing the popularity of the video) and the number of followers the channel had (representing the trustworthiness of the source). They found that more views increased the perceived credibility of the video, and more followers increased perceived authenticity of the video. This study highlights how individuals have a tendency to perceive deepfakes as more credible and authentic if others appear to do so too. Once again though, the lack of studies on this topic limits the generalizability of these results. Notably, there were no studies exploring technical factors which influence perceptions of deepfakes, such as the quality or resolution of the deepfake, clarity of audio, visual artifacts, or realism of the deepfake.

2.4.4 RQ4. What kind of deepfake technology is being used in current empirical deepfake research?

While not all the deepfakes used in the included studies were available for other researchers to view, it is clear that they varied considerably in terms of resolution, audio quality, accuracy of lip-syncing to voice, level of realism of the faces, and visual artifacts due to deepfake technology. As detailed in Table 1, they were taken from multiple different pre-made sources such as from the Canny AI art installation (Ahmed, 2021b; Ahmed et al., 2023a; Ahmed et al., 2023b; Hwang et al., 2021; Murphy & Flynn, 2022) and YouTube (Iacobucci et al., 2021; Murphy & Flynn, 2022; Vaccari & Chadwick, 2020; Jin et al., 2023) or they were produced by the researcher themselves using software such as the ZAO mobile app (Wu et al., 2021), Faceswap software (Lee & Shin, 2022), custom made by a VFX artist or similar professional (Hameleers, 2024; Hameleerse et al., 2022; Hameleers et al., 2024; Sharma et al., 2023) or Tacotron 2 software (Dobber et al., 2021). The deepfakes included in this review that

were drawn from the Canny AI art installation or YouTube could potentially be of a higher quality and more realistic since they were professionally produced by a team (Canny AI or Buzzfeed), but the downside to these studies is the lack of experimental control over the deepfakes used. The benefits for using pre-made deepfakes are that they enhance ecological validity as these deepfakes were made outside of an experimental setting, allow for comparison of results due to identical deepfake stimuli, and are freely available for other researchers to use to replicate the study allowing others to critically evaluate the work. Deepfakes which were produced by the researchers could be tailored towards answering specific research questions such as in Lee and Shin (2022) or Dober et al., (2021) where they manipulated specific faces or facial features and voices respectively rather than being constrained to designing experiments around pre-existing deepfakes. However, custom deepfakes give rise to issues of reporting as they are often not made readily available online (Dobber et al., 2021; Lee & Shin, 2022). This may be for ethical reasons due to researchers not wanting to contribute to the spread of misinformation online by uploading their custom deepfake, specifically if the deepfake was political in nature.

With deepfakes primarily being an audio-visual form of misinformation, their perceived credibility determines the quality of the deepfake. Only four studies from the 20 experimental studies in this review reported assessing the perceived credibility of the deepfake stimuli in their experiments as an indicator for quality (Dobber et al., 2021; Hameleers et al., 2022; Murphy & Flynn, 2022; Sharma et al., 2023), while five other studies assessed credibility as a dependent variable only (Hameleers, 2024; Hameleers et al., 2024; Hwang et al., 2021; Lu & Chu, 2023; Ng, 2023). Without assessing perceived credibility of deepfakes, it is difficult to interpret null results and assess video quality. Among the studies that reported assessing perceived credibility, Sharma et al. (2023) asked a group of students (separate to the study sample) before the study to distinguish between an authentic video and a deepfake. Dobber et

al. (2021) and Murphy and Flynn (2022) reported asking participants at the end of the study to rate the deepfakes they viewed on how authentic or credible the participant thought the deepfakes were, and Hameleers et al. (2022) assessed credibility indirectly by comparing the perceived credibility and realism scores of statements made in authentic videos to deepfaked ones.

Overall, an overarching pattern seen across the four research questions is the short-term nature of the outcome measures in all of the studies included in this review. None of these studies included any kind of longer-term follow-up, instead all assessed beliefs, memories, or behavioural intentions instantly after the deepfake was viewed. Perhaps of more importance is the fact that most of the included studies did not include any comparator, so could not investigate what it is about deepfakes that may be novel or particularly efficacious at affecting beliefs, memories, and behaviours. The research indicates that while deepfakes may sometimes have a measurable impact on beliefs, memories, and behaviours, it is unclear if these effects are lasting or indeed any greater than effects of less technologically sophisticated forms of misinformation.

2.5 Discussion

In total, our scoping review yielded 22 empirical papers that investigated the effects of viewing deepfakes, compared the deceptive ability of deepfakes to other misinformation mediums, or highlighted individual factors that are associated with vulnerability to deepfake effects. We found some evidence that deepfake exposure can manipulate beliefs about politicians and celebrities, enhance the credibility of misinformation messages, and induce false memories. However, the question of whether deepfakes are manipulative in a way that goes above and beyond existing forms of misinformation (such as text or doctored photos) is unclear, as many studies did not use experimental designs that allowed for such comparison.

Finally, we found that cognitive ability is emerging as a factor that is associated with susceptibility to deepfake misinformation. Overall, while this review indicates that viewing deepfakes can bring about measurable effects, but incidences of conflicting findings and variety in study design and outcome measures used in those studies, means it is too early to draw concrete conclusions. We simply do not have enough evidence to support the widespread concerns that are often voiced in deepfake discourse (Ajder et al., 2019; Albahar & Almalki, 2019; Chesney & Citron, 2019a; Chesney & Citron, 2019b).

To answer RQ1, we were interested in whether there was evidence of measurable effects of deepfakes on beliefs, memories, and behaviours. Eight papers found a measurable impact of deepfakes on beliefs, which broadly suggested that deepfakes may be able to enhance the vividness and credibility of misinformation messages, negatively impact attitudes towards political candidates, enhance mistrust towards real information and sources, enhance our self-esteem and perceived attractiveness, and negatively impact our views of deepfake resurrections. Two empirical studies found that deepfakes have a measurable impact on memories, capable of generating false memories for movies and entertainment videos. Finally, four studies identified a measurable effect of deepfakes on behavioural intentions, suggesting that deepfakes are more likely to be shared online due to their enhanced perceived credibility. However, these results are limited by the diversity of research topic, relatively low number of studies in each research area, and varying sources of deepfakes use in the experiments. We found that there is insufficient empirical data on the measurable effects of deepfake exposure to substantiate the circulating claims about deepfakes, such as jeopardizing national security, sabotaging political candidates, and undermining the foundation of truth in society (Ajder et al., 2019; Albahar & Almalki, 2019; Fallis, 2021).

Perhaps one of the largest gaps in this area of research is the lack of studies investigating any long-term effects of deepfake exposure on peoples beliefs, memories, and/or behaviours.

This type of research is crucial in quantifying the effects of deepfake exposure because initial tests of the effects of seeing deepfakes may not detect significant shifts in attitudes or memories. This is because there could be a delayed "sleeper effect" in the misinformation. A "sleeper effect" can be defined as an increase in the influential power of a persuasive message after a period of time has passed, commonly associated with a discounting cue (Kleinnijenhuis, 2020). The theory behind a "sleeper effect" is that when individuals are exposed to a piece of misinformation, they will encode the message itself, as well as any reason to discount it, into their memory. Over a period of time, the original meaning behind the message is disassociated with the reasons to discount the message. Therefore, when the individual is exposed to the message again, they are more likely to solely recall the meaning behind the piece of misinformation and is more likely to be influenced by it without the discounting reasons (Underwood & Pezdek, 1998). Indeed, studies have found the presence of sleeper effects in memory and attitudinal studies (Murphy et al., 2021; Lariscy & Weaver, 1999; Underwood & Pezdek, 1998). The potential presence of a "sleeper effect" suggests that deepfake exposure may have a significant effect on shifting beliefs, memories, or behaviours, but current deepfake research has not been designed to explore this question creating a knowledge gap.

Considering the substantial attention given to the possible political consequences of deepfakes (Ajder et al., 2019; Chesney & Citron, 2019a; Wilkerson, 2021), there appears to be a notable gap between these discussions and the amount of directly relevant empirical evidence. Broinowski (2022) has documented the history of alarmism at the advent of new technologies, from how telephones were speculated to eliminate privacy (Orben, 2020), to fears of a dystopian future with the arrival of the internet in the 1990s (Orben, 2020). Such periods of heightened concern have often been accompanied by extensive critical commentary on new technologies, sometimes with less attention to the social contexts in which those technologies are used. Johnson and Verdicchio (2017) describe a similar alarmism for AI technology as a

phenomenon known as “sociotechnical blindness” where individuals fixate on the abstract concept of AI, rather than the human beings and social institutions utilising it. A stronger focus on the social and institutional contexts in which deepfakes are produced and used may help complement existing discussion of their abstract risks and support more targeted empirical research on their harms. While our review calls for more empirical research on the harms of deepfake exposure, it also highlights the value of adopting a careful, evidence-led approach to assessing the harms of emerging technologies. This sentiment has been echoed by some emerging literature decrying technological alarmism (Orben, 2020).

For RQ2, we wanted to see if there was evidence that compares the effects of deepfakes with pre-existing forms of misinformation to identify whether deepfakes are uniquely persuasive and deceptive. We only found six studies which provided evidence for RQ2, and they suggest that deepfakes may not be uniquely effective at distorting memories compared to pre-existing forms of misinformation, but the evidence is mixed when assessing whether deepfakes enhance the credibility of misinformation messages more than pre-existing formats. The small number of studies including comparators is noteworthy, particularly because many empirical studies document some effect of deepfakes without directly testing whether that effect differs from effects observed with other misinformation formats. It is imperative that researchers clarify the unique position of deepfakes in misinformation literature, and we strongly recommend that future experiments include a less technically advanced comparator, such as simple text-based misinformation. This would help establish whether deepfakes should be treated as just another form of misinformation or whether they are indeed a special case (Chawla, 2019; Kietzmann et al., 2020; Vaccari & Chadwick, 2020).

In RQ3, we were interested in any evidence which outlines psychological, social, or technical factors which influence the effect of deepfakes on beliefs, memories, and behaviours. We found 13 studies which explored how psychological and social factors such as cognitive

ability, message congruency, and affect individuals' beliefs memories, and behaviours. Results from these studies broadly suggest that individuals with higher cognitive ability are less likely to consider deepfakes as credible and share them, while individuals who agree with the message behind misinformation presented as deepfakes are more likely to believe and share them. However, these results are obfuscated by poor methodology and therefore are difficult to draw conclusions from. Despite this, our findings on cognitive ability reflect results from broader misinformation literature i.e. cognitive ability is associated with reduced susceptibility to deepfakes. This has been demonstrated in existing research concerning misinformation in the form of news headlines, social media posts, or other text-based formats (Bronstein et al., 2019; Pennycook & Rand, 2020; Pennycook & Rand, 2021). Some of the studies included in this review suggest that deepfake-based misinformation is especially potent in spreading misinformation due to the multi-modal aspect of deepfakes enhancing the vividness and credibility of the misinformation (Ahmed & Chua, 2023; Hwang et al., 2021; Lee & Shin, 2022), but other studies dispute this (Hameleers, 2024; Hameleers et al., 2022; Murphy & Flynn, 2022; Murphy et al., 2023). The lack of comparative studies assessing the predictive power of cognitive ability for misinformation susceptibility in different formats means that we cannot deduce if there is anything special about deepfakes in this regard, or if these findings are merely an extension of what is already established as a predictive factor in misinformation susceptibility.

We were interested in identifying the types of deepfakes being used in deepfake empirical papers for RQ4. The experiments included in this review used deepfakes of varying quality from different sources, including Tacotron 2 and AI based lip synchronization techniques, Canny AI art installation, Faceswap software, ZAO deepfake app, and publicly available deepfakes taken from YouTube, Reddit and other internet sites. Many studies did not share the deepfakes used in the experiment, making it difficult to compare the quality of

materials between studies, though we did note that some of the deepfakes that were accessible were of quite poor quality. Little research has been done to identify the effect of the "quality" of a deepfake on its effectiveness to influence beliefs, memories, and/or behaviours. In fact, there is little consensus on what a "high quality" deepfake is (Lyu, 2020). It is possible that the quality of a deepfake significantly affects its influential ability as some research has demonstrated that the perception of deepfakes depends on its realism and human-likeness (Kaate et al., 2023). Despite this gap in literature, we do not advocate for the standardisation of deepfake quality in empirical research because researchers may choose specific software/deepfake stimuli for different reasons (accessibility, financial constraints, resource and time constraints) thus pushing for a standardized quality may act as a barrier restricting empirical deepfake research. We do however suggest checking the perceived credibility of deepfake stimuli in future experimental research. This allows future researchers to identify whether the deepfakes were effective at manipulating participants. We suggest pilot testing of the deepfake stimuli to assess their rate of identification compared to authentic videos before conducting the study and clear reporting of measures of video quality in the write-up.

Our findings suggest that despite widespread claims about the ease of access to deepfake technology (Albahar & Almalki, 2019; Westerlund, 2019; Yadlin-Segal & Oppenheim, 2021), researchers in this area clearly find it difficult to create custom high-quality, realistic deepfakes. This may be due to a lack of time, computing infrastructure, or expertise, but the end result is that very few researchers have used custom-created high-quality deepfakes in their studies (Hameleers 2024; Hameleers et al., 2022; Hameleers et al., 2024). Thus, our results may not reflect the true manipulative efficacy of high-quality deepfakes that may become more accessible in the future. Where researchers cannot create their own high-quality deepfakes, some have used existing YouTube videos (Murphy & Flynn, 2022), while others have resorted to asking participants to imagine hypothetical deepfakes (Fido et al., 2022).

Neither solution is without its drawbacks. Using existing videos incurs the risk that participants will have seen the video before, and the studies included in this review differed in whether they removed participants who had previously seen the video (Murphy & Flynn, 2022) or opted to retain them in the analyses (Vaccari & Chadwick, 2021). Asking participants to imagine a deepfake is clearly a difficult task and participants' ability to do so will be hugely affected by their technical literacy (Fido et al., 2022). For these reasons, we strongly encourage more collaboration between psychologists who are interested in testing the cognitive effects of deepfake exposure and those with more technical expertise who can build convincing custom deepfakes for use in these studies. We also encourage researchers to make their materials available to other researchers where possible, so that the quality of the deepfakes can be assessed. Only by establishing more rigorous and transparent methodologies will we gain insight into the effects of deepfakes.

Through reviewing existing research on the topic, we identified a number of methodological limitations that impede our ability to draw strong conclusions from the existing data. Firstly, while some studies in our review have included behavioural measures such as sharing intentions (Ahmed, 2021a; Ahmed and Chua, 2023; Hwang et al., 2021; Sharma et al., 2023), there is little empirical deepfake research exploring other real-world behavioural consequences of deepfakes such as voting intentions or intentions to vaccinate. Our review highlights an overall lack of research that investigates the effect of viewing deepfakes on people's behaviour. This finding is not unique to deepfake research – a recent review of misinformation research found that only 29% of empirical misinformation studies between the years 2016 – 2022 examined the impact of misinformation on behaviours or behavioural intentions (Murphy et al., 2023). Most of those studies only examined online behaviours or behavioural intentions (e.g. intent to share a news story on social media), so that only 11% of studies examined offline behaviours or behavioural intentions (e.g. intent to get vaccinated).

We echo the call by Murphy et al. (2023) and encourage researchers to examine behavioural effects of deepfakes, where we currently have little evidence but rampant speculation (Bastick, 2021; Chesney & Citron, 2019a; Chesney & Citron, 2019b). Of course, research in this area will need to follow best practice in terms of ethical practices, including effective debriefing of participants as soon as possible (Greene et al., 2023).

While we have highlighted a tendency of existing research to overlook the need for empirical literature in favour of alarmist speculations, we feel the need to also highlight the current real-world harms of deepfakes, and interestingly, the lack of empirical deepfake literature underpinning these real-world applications. An application of deepfakes that did not appear in the scoping review is pornography. Estimates suggest that 96% of existing online deepfakes are pornographic in nature (Ajder et al., 2019), it is therefore of note that there were no empirical studies quantifying the effects of deepfake porn exposure in the scoping review (though our search terms did not seek these studies out). Deepfake pornography can potentially pose a threat to the depicted person's reputation and mental health through its dissemination and subsequent distress, similar to how non-consensual spread of intimate images online can cause emotional distress (Henry et al., 2017; Karasavva & Noorbhai, 2021). Deepfake pornography may also pose threats to consumers, similar to how general pornography can negatively "shape sexual scripts" and influence how consumers behave during sexual encounters (Wright, 2011). Legislation and research are converging on deepfake pornography in an attempt to understand and regulate its impacts through policy (Albahar & Almalki, 2019; Gieseke, 2020; Hancock & Bailenson, 2021; Karasavva & Noorbhai, 2021). Rigorous, ethical research that gives us a clearer understanding of the effects of deepfake pornography is urgently required, as we attempt to develop effective policies and countermeasures.

There are several limitations to the current study. Firstly, due to the different interpretations and definitions of deepfakes in deepfake literature (Whittaker et al., 2023), it is

possible that some studies used deepfakes in an experimental manner but failed to describe them with the term “deepfake”. While we took all reasonable steps to include these studies in our review, it is possible that we could have missed them. Secondly, we did not include any preprints or conference abstracts in our search. This study is a scoping review of peer-reviewed literature on the topic of deepfake effects on beliefs, memories and behaviours and therefore those elements were omitted. However, it is possible that informative research related to our research questions could be found in preprints and conference abstracts, and therefore some information related to our research topic could have been missed.

In conclusion, the current scoping review maps the current state of empirical deepfake literature and provides specific suggestions for future empirical deepfake research. A key priority is an increase in empirical research quantifying the effects of deepfake exposure, alongside ongoing conceptual and theoretical work. We also advocate for more empirical deepfake studies to test the potential longitudinal effects of deepfake exposure, to test a broader range of behavioural implications of deepfake exposure and include comparators to deepfakes in experiments such as text or audio misinformation to investigate the unique effectiveness of deepfakes. While research in this field is clearly still in its infancy, this review highlights some persistent methodological issues as well. In general, we conclude that there are a great many papers decrying the risks and harms of deepfakes without the necessary empirical evidence to support these alarmist claims. We cannot hope to design interventions or implement regulations to combat these concerns unless we first quantify the issues. Therefore, we call for further (methodologically rigorous) research on the effects of deepfake exposure.

Acknowledgements

This publication has emanated from research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 13/RC/2094_2. Initials of authors who received the award: DC, JT, CL, and GM

2.6 References

Ahmed, S. (2021). Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics, 57*, 101508.

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

Ahmed, S. (2021). Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. *Personality and Individual Differences, 182*, 111074.

Ahmed, S., & Chua, H. W. (2023). Perception and deception: Exploring individual responses to deepfakes across different modalities. *Heliyon, 9*(10).

Ahmed, S., Ng, S. W., & Bee, A. W. (2023). Understanding the role of fear of missing out and deficient self-regulation in sharing of deepfakes on social media: Evidence from eight countries. *Frontiers in Psychology, 14*, 1127507.

Ahmed, S., Wang, Y., & Bee, A. W. (2023). Adjusting news accuracy perceptions after deepfakes exposure: Evidence from a non-Western context. *Telematics and Informatics, 84*, 102047.

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

Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology, 97*(22), 3242–3250.

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology, 8*(1), 19–32.

Balmas, M. (2014). When fake news becomes real: Combined exposure to multiple news sources and political attitudes of inefficacy, alienation, and cynicism. *Communication Research, 41*(3), 430–454.

Bastick, Z. (2021). Would you notice if fake news changed your behavior? An experiment on the unconscious effects of disinformation. *Computers in Human Behavior, 116*, 106633.

Broinowski, A. (2022). Deepfake nightmares, synthetic dreams: A review of dystopian and utopian discourses around deepfakes, and why the collapse of reality may not be imminent—Yet. *Journal of Asia-Pacific Pop Culture, 7*(1), 109–139.

Bronstein, M. V., 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*(1), 108–117.

Bryanov, K., & Vziatysheva, V. (2021). Determinants of individuals' belief in fake news: A scoping review—Determinants of belief in fake news. *PLoS ONE, 16*(6), e0253717.

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

Chesney, R., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*, 1753.

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Affairs, 98*, 147.

Cole, S. (2017, July 1). AI-assisted fake porn is here and we're all fucked. Vice. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

Coupe, J. (2020, September 24). Manipulating memories: Archives, History and Deepfakes [Video]. YouTube. https://www.youtube.com/watch?v=PQVFgqHAZ5M

de Saint Laurent, C., Murphy, G., Hegarty, K., & Greene, C. M. (2022). Measuring the effects of misinformation exposure and beliefs on behavioural intentions: A COVID-19 vaccination study. *Cognitive Research: Principles and Implications, 7*(1), 87.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics, 26*(1), 69–91.

Dunne, J. E. (2018, July 23). Cincinnati art museum to unveil new contemporary conceptual art in life: Gillian Wearing. Cincinnati Art Museum. https://www.cincinnatiartmuseum.org/media/206653/gillianwearing-final.pdf

Fallis, D. (2015). What is disinformation? *Library Trends, 63*(3), 401–426.

Fallis, D. (2021). The epistemic threat of deepfakes. *Philosophy & Technology, 34*(4), 623–643.

Fido, D., Rao, J., & Harper, C. A. (2022). Celebrity status, sex, and variation in psychopathy predicts judgements of and proclivity to generate and distribute deepfake pornography. *Computers in Human Behavior, 129*, 107141.

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.

Gieseke, A. P. (2020). “The new weapon of choice”: Law's current inability to properly address deepfake pornography. *Vanderbilt Law Review, 73*, 1479.

Greene, C. M., Nash, R. A., & Murphy, G. (2021). Misremembering Brexit: Partisan bias and individual predictors of false memories for fake news stories among Brexit voters. *Memory, 29*(5), 587–604.

Greene, C. M., & Murphy, G. (2021). Quantifying the effects of fake news on behavior: Evidence from a study of COVID-19 misinformation. *Journal of Experimental Psychology: Applied, 27(*4), 773.

Greene, C. M., Bradshaw, R., Huston, C., & Murphy, G. (2022). The medium and the message: Comparing the effectiveness of six methods of misinformation delivery in an eyewitness memory paradigm. *Journal of Experimental Psychology: Applied, 28*(4), 677.

Greene, C. M., de Saint Laurent, C., Murphy, G., Prike, T., Hegarty, K., & Ecker, U. K. (2023). Best practices for ethical conduct of misinformation research: A scoping review and critical commentary. *European Psychologist, 28*(3), 139.

Hameleers, M., Powell, T. E., Van Der Meer, T. G., & Bos, L. (2020). A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication, 37*(2), 281–301.

Hameleers, M., van der Meer, T. G., & Dobber, T. (2022). You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media + Society, 8*(3), 20563051221116346.

Hameleers, M. (2024). Cheap versus deep manipulation: The effects of cheapfakes versus deepfakes in a political setting. *International Journal of Public Opinion Research, 36*(1), edae004.

Hameleers, M., van der Meer, T. G., & Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior, 152*, 108096.

Hancock, J. T., & Bailenson, J. N. (2021). The social impact of deepfakes. *Cyberpsychology, Behavior, and Social Networking, 24*(3), 149–152.

Henry, N., Powell, A., & Flynn, A. (2017). Not just “revenge pornography”: Australians’ experiences of image-based abuse. A summary report. RMIT University.

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.

Iacobucci, S., De Cicco, R., Michetti, F., Palumbo, R., & Pagliaro, S. (2021). Deepfakes unmasked: The effects of information priming and bullshit receptivity on deepfake recognition and sharing intention. *Cyberpsychology, Behavior, and Social Networking, 24*(3), 194–202.

Jin, X., Zhang, Z., Gao, B., Gao, S., Zhou, W., Yu, N., & Wang, G. (2025). Assessing the perceived credibility of deepfakes: The impact of system-generated cues and video characteristics. *New Media & Society, 27*(3), 1651–1672.

Kaate, I., Salminen, J., Jung, S. G., Almerekhi, H., & Jansen, B. J. (2023). How do users perceive deepfake personas? Investigating the deepfake user perception and its implications for human–computer interaction. In *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter* (pp. 1–12).

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.

Kietzmann, J., Lee, L. W., McCarthy, I. P., & Kietzmann, T. C. (2020). Deepfakes: Trick or treat? *Business Horizons, 63*(2), 135–146.

Kleinnijenhuis, J. (2020). Sleeper effect. In *The International Encyclopedia of Media Psychology* (pp. 1–7). Wiley.

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience, 24*(11).

Lamar, K. (2022, May 31). The Heart Part 5 [Video]. YouTube. https://www.youtube.com/watch?v=uAPUkgeiFVY&ab_channel=KendrickLamarVEVO

Leask, J., Booy, R., & McIntyre, P. B. (2010). MMR, Wakefield and The Lancet: What can we learn? *Medical Journal of Australia, 193*(1), 5–7.

Lee, J., & Shin, S. Y. (2022). Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news. *Media Psychology, 25(*4), 531–546.

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

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory, 12*(4), 361–366.

Loomba, S., de Figueiredo, A., Piatek, S. J., de Graaf, K., & Larson, H. J. (2021). Measuring the impact of COVID-19 vaccine misinformation on vaccination intent in the UK and USA. *Nature Human Behaviour, 5*(3), 337–348.

Lu, H., & Chu, H. (2023). Let the dead talk: How deepfake resurrection narratives influence audience response in prosocial contexts. Computers in Human Behavior, 145, 107761.

Lyu, S. (2020). Deepfake detection: Current challenges and next steps. In *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)* (pp. 1–6). IEEE.

Mahmud, M. R., Bin Reza, R., & Ahmed, S. Z. (2023). The effects of misinformation on COVID-19 vaccine hesitancy in Bangladesh. *Global Knowledge, Memory and Communication, 72*(1–2), 82–97.

Messaris, P., & Abraham, L. (2001). The role of images in framing news stories. In *Framing Public Life* (pp. 231–242). Routledge.

Mihailova, M. (2021). To dally with Dali: Deepfake (inter)faces in the art museum. *Convergence, 27*(4), 882–898.

Murphy, G., Loftus, E. F., Grady, R. H., Levine, L. J., & Greene, C. M. (2019). False memories for fake news during Ireland's abortion referendum. *Psychological Science, 30*(10), 1449–1459.

Murphy, G., Lynch, L., Loftus, E., & Egan, R. (2021). Push polls increase false memories for fake news stories. *Memory, 29*(6), 693–707.

Murphy, G., & Flynn, E. (2022). Deepfake false memories. *Memory, 30*(4), 480–492.

Murphy, G., Ching, D., Twomey, J., & Linehan, C. (2023). Face/Off: Changing the face of movies with deepfakes. *PLOS ONE, 18*(7), e0287503.

Murphy, G., de Saint Laurent, C., Reynolds, M., Aftab, O., Hegarty, K., Sun, Y., & Greene, C. M. (2023). What do we study when we study misinformation? A scoping review of experimental research (2016–2022). *Harvard Kennedy School Misinformation Review*.

Nash, R. A., Wade, K. A., & Lindsay, D. S. (2009). Digitally manipulating memory: Effects of doctored videos and imagination in distorting beliefs and memories. *Memory & Cognition, 37*, 414–424.

Nash, R. A., Wade, K. A., & Brewer, R. J. (2009). Why do doctored images distort memory? *Consciousness and Cognition, 18*(3), 773–780.

Nash, R. A., Wade, K. A., Garry, M., Loftus, E. F., & Ost, J. (2017). Misrepresentations and flawed logic about the prevalence of false memories. *Applied Cognitive Psychology, 31*(1), 31–33.

Ng, Y. L. (2023). An error management approach to perceived fakeness of deepfakes: The moderating role of perceived deepfake targeted politicians' personality characteristics. *Current Psychology, 42*(29), 25658–25669.

Ognyanova, K., Lazer, D., Robertson, R. E., & Wilson, C. (2020). Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*.

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews, 5*, 1–10.

Orben, A. (2020). The Sisyphean cycle of technology panics. *Perspectives on Psychological Science, 15*(5), 1143–1157.

Pantserev, K. A. (2020). The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. In *Cyber defence in the age of AI, smart societies and augmented humanity* (pp. 37–55).

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality, 88*(2), 185–200.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences, 25*(5), 388–402.

Pickrell, J. E., McDonald, D. L., Bernstein, D. M., & Loftus, E. F. (2016). Misinformation effect. In *Cognitive illusions* (pp. 406–423). Psychology Press.

Pierri, F., Perry, B. L., DeVerna, M. R., Yang, K. C., Flammini, A., Menczer, F., & Bryden, J. (2022). Online misinformation is linked to early COVID-19 vaccination hesitancy and refusal. *Scientific Reports, 12*(1), 5966.

Qobrtay, A. (2022, March 3). “Star Wars” fans love CGI Luke Skywalker, but deepfake implications are dangerous. Daily Targum. https://dailytargum.com/article/2022/03/star-wars-fans-love-cgi-luke-skywalker-but-deepfake-implications-are

Rini, R., & Cohen, L. (2022). Deepfakes, deep harms. *Journal of Ethics & Social Philosophy, 22*, 143.

Sacchi, D. L., Agnoli, F., & Loftus, E. F. (2007). Changing history: Doctored photographs affect memory for past public events. *Applied Cognitive Psychology, 21*(8), 1005–1022.

Sharma, I., Jain, K., Behl, A., Baabdullah, A., Giannakis, M., & Dwivedi, Y. (2023). Examining the motivations of sharing political deepfake videos: The role of political brand hate and moral consciousness. *Internet Research, 33*(5), 1727–1749.

Shin, S. Y., & Lee, J. (2022). The effect of deepfake video on news credibility and corrective influence of cost-based knowledge about deepfakes. *Digital Journalism, 10*(3), 412–432.

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *PLOS ONE, 18*(10), e0291668.

Underwood, J., & Pezdek, K. (1998). Memory suggestibility as an example of the sleeper effect. *Psychonomic Bulletin & Review, 5*, 449–453.

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*(1), 2056305120903408.

Van Bavel, J. J., Harris, E. A., Pärnamets, P., Rathje, S., Doell, K. C., & Tucker, J. A. (2021). Political psychology in the digital (mis) information age: A model of news belief and sharing. *Social Issues and Policy Review, 15*(1), 84–113.

Van Der Linden, S. (2022). Misinformation: Susceptibility, spread, and interventions to immunize the public. *Nature Medicine, 28*(3), 460–467.

Wade, K. A., Garry, M., Don Read, J., & Lindsay, D. S. (2002). A picture is worth a thousand lies: Using false photographs to create false childhood memories. *Psychonomic Bulletin & Review, 9*(3), 597–603.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11).

Whittaker, L., Mulcahy, R., Letheren, K., Kietzmann, J., & Russell-Bennett, R. (2023). Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda. *Technovation, 125*, 102784.

Wilkerson, L. (2021). Still waters run deep (fakes): The rising concerns of "deepfake" technology and its influence on democracy and the First Amendment. *Missouri Law Review, 86*, 407.

Wu, L., Morstatter, F., Carley, K. M., & Liu, H. (2019). Misinformation in social media: Definition, manipulation, and detection. *ACM SIGKDD Explorations Newsletter, 21*(2), 80–90.

Wu, F., Ma, Y., & Zhang, Z. (2021). "I found a more attractive deepfaked self": The self-enhancement effect in deepfake video exposure. *Cyberpsychology, Behavior, and Social Networking, 24*(3), 173–181.

Wright, P. J. (2011). Mass media effects on youth sexual behavior assessing the claim for causality. *Annals of the International Communication Association, 35*(1), 343–385.

Yadlin-Segal, A., & Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. *Convergence, 27*(1), 36–51.

Chapter 3 Encountering Deepfakes: A Thematic Analysis of Comments on a Music Video featuring Deepfake Content

This paper was published in the proceedings of CSCW 2025.

Ching, D., Twomey, J., Linehan, C., & Murphy, G. (2025). Encountering Deepfakes: A Thematic Analysis of Comments on a Music Video featuring Deepfake Content. *Proceedings of the ACM on Human-Computer Interaction, 9*(7), 1-23.

Study Motivation

This study served as one of the first qualitative studies exploring responses to deepfakes in a naturalistic environment outside of a laboratory setting. The music video that became the focus of this study that featured deepfakes, Kendrick Lamar's "The Heart Part 5", had just been released at the time and garnered millions of views. We used this music video as an opportunity to study how people experienced deepfakes in an entertainment environment. This research was essential due to the limited understanding of how deepfakes are encountered and experienced in everyday life. This gap in qualitative knowledge about deepfake experiences was highlighted in Chapter 2 where no qualitative studies had been found exploring deepfake effects. It is important to investigate not only the quantitative effects of deepfakes on beliefs, memories, and behaviours, but also the qualitative effects, including individuals' responses and attitudes toward deepfakes. Combining both approaches allows for a more nuanced understanding of deepfakes beyond isolated controlled responses to deepfakes in experimental settings which serves to inform the design of future

experimental studies. Despite the widespread fear and sensationalism surrounding deepfakes, it is concerning that so little is known about how individuals respond to them in real-world contexts, a setting where they are most likely to come across these deceptive media.

Author Contribution

Didier Ching, Conor Linehan, and Gillian Murphy conceptualized the study. Didier Ching screened the comments used for the thematic analysis and conducted the thematic analysis. John Twomey assisted technically by extracting the comments from the YouTube video and contributed to the analysis process by reviewing the comments and codes as a secondary reviewer. Conor Linehan and Gillian Murphy also contributed to the thematic analysis and shaping the final thematic structure. Didier Ching wrote the original draft and edited the document based on feedback from John Twomey, Conor Linehan, and Gillian Murphy. The final version of the manuscript was approved by every author.

3.1 Abstract

Much early research on deepfakes has focused on speculating about potential future harms of the technology. Everyday uses and experiences of existing deepfake content are, surprisingly, comparatively understudied. For example, there is little research that seeks to understand how people encounter and respond to deepfake content in today's entertainment media. In this study, we conducted a reflexive thematic analysis of user's comments made in response to a Kendrick Lamar YouTube music video that featured deepfake content. In total, we analysed 1467 comments. Our findings are described by six themes; advancing realism of deepfakes, seeing is not believing, predicting dystopian threats, enhancing the experience of the song, strong emotional reactions, and 'a toy for play'. People generally enjoyed the creativity and emotional impact that deepfakes brought to the video but those who were unaware of deepfake technology were left confused and doubting themselves. This study offers a naturalistic view of how deepfakes are encountered in everyday life.

3.2 Introduction

"Deepfakes" are a form of video-editing technology that uses AI neural networks to digitally superimpose the face of one person onto another in video format (Chawla, 2019). The first ever recorded instances of deepfakes occurred in 2017 on the social networking site Reddit, where the faces of female celebrities were superimposed onto the bodies of actresses in pornographic videos (Cole, 2017). Commentators have suggested that deepfakes have the potential to facilitate a range of different harms to people, institutions and society in the near future. This technology has often been hypothesised to be able to spread misinformation to the public leading to significant consequences such as misinform citizens on important events (Albahar & Almalki, 2019; Westerlund, 2019), manipulating election results through the defamation of political candidates (Chesney & Citron, 2019), and undermining epistemic trust in society due to citizens being unable to differentiate between real and false information (Fallis, 2020). While these hypothesised harms have not yet come to pass, real-world impact of deepfakes have been documented in the form of financial fraud, where malicious actors use deepfake technology to pose as influential people to trick individuals into paying money (Bhati, 2023; Brewster, 2023) and non-consensual deepfake porn of mostly female celebrities which can cause psychological and reputational harm (Maddocks, 2020).

While there have been many research papers discussing the potential harmful effects of deepfake technology, many gaps in knowledge still exist within deepfake literature that are hindering our understanding of this relatively new technological advancement. Currently, there is a lack of empirical literature to address fundamental questions regarding the potential psychological effects that deepfake exposure may have on an individual's beliefs, memories, and behaviour without which it is difficult to ascertain the impact of deepfake technology. Furthermore, existing empirical deepfake literature does not explore how deepfakes are experienced in everyday settings and are usually experimental in nature. The lack of

observational and naturalistic investigations of deepfakes indicates that we may not have a proper understanding of the everyday experience of deepfakes. This everyday experience is valuable because it shows us how individuals are likely to react when they encounter faked videos “in the wild”. For example, are individuals sceptical of seeing deepfakes or are they likely to blindly believe in the contents of the faked video?

One common way that many individuals experience deepfakes in their everyday lives is through entertainment and media. For example, some music videos which incorporated deepfakes on YouTube have garnered 44 million views (Lamar, 2022) and deepfakes incorporated into awareness and educational videos on YouTube have almost 10 million views (Buzzfeed, 2018). Channels that use deepfakes as a form of entertainment on other large social media platforms such as TikTok have garnered millions of followers such as “Metaphysic.ai” which posts deepfakes of Tom Cruise performing different tasks. Despite deepfake technology beginning to be incorporated into television (David, 2022; Gamage et al., 2022), music videos (Iamtherealak, 2020; Lamar, 2022; Wilson, 2021) and museums (Mihailova, 2021), there are very few empirical studies of how people experience deepfakes when they are encountered recreationally and in naturalistic settings. The proliferation of deepfake content in the entertainment industry and in the everyday lives of the public warrants a closer examination into how individuals live with and interact with this technology.

In the current paper, we study an instance of creative deepfake use in popular media. Specifically, on the 9th of May 2022, artist Kendrick Lamar released a music video on YouTube titled “The Heart Part 5” where he utilised deepfake technology to superimpose the faces of various influential men, some of whom are deceased, onto his own face while rapping (Lamar, 2022). The nature of this music video offered a perfect opportunity to explore how people think, feel, and respond to deepfake use in an entertainment environment by qualitatively analysing the comments in the comment section of the video. The naturalistic

environment in which we obtained our data provides a unique opportunity to examine the everyday responses and experiences people have with deepfake technology. We have the opportunity to understand how people encounter, understand, and experience the technology "in the wild." (Rogers, 2023). We analyse comments that they offer freely, without the self-conscious awareness that they are being observed. Therefore, our research question is: "how did YouTube commenters discuss the experience of encountering deepfakes in the music video by Kendrick Lamar".

The contribution of this paper is in describing and understanding the lived experience of encountering deepfake technology in a naturalistic environment. In the spirit of third wave HCI, we aim to explore deepfake technology as it is experienced by the everyday consumer in their own situated context. As deepfake technology is becoming increasingly pervasive in the entertainment media that the everyday person consumes, it is vital to understand the responses and interactions of those who are consuming and using this technology. We contribute to HCI in multiple different ways. Firstly, we provide an unprecedented in-depth analysis and understanding of how everyday individuals perceive and interact with deepfake technology in a naturalistic environment. Secondly, we discuss possible explanations for these interactions through psychological theory. Thirdly, we outline implications that these understandings have for designers, technologists, and policymakers on the use of deepfakes in everyday entertainment settings.

3.3 Background

3.3.1 Research on Deepfakes

"Deepfakes" are a form of digital video manipulation which first emerged on Reddit in 2017 (Cole, 2017). Using AI neural network technology, the facial features of one person can

be superimposed onto another person in a video format, essentially depicting one person doing something that they never did in a relatively convincing manner (Albahar & Almalki, 2019; Temir, 2020). The resulting video is known as a "deepfake". Much research on deepfake technology sets out to speculate on and discuss the potential harms that deepfake technology poses to individuals and society (Albahar & Almalki, 2019; Chesney & Citron, 2019; Fallis, 202; Westerlund, 2019). Those papers often do not involve empirical analyses and discuss deepfake technology in legal (Chesney & Citron, 2019), philosophical (Fallis, 2020), and societal (Westerlund, 2019) contexts. Existing research often discusses the potential of deepfake technology to spread misinformation. For example, Chesney and Citron (2019) assess the dangers of deepfakes in a legal and political environment by discussing their causes and consequences. While they also discuss possible pro-social uses of deepfake technology such as in education by recreating historical figures to teach students or in artistic environments to resurrect deceased individuals to play themselves in films, the majority of the paper focuses on discussing the harmful effects of this technology. Chesney and Citron (2019) speculate that deepfakes could lead to the erosion of trust in news media as people would be unable to differentiate between real and fake news, as well as the jeopardization of national security by potentially spreading misinformation in military settings. This emphasis on the detrimental applications of this technology is frequently seen in literature (Ajder et al., 2019; Albahar & Almalki, 2019; Westerlund, 2019).

Empirical work on deepfake technology often focuses on detection and countermeasures against the aforementioned threats of deepfakes (Li et al., 2020; Lees et al., 2021; Qian et al., 2021). As deepfakes become increasingly realistic, these studies aim to explore different methods of detecting visual artifacts such as ones due to errors in geometric or illumination estimation (Matern et al., 2019). There are significantly lower number of studies that focus on empirically evaluating the harms of existing deepfake technology. However, there

are examples of studies that investigated the psychological impacts of deepfake exposure (Dobber et al., 2021; Lee & Shin, 2022). For example, Dobber et al. (2021) exposed participants to a self-made deepfake of a politician speaking inappropriate things. They measured participant's attitudes towards the politician among those who saw the deepfake and compared them to those who didn't. This study aimed to explore whether deepfakes can manipulate people's attitudes in a political context and found that deepfakes only affected individuals who would be particularly offended by the deepfake (i.e., Christians would be more offended by a deepfake of a Christian politician).

Another subset of research on deepfakes focuses on its applications in pornography, where the faces of mostly women are superimposed onto actresses in pornographic videos. Indeed, current statistics estimate up to 96% of existing deepfakes on the internet are pornographic in nature (Ajder et al., 2019). Current research in this area focuses on identifying the harms of deepfake technology against women (Gosse & Burkell, 2020; Maddocks, 2020) and how legislation is currently equipped to deal with this issue (Gieseke, 2020; Johnson & Verdichhio, 2017; Kugler & Pace, 2021). Gosse and Burkell (2020) empirically analysed how news media discusses the harmful impacts of deepfakes and found that the harms deepfake pornography poses to women (invasion of privacy, defamation, harassment, blackmail) are often given less attention by news media compared to speculations surrounding the hypothetical harms of deepfakes such as political sabotage and misinformation spread. These findings highlight a concerning trend in deepfake research regarding the overemphasis in discussing the hypothetical futures of deepfake technology, rather than their current impacts and experiences.

While these psychological studies play a role in quantifying the impact of deepfakes, they often lack insights from a psychological perspective known as "everyday psychology" or "critical psychology". This perspective situates human actions and reactions in their own

complex subjective, social, and environmental context and emphasises the value of empirically investigating psychological phenomena from the viewpoint of an individual and how they experience local situations, daily life practices, and technologies on an everyday basis (Schraube & Højholt, 2015). Some HCI and CSCW literature have adopted this perspective in their research as well (Deng et al., 2024; Magee et al., 2017) and this train of thought largely echoes the third wave of HCI research which focused on understanding the spread of technology from the workplace to the everyday life and culture of users (Bødker, 2006). This perspective holds value when studying the psychological impact of viewing deepfakes because it urges psychological research to focus on investigating the influence of deepfakes outside of a laboratory and controlled environment and into the everyday environments that individuals who are likely to encounter deepfakes live in.

Methodologically, few empirical studies have investigated the human experience of deepfakes in a naturalistic environment where we do not expect to see deepfakes. Some HCI studies have approached this issue indirectly by examining how individuals in naturalistic environments discuss and conceptualize deepfakes (Gamage et al., 2022), but none have examined how individuals would respond to deepfake exposure in their own personal environments. It is vitally important to explore deepfakes in a day-to-day context as it gives us insight into how deepfakes are realistically experienced, rather than how we speculate deepfakes might be experienced.

3.3.2 Deepfakes in Entertainment

Despite the many malicious applications of deepfakes, people have also begun to utilise deepfakes in constructive ways such as in entertainment and popular media. In the film and TV industries, deepfake technology has been used to replace the faces of actors with younger versions in an attempt to represent them in a de-aged form. For example, in the popular Star

Wars TV series “The Book of Boba Fett”, a young version of Mark Hamill’s face (the original actor for Luke Skywalker) was superimposed onto another actor using deepfake technology to represent the character of Luke Skywalker (Gamage et al., 2022). Fans reportedly had mixed responses to seeing a de-aged Mark Hamill (David, 2022).

Deepfake technology has also been used to recreate the late painter Salvador Dali on digital screens in the museum and interact with visitors. Director of this project Hank Hine has suggested that the deepfake is meant to create a strong emotional link between the visitor and Dali, as well as humanizing the deceased (Mihailova, 2021). This example highlights the potential of deepfakes to create tangible versions of individuals.

Furthermore, creative uses of deepfakes have also permeated the music industry with songs such as “Family Tree” by AK and “Self” by Steven Wilson utilising deepfake technology to superimpose faces of celebrities onto their own whilst singing (Iamtherealak, 2020; Wilson, 2021). Both videos have over 500,000 views at the time of writing and gained considerable publicity (Blistein, 2021; Hadusek, 2021). Responses to these videos were generally positive, but some acknowledged the controversial nature of the technology (Archanoth 2020).

Despite the abundance of examples of deepfake technology being used in entertainment and artistic environments, there has been little empirical research analysing how people respond and experience deepfakes in these contexts. In one example, Murphy et al., (2023) found that individuals had creative and ethical concerns surrounding deepfakes in film, specifically the use of deepfakes to recast actors. Using a mixed-methods approach, participant’s memories of actors cast in remade films were tested and their opinions on deepfake technology being used to recast actors in movies were explored. Their study aimed to investigate whether seeing a deepfake video of a fictitious film remake (e.g., Charlize Theron as Captain Marvel instead of Brie Larson) could elicit false memories from participants. Results showed that deepfake exposure was capable of eliciting false memories in participants.

Participants also raised concerns surrounding how the artistic integrity of films would be disrespected and the shared social experience of films would be ruined. Discomfort was also expressed about the level of control and customizability that deepfakes offered. Murphy et al. (2023) encourages further research into understanding deepfakes as a cultural phenomenon rather than viewing deepfakes as a uniquely advancing "alien technology".

3.3.3 Deepfake Resurrection

The ability for deepfake technology to be used to resurrect deceased individuals for entertainment purposes has frequently been discussed in recent years as deepfakes of deceased individuals are being spread through different media (Lees et al., 2021; Lu & Chu, 2023). While reanimating individuals with deepfake technology may seem novel due to its hyper-realism, the digital "resurrection" of deceased public figures is nothing new. There have been many attempts to bring deceased celebrities back to life on the screen such as when the late Elvis Presley sang a duet with Celine Dion on the TV show "American Idol" in 2007 (Ngak, 2012), or when the late rapper Tupac performed with Snoop Dogg during the Coachella festival in 2012 (Ngak, 2012). Sherlock (2013) wrote a paper discussing digital resurrections and has hypothesised that digital resurrection of influential figures serves to contextualise these figures to the modern world, foster para-social and intimate relationships between these figures and the people who admire them and acts as a memorial for the deceased. An emerging aspect of deepfake technology that could potentially appeal to individuals is the ability to offer a glimpse of people in the past, similar to how Elvis or Tupac were portrayed, or de-ageing individuals (Mark Hamill as Luke Skywalker).

A few recent studies have endeavoured to study the application of deepfakes in digital resurrections. For example, Lees et al., (2021) aimed to report the creative, technological, and legal ramifications of the "Virtual Maggie" project where the deceased British prime minister

Margaret Thatcher was recreated on the digital screen to play herself in a period drama. They report insights from the process of the Virtual Maggie project and describe how deepfake technology has complicated the direction of screen performances as it adds another layer of consideration for filmmakers artistically. These challenges are further exacerbated by the deepfakes not being able to portray historical figures realistically due to the lack of high-quality imagery. Furthermore, this report highlights legal issues surrounding the reputational damage that deepfake resurrections can cause, as well as the potential defamation of deceased individuals. However, Lees et al. (2021) did not report or investigate on how deepfakes were experienced by the participants.

Ethical concerns are frequently raised in digital resurrection literature. Digital resurrections are described in literature as akin to controlling their deceased image like puppets (Pantserev, 2020) and ultimately disrespects the dignity of the deceased (Bode, 2007). Even when digital resurrections through deepfakes were used pro-socially to raise awareness for a mass shooting in America by depicting a deceased victim advocating for gun regulation (De Ruiter, 2021), some people still expressed discomfort to seeing the deceased victim being manipulated (Lu & Chu, 2023). While there is no empirical research on digital resurrections facilitated through deepfakes and how people experience them, digital resurrection literature in general highlights an overall discomfort surrounding the ethical implications of digitally recreated individuals.

It is evident that a proper exploration of perceptions and experiences surrounding deepfake technology is necessary to understand how individuals are affected by deepfake technology. The academic community's current perception of deepfake technology is one categorised by hypothetical dystopian outcomes and speculative harmful effects of deepfake exposure. Empirical evidence to substantiate these claims is yet to surface, and therefore

accurate predictions about the effects and experience of seeing deepfakes cannot be made. A naturalistic exploration of the experience of deepfakes is necessary to highlight how an individual encounters deepfake technology in their everyday lives. The proliferation of deepfake content in entertainment media offers us a perfect environment to examine individual responses and reactions to deepfakes in a naturalistic way. Despite this proliferation, it is yet unclear how individuals experience deepfake content in entertainment settings. Therefore, our research question is: "how did YouTube commenters discuss the experience of encountering deepfakes in the music video by Kendrick Lamar".

3.4 Methods

3.4.1 Music Video

The music video in which this study thematically analysed comments from is called "The Heart Part 5" by artist Kendrick Lamar. The video was uploaded to YouTube on the 9th May 2022 and can be found here: https://www.youtube.com/watch?v=uAPUkgeiFVY&ab_channel=KendrickLamarVEVO. The video features Kendrick Lamar rapping and simultaneously changing his face into various famous men including OJ Simpson, Will Smith, late Kobe Bryant, Kanye West, late Nispey Hussle, and Jussie Smollett. From this list of men, some are famous actors, singers, and sports players and they are all influential or controversial in their own way. In the music video, the lyrics of the song coincide with each of the relevant deepfaked faces. For example, when the face of the recently deceased rapper Nipsey Hussle replaced Kendrick Lamar's, the lyrics "To my brother, to my kids, I'm in heaven" were spoken. This pattern follows throughout the music video. The primary reason for choosing this music video as a focus for our study was the number of views and attention it was getting upon release. Garnering several million views on

release up to 44 million views on the date of writing this paper (Lamar, 2022), this music video was one of the first pieces of entertainment media to incorporate deepfake technology and reach millions of views and achieve global attention (Hogan, 2022; Pantserev, 2020; Squires, 2022).

3.4.2 Data Extraction

Data was collected by extracting comments from the "The Heart Part 5" music video using YouTube API. All comments from the music video were extracted on the 4th August 2022, totalling 66769 comments. Replies to these comments were excluded due to the context-based nature of comment replies and therefore would be difficult to interpret and use for the analysis. We aimed to collect data from the music video which referred to the visual deepfake aspect as it is the phenomena of concern. We aimed to extract and analyse approximately 1500 comments from the 66769 total comments. This approximate number of comments was found by referring to existing studies that used YouTube comments for a thematic analysis (Ferrey et al., 2023; King & McCashin, 2022; McLellan et al., 2022). No studies existed that investigated deepfakes or misinformation using thematic analysis through YouTube comments and therefore the aforementioned studies were used as an approximate guideline as they produced a sufficiently in-depth thematic analysis from their comments. These studies analysed between 1000-1500 comments and therefore we similarly aimed to retain this number of comments for our analysis. We also considered this number of comments as appropriate for time and resource constraint reasons.

3.4.3 Inclusion Criteria

Due to many comments potentially not being useful to the analysis of people's reactions to deepfakes e.g., commenting an emoji or commenting something unrelated to the video or deepfakes, we implemented an inclusion/exclusion criterion to identify relevant comments. We

decided to include comments that referred to: the visual component of the music video (e.g., *"This video is lit how his face changes is so sick"*), to deepfakes or video-editing technology (e.g., *"Yoo that deepfake is crazy"*), to any of the men included as deepfakes in the video (e.g., *"Lol "Juicy Smoo-yay"* making an appearance"), and to any form of deception or creativity (e.g., *"He's basically showing you how they use technology to fool the fuck out you"*). We excluded comments that only referred to the song and/or only to Kendrick Lamar and not to the music video in particular such as *"This song is fire"* or *"Kendrick you're amazing"*. These criteria isolate relevant comments for analysis.

3.4.4 Data Selection Process

Our aim was to reach 1500 relevant comments for our analysis. We wanted the 1500 comments to also be representative of the overall 66769 and thus several steps needed to be taken to ensure a relevant and representative sample.

1. The approximate ratio of relevant to irrelevant comments was first established by manually applying the inclusion/exclusion criteria above to the first 4000 comments in the dataset. We found that approximately 1 in 6 comments being made were relevant for our analysis.
2. On this basis, to reach our target of 1500 relevant comments, we needed a 9000 (1500 x 6) representative sample to which when we apply our inclusion/exclusion criteria, we would reach our target relevant sample
3. To generate 9000 comments representative of the 66769 total number of comments, we extracted every 7th comment from the 66769 total ($66769 \div 9000 \approx 7$). This process of extracting comments was considered better than simply extracting the first 9000 comments as we would be able to collect comments from throughout the period that the

music video was available for viewing and commentary since its release to the date of data extraction. We carried out this process in order to avoid potential bias from commenters who are committed fans of Kendrick Lamar that would presumably dominate the comments section in the first few days of the music video being released.

4. Finally, we manually applied our inclusion/exclusion criteria to the 9000 comments extracted as in point 3 to arrive at our final relevant comments sample of 1467.

3.4.5 Analysis

We used reflexive thematic analysis to analyse the comments from the music video. A reflexive thematic analysis was considered appropriate due to its exploratory and flexible nature (Braun & Clarke, 2006). The exploratory aspect of thematic analysis is especially suitable for analysing this creative application of deepfakes due to the scarcity in creative deepfake research and lack of strong theoretical work on the topic. As well as this, YouTube comments tend to be short and fragmented (Schultes et al., 2013) and an in-depth analysis of the comments would be difficult and inappropriate, therefore a thematic analysis was considered suitable for this data. The comments were transferred to an excel file and we followed the thematic analysis process in Clarke et al. (2015). Firstly, we familiarized ourselves with all the comments through several readings. At this stage, we made notes and our own comments throughout the data to gain a surface-level understanding of the commenters reactions to the music video. We then generated initial codes for each comment. This second stage involves labelling each comment with a shorthand word or phrase which captures the idea of the comment itself. The comments yielded both semantic (surface-level meaning of the comment) and latent (the meaning behind the comment) codes. These codes then underwent several iterations between the primary and secondary investigator before proceeding to the third stage in which we began collating the codes together to form meaningful themes that

answer our research question. In the fourth stage, these themes were iteratively discussed and developed to ensure that a coherent narrative was developed from the themes. In the fifth stage, these themes were then defined, named, and finalized for the write-up process. The final stage is the writing of the results in line with the principles of Braun and Clarke (2006).

Guidance from Clarke et al. (2015) was considered and implemented throughout the analysis process. To explore the experiences and reactions people had to seeing deepfakes in the Kendrick Lamar video, we decided to code information regarding the commenter's direct experiences, rather than code facts about deepfakes themselves. This epistemological difference spawns from the lack of theory surrounding deepfakes in entertainment to inform our interpretations of the comments, therefore we needed to focus on a descriptive and essentialist approach to our reflexive thematic analysis which emphasises the experiences, meanings, and reality of the commenters (Clarke et al., 2015).

Secondly, the reflexivity of the research team was taken into serious consideration, with reflexivity being the acknowledgement of our roles as researchers in the analysis. When analysing each comment, it was important for us as researchers to consider how our assumptions and prior experiences could shape and influence the research process. This process is crucial for the transparency and quality of the analysed data (Braun & Clarke, 2006). We had to consider our role as researchers in the analysis compared to the people who were commenting on the video. Commenters often had different vernacular to the primary investigator, and they often used emojis throughout their comments leading to a certain level of interpretation required when reading and interpreting the comments. The primary investigator is also an Irish national with Chinese ethnicity who is unfamiliar with the culture from which the music video derives as well as the slang and vernacular used in American rap music. Collectively, the thematic analysis was conducted with a relatively unfamiliar perspective on rap culture. Furthermore, the primary investigator analysed responses from

commenters from the perspective of a deepfake researcher, which we considered beneficial, as it allowed us to apply our knowledge of deepfake literature to better understand commenter's reactions. A secondary investigator was also involved in the coding process. While Braun and Clarke (2006) suggest that it is unnecessary to include a secondary investigator, this was done to ensure that the unfamiliarity with the vernacular and culture of the music video did not confound the meaning of the codes, rather than to ensure that the codes were "reliable". Inter-rater reliability for the codes was not assessed because the meaning and data generated from the comments are contextual and situational in reflexive thematic analysis, therefore researcher subjectivity is a vital component in interacting with the comments to produce insightful data rather than a factor to be controlled for (Braun & Clarke, 2013).

3.5 Results

Responses to the music video were often short and reactionary, but some comments included more depth and analysis. Six main themes were developed from the data: Advancing realism of deepfakes, Seeing is not believing, Predicting dystopian threats of deepfakes, Pairing deepfakes with lyrics enhances the experience of the song, Deepfakes eliciting strong emotional reactions, and Treating deepfakes as a toy for play. These six themes give us an insight into people's reactions to different aspects of encountering entertainment deepfakes.

3.5.1 Advancing Realism of Deepfakes

In our dataset, people frequently commented that video-editing technology has continued to advance in terms of the realism of content produced. Many commenters were impressed by how realistic the deepfakes in the music video were, often highlighting the realism of particular celebrity faces used in the music video, for example: *"The will smith face*

looked too real" or *"You can't tell me that wasn't Ye"*. This sentiment was taken further with some commenters comparing the deepfake face swapping to science fiction ideas such as shapeshifting or cloning due to how realistic the deepfake faces seemed to be *"Shape shifting reptilian confirmed"* or *"He looks like Kanye and kid cudi, he's definitely cloned"*. These responses reflect a perception of hyper-realism surrounding deepfakes and reinforce literature decrying the harms of deepfake technology for their ability to realistically defame and misrepresent information in video format (Albahar & Almalki, 2019; Westerlund, 2019).

While some commenters acknowledged the realism of deepfakes, other commenters responded by comparing them to existing examples of realistic video editing of faces in entertainment. For example, commenters compared "The Heart Part 5" to the ending of Michael Jackson's music video "Black or White" where multiple different people morph and transition into one another: *"This is amazing. I remember Michael Jackson was the first to use face morphing tech on Black Or White years ago. Deepfake is a bit different but Lamar made it work!"*. Commenters demonstrated a familiarity with video-editing technology and an acknowledgement of the advancements that realistic video-editing technology has made. These responses may reflect a perception of evolution of video-editing technology from more primitive video-editing techniques such as "morphing" in Michael Jackson's "Black or White" music video in 1991 (Chin, 2011) and "rotoscoping" in Celine Dion's duet with Elvis on "American Idol" in 2007 (Boudreault-Fournier, 2016) to using AI neural network technology to superimpose precise faces onto one another. This evolution situates deepfake technology as the next step in video-editing technology, rather than an independent advancement.

Despite the current perception of deepfakes and their advancing level of realism, commenters highlighted aspects of deepfakes from "The Heart Part 5" which could be improved. The disparity in lip-syncing between the lips of the deepfake faces and the lyrics spoken throughout the music video was noticed and criticised by commenters: *"Lip-syncing*

being off was pretty noticeable and got progressively worse"/ *"His mouth animation and the song is slightly off."*. Pointing out these flaws in the deepfakes seems to highlight the importance of aural and visual accuracy in deepfakes. While research is scarce regarding the significance of accurate aural and visual components in deepfakes, Stein and Ohler (2018) have found that distorting facial and aural features of embodied agents (virtual or robotic representations of digital systems) leads to reduced levels of credibility and attractiveness towards that agent. Similarly, Kaate et al. (2023) discuss how glitches, distortions, and vocal issues can reduce the overall "realness" and "human-likeness" of a deepfake possibly affecting immersion. It is possible that aural and visual inaccuracies break the immersion of the technology and thus viewers are quick to respond with criticism.

3.5.2 Seeing is not Believing

Many viewers seemed unaware of the existence of deepfake technology and its capabilities. When these viewers subsequently saw the deepfakes in the music video, their reactions and responses offered interesting insights into how individuals with little to no knowledge of deepfakes interact with this technology. Firstly, most responses were characterised by confusion, doubt, and uncertainty. Commenters were confused by the face morphing and responded by doubting themselves and what they were seeing: *"Is it me or does he look like will smith for a sec", "am I the only one that sees Kanye here?"*, or *"Why does his face at some point look like will Smith's?"*. These responses are indicative of the commenters not knowing they were seeing a deepfake as they would not have been confused by the appearance of the faces if they were aware that deepfakes were possible. It is important to note here that commenters are still questioning what they are seeing as they understand that something unusual is happening here with the faces in the music video.

Interestingly, some commenters responded with confusion but also with a potential way to explain the visuals. For example: *"Damn son. How did he get all those cameos though???"*. This response is particularly interesting because the commenter seems to be attempting to comprehend a technology that they are unfamiliar with by framing and explaining the visuals through familiar means (i.e., a cameo). This response is reminiscent of an everyday psychology concept known as "familiarization" where the everyday individual often attempts to anchor unknown and unaccustomed concepts and experiences in familiar terms to better understand them. Individuals in our dataset are attempting to ground these experiences in their own reality (Emiliani & Passini, 2016). Similarly, responses from other commenters reflected this concept of familiarization responding with: *"Kobe Face at 3:58 or am I trippin?"*, *"Am I high or Kendrick has different faces in this video. Including Kanye, Will smith, Nipsey, etc"*, or *"Was his face changing to different people or am I too high"*. Commenters seem to be confused by the deepfakes to the point where they could only explain the experience by likening it to taking psychedelic substances because they were unaware of the capabilities of deepfakes.

Finally, some commenters did not respond with any form of confusion. These commenters simply asserted that they are seeing some faces and similarities between Kendrick Lamar and the deepfake faces in the video: *"Kendrick looks like kanye west a bit in this video. Lol"*, *"Kendrick looks like will smith here when he's looking dead at the camera lol"*, or *"He kinda of looks like nipsey"*. These responses seem to indicate that the commenters don't see anything unusual about the deepfakes and the stark face changes are dismissed as Kendrick Lamar simply looking alike to them. The fact that the faces are changing seem to go completely unnoticed with this subgroup of commenters as they are not even questioning the visuals and accepting them as a coincidence that Kendrick Lamar and the deepfake faces are similar. This is potentially concerning as it seems to highlight that some individuals who are unaware of the existence of deepfake technology may not question the deepfake visuals they see.

The lack of awareness of the existence of this technology and its capabilities can be read as a lack of digital literacy among some of the commenters. Despite the fact that the deepfakes in the music video were not meant to deceive or misinform but to entertain, they still caused a significant amount of confusion, doubt, and uncertainty in some commenters. This can potentially be concerning as a recent study on the effects of deepfake exposure on trust in media and news showed that seeing deepfakes may not necessarily misinform individuals, but they help increase levels of doubt and cynicism towards media and news sources (Vaccari & Chadwick, 2020). Finally, and perhaps most worryingly, some individuals were completely unaware of the deepfakes in the music video. The deepfakes going completely unnoticed by some commenters is concerning because it could potentially suggest that some individuals that encounter deepfakes in their everyday lives may see them as normal videos without question, and thus may be more likely to believe in the veracity of the video.

3.5.3 Predicting Dystopian Threats of Deepfakes

Commenters often expressed their concerns over the current and potential harms of deepfakes. They often depicted a dystopian future where deepfakes have eroded the public's trust in news and media: *"This shit is eerie I don't like this Deepfake technology it can make lies appear real to the masses when used in the media"*. These fears and speculations were also associated with negative predictions concerning the future of technology: *"This technology kinda scary though if used for the wrong bs like framing folks etc"* and the potential malicious actors using this technology: *"Just imagine ; how many times the government has used this facial features to confuse mankind?"*. These responses towards deepfakes are akin to the speculations found in deepfake literature decrying the potential of deepfakes to cause epistemic harm (Fallis, 2020) and to defame individuals (Chesney & Citron, 2019).

A number of commenters considered deepfake technology harmful enough to require awareness to be raised around this technology. Commenters considered how the music video warns people of the simplicity surrounding the use of this technology *"CGi technology.. crazy thing is, he bringing awareness of how easy it is to manipulate a face now, thanks to all these cameras."*. Commenters also speculated that deepfakes could affect the entertainment industry by replacing actors with the faces of anybody we want to with deepfakes in the future: *"but man these effects make me wonder... do we need actors and actresses anymore. within a few years we can just create what we want. so weird."*. These responses are similar to a study conducted by Murphy et al. (2023) exploring how people feel about potentially using deepfakes of preferred actors to replace the faces of actors in movies. Similar to the majority of existing deepfake literature and discussion, these responses and concerns reflect an overall focus on the hypothetical dystopian harms of deepfakes rather than the realistic and modern capabilities of deepfake technology (Chesney & Citron, 2019; Westerlund, 2019).

3.5.4 Pairing Deepfakes with Lyrics Enhances the Experience of the Song

A number of commenters mention how deepfakes have enhanced their understanding and enjoyment of the song: *"Lol I've been listening but it's the first I watched the video now I see the deep fake and I understand the song better now"*. Specifically, one aspect of the music video that commenters responded positively to was how the lyrics of the song coincide with relevant deepfake faces. This pairing between lyrics and deepfakes were often described as creative: *"Yes the way he changed to Will smith and said hurt people hurt people too!! That's real! Such a creative video and unique no doubt!"* and *"When he said When hurt people hurt people and switched to Will..... I mean damn. Only a real genius can sum up that entire situation and so much more with just 5 words. Absolute Legend"*. Some commenters perceived

this pairing as enhancing the message behind the song: *"Did y'all catch when he said in a world where hurt ppl hurt ppl..... And it was will Smith face.. .... Message"*. Commenters were extremely engaged in identifying the meaning that was generated through this pairing by asking other commenters about the meaning behind the faces, further speaking to the level of meaning produced by combining lyrics with the deepfakes: *"I understood the NIP transition, but what did the other face transitions represent?"*.

Commenters often reported that the pairing of lyrics and faces elicited strong emotional responses: *"Look what I've done for you when he's Will Smith and says those words, I get chills"* and *"I got chills watching this, especially when the morphing correlated with the words he spit… genius!"*. This seems to speak to the emotional and creative impact that this pairing generated. Some commenters also found this pairing to be humorous as well when referring to certain celebrities *"Dat will Smith part look what I done for u Jada punching the air rn lol"*. These responses highlight the ability of deepfakes to artistically enhance the experience of the song by eliciting strong emotional reactions. This is similar to how deepfakes have been suggested to enhance the experience of other artistic expressions such as the resurrection of Salvador Dali through deepfakes in the Dali Museum (Mihailova, 2021). It has been posited that deepfakes invoke a strong emotional involvement between the observer and the art, promote empathy for the represented individual, as well as humanize them (Mihailova, 2021). Similarly, deepfakes in the music video serve to invoke empathy in viewers, as well as produce a more vivid emotional experience for the viewer.

3.5.5 Deepfakes Eliciting Strong Emotional Reactions

Closely related to the theme above, merely seeing the deepfakes and the identities that they represent elicited strong emotional reactions in the commenters ranging from sadness: *"The song and video is so powerful and refreshing. It made me cry the first time I watched it"*

to admiration *"he is on another level, man this visual is moving bless up Kendrick"* and laughter *"I thought the face after Will Smith's would be Chris Rock's"*. This wide range of emotional responses highlights how powerful deepfakes and the ability to represent identities can be.

Commenters also responded with feelings of sympathy: *"will Smith one hit different damn I feel for him"*, but especially for the deepfake faces of the people who were deceased. Some commenters exhibited concern for the family members of the deceased men and how they may feel: *"Imagine how nipsey's family feels while watching this. Damn that would be tough to watch and hear"*. These responses reflect the sentiments of Mihailova (2021) positing that deepfakes can be used to humanize and invoke a sense of empathy in viewers.

Interestingly, many commenters viewed deepfakes of the deceased men as Kendrick embodying the men and "speaking" through them: *"WTF just happened?? That last verse nipseys spirit really spoke to Kendrick homie that shit was phenomenal"*. To some commenters, there was a spiritual element to Kendrick using deepfakes to recreate the deceased men: *"the visual transformation from kobe to nipsey tho. (...) dope way to channel their spirits. now THIS is homage"*. Many commenters viewed Kendrick using deepfakes of the deceased men as a means to pay tribute and memorialize them: *"Thank you Kendrick for paying tribute to our brother nipsey hussle. He talked to us for the last time now"* and *"Thank you for keeping Nipsey's memory alive Kendrick"*. It is clear that some commenters were expressing a large amount of sadness from seeing deepfakes of the deceased men as well: *"Man that Kobe and Nipsey part had my chest hurting"*. Some commenters were perhaps exhibiting signs of parasocial grief, a phenomenon in which people grieve for deceased famous people whom they have never met such as celebrities, sports players, or actors (Bingaman, 2022).

However, some commenters were offended by the use of the deceased men in the music video, citing that their use was simply *"begging for views"* and that it was *"disrespectful"* for

Kendrick to use them in this manner. It was even viewed as Kendrick not *"letting"* the deceased rest by digitally resurrecting them. These responses reflect a similar sentiment regarding the ethical issues of digital resurrection. Sargeant (2017) echoes the responses of the viewers by positing that resurrection of a deceased person through technology can be seen as dishonouring of the dead, almost like puppeteering a deceased body. Bode (2007) also highlights how digital resurrection can be seen as "exploitative" and presents moral dilemmas regarding consent as resurrected public figures can no longer consent to their image being used or how it is portrayed.

3.5.6 Treating Deepfakes as a Toy for Play

Many commenters derived enjoyment from simply trying to identify all the faces in the music video: *"I got Kanye, Will Smith and snoop (?) who else?"* or *"Can someone please tell me all the faces seen in the video? I recognize some of them, but not all and it's really bugging me"*. This form of enjoyment is born from the inherent attributes of deepfakes, which is to swap the identities of different people and seems to highlight how most people are often drawn to the dystopian and hypothetical harms of emerging technology and subsequently neglect ameliorative qualities that the technology could have (Broinowski, 2022). While the hypothetical harms of deepfakes focus on the total erosion of truth in society (Chesney & Citron, 2019), it is evident that many people simply enjoy deepfakes as a form of day-to-day entertainment.

Many commenters seemed to take a playful and participatory approach to the deepfakes, debating and questioning the choice of faces: *"This track is amazing but juicy smoolyay and Will Smith faces should not have been anywhere close to this video. It sadly takes away from how great this.message was. They are both shallow and shameful examples of the culture"*. Other commenters expressed how they would have preferred if certain other faces appeared in the video: *"I would have liked to see George Floyd @ 3:35; after those deep breaths !"*. By

responding with names of different public figures, commenters seemed to be embracing the choice and freedom offered by deepfake technology, seemingly viewing the possibilities as infinite. While some research has suggested that the huge amount of choice offered by deepfakes in entertainment can be overwhelming and uncomfortable, such as the case of personalized deepfake recasting of films (Murphy et al., 2023), in this case it was viewed as a positive.

Commenters were also entertained by how deepfakes could represent a person performing an action or saying words that they wouldn't normally do or say *"Will Smith is not happy with you Kendrick. Will Smith don't have to cuss in his raps to sell record. Now Kendrick made him do it"*. Commenters seemed to enjoy the level of manipulation that deepfakes can provide, from choosing who they want to see as deepfakes, to relishing the deepfake identities performing actions or speaking words that they normally wouldn't do or say. This level of control is reminiscent of discourse around deepfake pornography, where consumers create "designer porn" that fits their preference, without the consent or awareness of those featured (Kikerpill, 2020). While the enjoyment of deepfake faces in music videos is, on its face, more playful and less obviously harmful than non-consensual deepfake pornography, the issues of control and consent are still central. Indeed, it has been said that deepfakes are inherently morally problematic as the use of our images and audio to generate deepfakes is often done without the consent of the individual (Deng et al., 2024), regardless of whether the deepfakes generated are for entertainment or pornographic purposes. The control over an individual's representation that deepfakes offer also infringes on the privacy rights of individuals as deepfakes represent people's likeness in situations in which they did not consent to being in (Ramazotti, 2021). It can be argued that what makes deepfakes so potentially harmful is also what can make them so appealing.

Not all commenters enjoyed seeing deepfakes in the music video, with some considering them as cheap way to garner publicity: *"The song really good but the different celebrity faces he puts throughout the video are stupid af. Guess he wanted the vid to get more views cause it's a dumb decision"*. Some commenters viewed the deepfakes: "overrated". These commenters seemed to view deepfakes as an unnecessary gimmick that does not enhance the artistic value of the video.

3.6 Discussion

To reiterate, our research question was "how did YouTube commenters discuss the experience of encountering deepfakes in the music video by Kendrick Lamar?". Our results yielded a multitude of answers to this question which illustrate the complex everyday experiences of and responses to seeing deepfakes in an entertainment context. Firstly, the majority of the responses were positive in nature, voicing support and praise for the deepfakes included in the music video. The responses often commended the creativity surrounding how the deepfakes were utilized and the emotional impact that seeing the deepfakes had on the commenter. Commenters actively engaged with the deepfakes in a playful manner by attempting to guess who the faces were and theorizing about the meaning behind the inclusion of certain faces in the video. Secondly, a subset of commenters from our results discussed the deepfakes from a technical perspective by critiquing the quality of the deepfakes in the videos and comparing deepfake technology to pre-existing examples of video-editing technology, thus highlighting a level of awareness and knowledgeability surrounding deepfake technology and its progression. Thirdly, some individuals responded to seeing deepfakes by discussing the potential harms and threats that this technology could potentially bring, often predicting potential dystopian outcomes such as a complete collapse of epistemic truth in society.

Evidently, responses to and experiences of seeing deepfakes in the music video by Kendrick Lamar vary considerably in content and valence.

The above findings only show how individuals who are aware of deepfake technology would respond. Individuals in society often have varying levels of digital literacy and may not be aware of the technological capabilities of deepfakes. Our results also show the experiences of individuals who are exposed to deepfakes in the music video while simultaneously being unaware to the existence of this technology. These experiences are often characterized by confusion and self-doubt when exposed to the technology, questioning their own senses and speculating on how the deepfake component of the video was created. Some commenters did not even notice that the faces were changing and dismissed the deepfakes as being coincidentally alike to Kendrick Lamar's face. These findings are highlighting how individuals in society may respond to deepfakes if they are unaware of the technology: by doubting their own senses, attempting to explain the visual confusion in familiar terms, or simply not noticing the deepfakes. Our findings not only highlight whether individuals are in support of using deepfakes in parts of the entertainment they consume, but they also provide a nuanced view of what aspects of deepfake technology is appealing to individuals, what individuals are concerned about surrounding this technology, and how individuals who are unaware to the existence of this technology would respond to them.

3.6.1 Deepfake Awareness

In our study, we found that commenters responded to and experienced deepfakes very differently depending on their level of awareness of deepfake technology. We found an interesting contrast between individuals who were considerably aware of deepfake technology, and those who were uncertain and confused by it. In our first theme, certain commenters compared deepfakes to existing video-editing examples and even began praising and criticising

aspects of the technology. These responses highlight a familiarity with the technology in the casual manner in which the technology was discussed. Commenters are demonstrating that they know this technology exists and the individuals that appeared in the music video were digitally created. This theme also emphasises how some commenters who are aware of deepfake technology may not see it as something independently novel, but as an evolution of pre-existing visual-editing technology when they mention how deepfakes are similar to visual effects in entertainment media in previous years.

Those who did not explicitly demonstrate awareness of deepfake technology tended to discuss the video somewhat differently. Responses were often characterized by confusion, doubt, and uncertainty. Rather than clearly identifying the video as manipulated, commenters often appeared to question their own perception or relied on familiarisation strategies (Schraube & Højholt, 2015) to make sense of what they were seeing. Surprisingly, some commenters did not even notice the deepfakes and simply stated how similar Kendrick Lamar looked to the celebrities that were being deepfaked. Notably, we did not identify any comments within this group that explicitly suggested the visuals were fake or edited. This finding, or lack of a finding, seemed to be an underlying thread throughout these responses: they assumed that what they were seeing was truthful to some extent. Instead of suspecting the visuals were edited, individuals responded with questioning their own senses, explanations through familiar concepts, or assumed the faces coincidentally looked like Kendrick Lamar instead of the veracity of the video. This seems to suggest that individuals who do not know that deepfake technology exists assume that what they are seeing is the truth. Our result may potentially be explained by Truth-default theory (TDT) which supposes that individuals naturally assume that information they receive is truthful (Levine, 2014). TDT posits that humans naturally presume other individuals are honest due to lack of information suggesting otherwise as well as dismissing the possibility of deception unless signs of deception were detected. Commenters

in our study who were unaware of deepfake technology may have assumed that there was no reason to worry about being deceived by a music video from Kendrick Lamar. This theory and subsequent supporting research are typically based off individuals receiving textual and verbal information/misinformation from an individual and our study may indicate that this effect could be similar with visual information.

This finding contributes to broader deepfake literature by highlighting how individuals react differently to deepfakes depending on their level of awareness to the technology in their day-to-day lives. Commenters who were not explicitly aware the deepfakes often reported confusion and doubt over their senses, as well as potentially defaulting to a truth bias towards the deepfakes. The implication of this finding for the broader HCI community is that there is a need to mitigate feelings of doubt and uncertainty, as well as inform digitally illiterate members of the public about deepfakes as they may be especially vulnerable to misinformation if it is presented as a realistic deepfake. This may need to come in the form of increased media literacy education. Understanding deepfake technology and being aware of its existence may serve as a way to protect individuals from being influenced by deepfakes (Hwang et al., 2021). This responsibility may need to fall on social media platforms to have warnings or informative text about deepfake technology before viewing (which have been shown to be effective at dispelling misinformation (Magee et al., 2017) or host educational programmes among their users to improve their literacy skills, which Google and Facebook have done (Lee, 2019). However, it is also important to raise awareness of deepfake technology without undermining individual's trust in legitimate media (Twomey et al., 2023). Therefore, a balance must be achieved with informing the public without overexaggerating the impact of deepfakes.

3.6.2 Implications for Deepfakes in Entertainment

Deepfakes also seem to be discussed and engaged with in a playful manner when encountered in an everyday environment. Our findings suggest that a large subset of commenters actively enjoyed engaging with deepfake technology, often suggesting faces they would like to see, attempting to identify all the different faces in the video, and attempting to make sense of the meaning behind including each face. These responses are indicative of the recreational aspect of deepfake technology I.e., allowing viewers to see their desired idols doing or saying anything. Ultimately, the aforementioned responses reflect how individuals respond to deepfakes on a casual and naturalistic level. These responses are to be expected when interpreted from the point of view of viewing a music video from an artist that the commenter presumably is favourable to. This study is the first empirical study to show that in a naturalistic entertainment environment, many individuals enjoy viewing deepfakes.

Our study has discovered some aspects of deepfakes which may enhance enjoyment of media for individuals. We have found that deepfakes paired with lyrics and words can enhance the underlying meaning of the message. Practical examples of this can be seen with the resurrection of Salvador Dali in the Dali Museum (Mihailova, 2021) or with Buzzfeed's video "You Won't Believe What Obama Says In This Video!" (Buzzfeed, 2018). Deepfakes can also increase immersion with the music video, as well as enhance its emotional impact. This sentiment is reflected in literature describing how visual stimuli can impact how the music is perceived (Ebendorf, 2007) and can enhance the overall reception of the music (Boltz, 2009; Boltz, 2013). However, our results also suggest that when deepfakes are produced poorly (I.e., lyrics not synchronised with lip movements or digital artifacts on the deepfake faces), viewers are quick to criticise the technical inaccuracies leading to potentially reduced enjoyment. This may possibly be attributed to a break in immersion as the illusion of seeing and hearing the deepfaked individual is broken. The findings provide evidence for the recreational aspect of

deepfakes and improves our understanding of how aspects of deepfake technology could enhance enjoyment for individuals.

There are several implications of our findings which can inform artists or individuals in the entertainment industry on how deepfakes can be incorporated and used in their media. Our findings suggest that deepfakes can help carry and amplify messages when conveyed in conjunction with the deepfake. This can be utilised in the edutainment industry (educational content which is entertaining) to promote awareness of deepfake technology or promote other educational content in an engaging and playful manner. There is potentially an avenue for designers to use this technology and lessons learned in this study to reach communities with a more impactful version of a message. The high customisability of deepfakes may also add to the reach this technology has. Some designers in the marketing industry have leveraged similar ideas to design advertisements with a multinational and impactful reach (Oakes, 2019). As well as this, our findings surrounding individual's enjoyment of the embodiment of individuals using deepfake technology can be harnessed in artistic industries such as museums. This is because deepfakes can portray historical artifacts or people in a realistic and life-like way that no other media can emulate. Indeed, this can already be seen in the Dali Museum (Mihailova, 2021) or in the Gipsoteca Museum (Zaramella et al., 2023), but the entertainment and educational value of these applications have yet to be empirically tested.

3.6.3 Exaggerated Deepfake Fears Overshadowing Current Threats

The general sentiment surrounding the threat of deepfake technology from the YouTube comments was that deepfakes are overwhelmingly deceptive to the point where they can perfectly emulate any individual to generate and spread copious amounts of misinformation thus eroding trust in governmental bodies and news sources in society. While this train of

thought is not uncommon among literature (Ajder et al., 2019; Albahar & Almalki, 2019; Chesney & Citron, 2019; Sylvester, 2021), it can potentially overshadow the existing issues of deepfake technology and detract from efforts to combat their current harms. Many topical issues surrounding deepfakes which actively harm individuals in society remain unresolved and under addressed. For example, there has been an increase in financial crimes and fraud using deepfake technology. Malicious actors have been shown to use deepfake technology to emulate important financial figures such as the company director of a bank (Brewster, 2023) or emulate friends or family to solicit funds (Bhati, 2023). As well as this, at the time of writing this paper, actors in Hollywood were recently on strike in protest of the potential use of AI and deepfake technology to imitate actor's likeness in movies (Ash, 2023; Watercutter & Bedingfield, 2023). This strike highlights issues surrounding the licencing of one's identity and the continuation of their livelihood in a society where deepfake technology and artificial intelligence could potentially replace individuals in art and entertainment media. Finally, deepfake pornography is another increasingly concerning issue which is often under-represented when discussing issues of deepfakes. The nature of deepfake pornography robs individuals (mostly women) of their agency and privacy, and humanity when they are non-consensually depicted in pornographic videos, which can lead to psychological distress (Gieseke, 2020; Maddocks, 2020). It is understandable that against the backdrop of a music video designed for entertainment, these concerns were not articulated or mentioned. However, amongst those who did attempt to criticise or raise awareness of the concerning development of deepfake technology, very few of these imminently alarming issues were mentioned in the comments we examined, and we hypothesise that this may be due to dystopian predictions of deepfake threats such as the total erosion of truth in society overshadowing existing threats (Albahar & Almalki, 2019; Chesney & Citron, 2019).

Individuals in the YouTube comments who focused on the threatening aspects of deepfakes overlooking existing consequences of deepfakes highlights an overall larger issue with perceptions around this technology. The perception that deepfakes will cause irreparable harm to truth in society and destabilize political structures is unfounded, unrealistic, and harmful to existing efforts to combat real world issues surrounding deepfakes. This perception may be due to a form of alarmism over emerging technologies which is a well-documented human response. Broinowski (2022) has described how different technologies in the past such as telephones and the internet have similarly spurred on discussions surrounding their potential dystopian applications, which ultimately did not come to pass. This suggests that individuals are quick to jump to the worst-case scenario for emerging technologies that we may yet to fully understand. Johnson and Veridicchio (2017) hypothesise that exaggerated fears surrounding AI technologies may be due to several factors, one of which is an unrealistic expectation of the speed of technological advancement necessary to achieve the dystopian outcomes they fear. Individuals underestimate the technological advancement necessary for deepfakes to reach an undetectable and completely deceptive form capable of crippling epistemic trust in all sources of information. This ultimately leads them to fixate on the worst-case scenario. The implication of this is that an abundance of research and literature will be focused on discussing scenarios regarding deepfakes that may ultimately never come to pass, while simultaneously detracting attention from more pressing issues.

3.6.4 Limitations and Future Work

This study also has several limitations. Firstly, results from this study are somewhat restricted to comments from individuals who are largely supporters of Kendrick Lamar and his music as they are engaging with his music video on his YouTube channel. This would evidently bias many of the comments we analysed to produce the results of this study as it may not be

representative of how people who are neutral to Kendrick Lamar would respond to deepfakes in their music video. We attempted to reduce the bias from fans of Kendrick Lamar by sampling comments throughout the timespan of the music video's release as described in the methods section rather than exclusively from the comments on release of the video or from top comments (comments with the highest like to dislike ratio).

Secondly, only investigating people's responses to deepfakes in a YouTube music video limits the generalizability of the results to deepfakes across all types of platforms, entertainment, and art. Even within YouTube, only approximately 12% of YouTube users leave a comment in the videos they watched according to a survey conducted by Schultes et al., (2013). Future work should emphasise expanding the scope of analysis to a broader population to consider perspectives on deepfakes in entertainment from various backgrounds. Taken together, the generalizability of the results may be limited but the aim of thematic analysis is to produce rich information in which this paper has ultimately produced.

While the thematic analysis of YouTube comments was appropriate for this study, YouTube comments are typically short and lacking in nuance and detail (Schultes et al., 2013). Future work should focus on a more in-depth analysis of the reasons behind people's enjoyment of deepfakes in a variety of entertainment media. Future work could also qualitatively analyse reactions from popular YouTube videos to entertainment media which features deepfakes to obtain richer data.

3.7 Conclusion

This goal of this study was to thematically analyse the comments of a YouTube music video which included deepfakes to explore viewer's natural reactions and responses to an instance of deepfake usage in entertainment. In general, we can see that deepfakes can have

potentially positive qualities which can enhance people's experience of art and entertainment. Individuals who are unaware of deepfake technology may be faced with doubt and uncertainty when exposed to deepfakes with potentially a truth bias to the visual effect. Discourse surrounding the dystopian impacts of deepfakes have largely overshadowed the recreational aspects of deepfakes and their current harms, therefore we urge researchers to focus on the realistic outcomes of deepfakes and how they affect society today, rather than tomorrow.

Acknowledgements

This work was supported with the financial support of the Science Foundation Ireland grant 13/RC/2094_2 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero -the Science Foundation Ireland Research Centre for Software (www.lero.ie).

3.8 References

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

Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, (November 2019), 3242-3250.

Ash, J. (2023, November 26). George Clooney supports actor's union strike, says it's an 'inflection point in our industry'. *Fox News*. https://www.foxnews.com/entertainment/george-clooney-supports-actors-union-strike-inflection-point-industry

Archanoth. 2020. Steven Wilson - SELF (Official Video). Reddit, Retrieved September 1, 2023, from https://www.reddit.com/r/progmetal/comments/la3bdk/steven_wilson_self_official_video/?rdt=62352

Bhati, D. (2023, July 17). Kerala man loses Rs 40,000 to AI-based deepfake WhatsApp fraud. *India Today*. https://www.indiatoday.in/technology/news/story/kerala-man-loses-rs-40000-in-ai-based-deepfake-whatsapp-fraud-all-about-the-new-scam-2407555-2023-07-17

Bingaman, J. (2022). "Dude I've never felt this way towards a celebrity death": Parasocial grieving and the collective mourning of Kobe Bryant on Reddit. *OMEGA-Journal of Death and Dying*, 86(2), 364-381. https://doi.org/10.1177/0030222820971531

Blistein, J. (2021, August 28). Steven Wilson becomes Bowie, Biden, Trump, and more in new 'Self' video. *Rolling Stone*. https://www.rollingstone.com/music/music-news/steven-wilson-self-music-video-deep-fake-technology-1121914/

Bødker, S. (2006). When second wave HCI meets third wave challenges. In *Proceedings of the 4th Nordic conference on human-computer interaction: Changing roles* (pp. 1-8).

Boudreault-Fournier, A. (2016). Microtopia in counterpoint: Relational aesthetics and the Echo Project. *Cadernos de Arte e Antropologia*, 5(1), 135-154. https://doi.org/10.4000/cadernosaa.1060

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2), 77-101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. Sage.

Brewster, T. (2021, October 14). Fraudsters cloned company director. *Forbes*. https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/

Bode, L. (2007). 'Grave robbing' or 'career comeback'? On the digital resurrection of dead screen stars. *History of Stardom Reconsidered*, 36-40.

Boltz, M. (2013). Music videos and visual influences on music perception and appreciation: Should you want your MTV? *The Psychology of Music in Multimedia*, 217-234.

Boltz, M., Ebendorf, B., & Field, B. (2009). Audiovisual interactions: The impact of visual information on music perception and memory. *Music Perception*, 27(1), 43-59.

Broinowski, A. (2022). Deepfake nightmares, synthetic dreams: A review of dystopian and utopian discourses around deepfakes, and why the collapse of reality may not be imminent—yet. *Journal of Asia-Pacific Pop Culture*, 7(1), 109-139. https://doi.org/10.5325/jasiapacipopcult.7.1.0109

Buzzfeed. (2018, March 29). "You won't believe what Obama says in this video!" *YouTube*. https://www.youtube.com/watch?v=cQ54GDm1eL0&ab_channel=BuzzFeedVideo

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

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 1753-1819. https://doi.org/10.15779/Z38RV0D15J

Chin, E. (2011). Michael Jackson's panther dance: Double consciousness and the uncanny business of performing while black. *Journal of Popular Music Studies*, 23(1), 58-74.

Clarke, V., Braun, V., & Hayfield, N. (2015). Thematic analysis. In *Qualitative psychology: A practical guide to research methods* (pp. 222-248). Sage.

Cole, S. (2017, July 22). AI-assisted fake porn is here and we're all fucked. *Vice*. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

David, J. (2022, September 12). 'Star Wars' fans divided on whether they want to see deepfake Luke Skywalker again. *WGTC*. https://wegotthiscovered.com/tv/star-wars-fans-divided-on-whether-they-want-to-see-deepfake-luke-skywalker-again/

Deng, Y., Chen, Z., He, C., Lu, Z., & Li, B. (2024). Persuasion or insulting? Unpacking discursive strategies of gender debate in everyday feminism in China. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1-19).

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

Diaz, A.-C. (2020, January 12). Parkland victim Joaquin Oliver comes back to life in heartbreaking plea to voters. *Adage*. https://adage.com/article/advertising/parkland-victim-joaquin-oliver-comes-back-life-heartbreaking-plea-voters/2285166

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, 69-91. https://doi.org/10.1177/1940161220944364

D'Rozario, D. (2016). The market for 'Delebs' (dead celebrities): A revenue analysis. *Journal of Customer Behavior*, 15(4), 395-414.

Ebendorf, B. (2007). The impact of visual stimuli on music perception (Doctoral dissertation).

Emiliani, F., & Passini, S. (2016). Everyday life in social psychology. *Journal for the Theory of Social Behavior*, 47(1), 83-97. https://doi.org/10.1111/jtsb.12109

Fallis, D. (2020). The epistemic threat of deepfakes. *Philosophy & Technology*, 1-21. https://doi.org/10.1007/s13347-020-00419-2

Ferrey, A., Ashworth, G., Cabling, M., Rundblad, G., & Ismail, K. (2023). A thematic analysis of YouTube comments on a television documentary titled 'Diabulimia: The World's

most dangerous eating disorder'. *Diabetic Medicine*, 40(5), e15025. https://doi.org/10.1111/dme.15025

Gamage, D., Chen, J., & Sasahara, K. (2021). The emergence of deepfakes and its societal implications: A systematic review. *Proceedings of the 2021 Truth and Trust Online Conference*, 1(1), 28-39.

Gamage, D., Ghasiya, P., Bonagiri, V., Whiting, M., & Sasahara, K. (2022). Are deepfakes concerning? Analyzing conversations of deepfakes on Reddit and exploring societal implications. In *CHI Conference on Human Factors in Computing Systems* (pp. 1-19). https://doi.org/10.1145/3491102.3517446

Gieseke, A. P. (2020). "The new weapon of choice": Law's current inability to properly address deepfake pornography. *Vand. L. Rev.*, 1479-1514.

Gosse, C., & Burkell, J. (2020). Politics and porn: How news media characterizes problems presented by deepfakes. *Critical Studies in Media Communication*, 37(5), 497-511. https://doi.org/10.1080/15295036.2020.1832697

Hadusek, J. (2021). Steven Wilson transforms into Brad Pitt, David Bowie, Donald Trump, and more in "Self" video: Watch. *Consequence*. Retrieved August 29, 2023, from https://consequence.net/2021/02/steven-wilson-self-music-video/

Hogan, M. (2022). How Kendrick Lamar's "The Heart Part 5" video subverts deepfake technology. *Pitchfork*. Retrieved July 28, 2022, from https://pitchfork.com/thepitch/kendrick-lamar-the-heart-part-5-video-deepfake-technology/

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*, 188-193.

Iamtherealak. (2020). AK – FAMILY TREE (Prod. AK). Retrieved September 24, 2022, from https://www.youtube.com/watch?v=AK_FAMILY_TREE

Kaate, I., Salminen, J., Jung, S.-G., Almerekhi, H., & Jansen, B. J. (2023). How do users perceive deepfake personas? Investigating the deepfake user perception and its implications for human-computer interaction. In *Proceedings of the 15th Biannual Conference of the Italian SIGCHI Chapter* (pp. 1-12).

Jenner, R. (2022). How Kendrick Lamar and Nipsey Hussle's yearslong friendship informed "The Heart Part 5". *GQ*. Retrieved August 29, 2022, from https://www.gq.com/story/kendrick-lamar-nipsey-hussle-the-heart-part-5

Johnson, D. G., & Verdicchio, M. (2017). AI anxiety. *Journal of the Association for Information Science and Technology*, 68(9), 2267-2270. https://doi.org/10.1002/asi.23867

Karasavva, V., & Noorbhai, A. (2021). The real threat of deepfake pornography: A review of Canadian policy. *Cyberpsychology, Behavior, and Social Networking*, 203-209. https://doi.org/10.1089/cyber.2020.0272

Kikerpill, K. (2020). Choose your stars and studs: The rise of deepfake designer porn. *Porn Studies*, 7(4), 352-356. https://doi.org/10.1080/23268743.2020.1765851

King, C., & McCashin, D. (2022). Commenting and connecting: A thematic analysis of responses to YouTube vlogs about borderline personality disorder. *Internet Interventions*, 28, 100540.

Kugler, M., & Pace, C. (2021). Deepfake privacy: Attitudes and regulation. *Northwestern University Law Review*.

Lamar, K. (2022). Kendrick Lamar - The Heart Part 5. Retrieved May 31, 2022, from https://www.youtube.com/watch?v=uAPUkgeiFVY&ab_channel=KendrickLamarVEVO

Lee, D. (2019). Facebook launches Digital Literacy Library to help young people use the internet responsibly. *The Verge*. Retrieved August 2, 2019, from https://www.theverge.com/2018/8/2/17643586/facebook-digital-literacy-library-educators-schools

Lee, J., & Shin, S. Y. (2022). Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news. *Media Psychology*, 1-16. https://doi.org/10.1080/15213269.2021.2007489

Levine, T. R. (2014). Truth-Default Theory (TDT). *Journal of Language and Social Psychology*, 33(4), 378-392. https://doi.org/10.1177/0261927X14535916

Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. (2020). Face X-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 5001-5010).

Lees, D., Bashford-Rogers, T., & Keppel-Palmer, M. (2021). The digital resurrection of Margaret Thatcher: Creative, technological, and legal dilemmas in the use of deepfakes in screen drama. *Convergence*, 27(4), 954-973.

Lu, H., & Chu, H. (2023). Let the dead talk: How deepfake resurrection narratives influence audience response in prosocial contexts. *Computers in Human Behavior*. https://doi.org/10.1016/j.chb.2023.107761

Matern, F., Riess, C., & Stamminger, M. (2019). Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)* (pp. 83-92). IEEE.

Maddocks, S. (2020). 'A deepfake porn plot intended to silence me': Exploring continuities between pornographic and 'political' deepfakes. *Porn Studies*, 7(4), 415-423. https://doi.org/10.1080/23268743.2020.1757499

Magee, R., Agosto, D., & Forte, A. (2017). Four factors that regulate teen technology use in everyday life. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (pp. 511-522).

Martel, C., & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, 101710.

McLellan, A., Schmidt-Waselenchuk, K., Duerksen, K., & Woodin, E. (2022). Talking back to mental health stigma: An exploration of YouTube comments on anti-stigma videos. *Computers in Human Behavior*, 131, 107214. https://doi.org/10.1016/j.chb.2022.107214

Metaphysic.ai. (2023). Deeptomcruise. *TikTok*. Retrieved November 26, 2023, from https://www.tiktok.com/@deeptomcruise?lang=en

Mihailova, M. (2021). To dally with Dalí: Deepfake (inter)faces in the art museum. *Convergence*, 27(4), 882-898. https://doi.org/10.1177/13548565211029401

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

Ngak, C. (2012). Tupac Coachella hologram: Behind the technology. Retrieved June 23, 2023, from https://www.cbsnews.com/news/tupac-coachella-hologram-behind-the-technology/

Qian, Y., Yin, G., Sheng, L., Chen, Z., & Shao, J. (2020). Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *European Conference on Computer Vision* (pp. 86–103). Springer.

Oakes, O. (2019). 'Deepfake' voice tech used for good in David Beckham malaria campaign. *Campaign.* Retrieved September 21, 2024, from

https://www.campaignlive.co.uk/article/deepfake-voice-tech-used-good-david-beckham-malaria-campaign/1581378

Qobrtay, A. (2022). 'Star Wars' fans love CGI Luke Skywalker, but deepfake implications are dangerous. *The Daily Targum.* Retrieved August 24, 2023, from https://dailytargum.com/article/2022/03/star-wars-fans-love-cgi-luke-skywalker-but-deepfake-implications-are

Pantserev, K. (2020). The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. In *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity* (pp. 37–55). https://doi.org/10.1007/978-3-030-35746-7_3

Pete, P. (2022). All of the deepfakes in Kendrick Lamar's "The Heart Part 5" video. *Hiphopcanada.* Retrieved September 19, 2023, from https://hiphopcanada.com/kendrick-lamar-heart-part-5-deepfakes/

Ramazotti, G. (2021). When Elvis Presley joined Celine Dion on stage for a mind-blowing duet 30 years after his death (Yes, really). *Smoothradio.* Retrieved November 22, 2022, from https://www.smoothradio.com/artists/elvis-presley/celine-dion-duet-hologram-american-idol-2007/

Reid, S. (2021). The deepfake dilemma: Reconciling privacy and First Amendment protections. *University of Pennsylvania Journal of Constitutional Law, 23,* 209.

Roberts, R. (2023). You're only mostly dead: Protecting your digital ghost from unauthorized resurrection. *Federal Communications Law Journal, 75*(2), 273–296.

Rogers, Y., & Marshall, P. (2017). *Research in the wild.* Morgan & Claypool Publishers.

Sargeant, A. (2017). The undeath of cinema. *The New Atlantis, 53*(Summer/Fall), 17–32.

Schraube, E., & Højholt, C. (2015). *Psychology and the conduct of everyday life.* Routledge.

Schultes, P., Dorner, V., & Lehner, F. (2013). Leave a comment! An in-depth analysis of user comments on YouTube.

Sherlock, A. (2013). Larger than life: Digital resurrection and the re-enchantment of society. *The Information Society, 29*(3), 164–176. https://doi.org/10.1080/01972243.2013.777302

Squires, B. (2022). Kendrick Lamar brings deepfake deep truths to 'The Heart Part 5' music video. *Vulture.* Retrieved December 28, 2022, from https://www.vulture.com/2022/05/kendrick-lamar-deepfakes-the-heart-part-5-nipsey-hussle.html

Wilson, S. (2021). *Steven Wilson - SELF (Official Video).* Retrieved May 2, 2023, from https://www.youtube.com/watch?v=link_placeholder

Sylvester, S. (2021). Don't let them fake you out: How artificially mastered videos are becoming the newest threat in the disinformation war and what social media platforms should do about it. *Federal Communications Law Journal, 73*(3), 369–392.

Stein, J.-P., & Ohler, P. (2018). Uncanny… but convincing? Inconsistency between a virtual agent's facial proportions and vocal realism reduces its credibility and attractiveness, but not its persuasive success. *Interacting with Computers, 30*(6), 480–491.

Temir, E. (2020). Deepfake: New era in the age of disinformation & end of reliable journalism. *Selçuk İletişim, 13*(2), 1009–1024. https://doi.org/10.18094/JOSC.685338

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *PLOS ONE, 18*(10), e0291668.

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*(1). https://doi.org/10.1177/2056305120903408

Watercutter, A., & Bedingfield, W. (2023). Hollywood actors strike ends with a deal that will impact AI and streaming for decades. *Wired.* Retrieved November 26, 2023, from https://www.wired.com/story/hollywood-actors-strike-ends-aistreaming

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11).

Zaramella, M., Amerini, I., & Russo, P. (2023). Why don't you speak?: A smartphone application to engage museum visitors through deepfakes creation. In *Proceedings of the 5th Workshop on the analySis, Understanding and proMotion of heritAge Contents (SUMAC '23)* (pp. 1–9). ACM. https://doi.org/10.1145/3607542.3617359

Chapter 4 Deepfake Misinformation Effects on Voters in the 2024 US Presidential Election

This paper is under review at *Advances in Psychological Science Open* (APSO).

Ching, D., Twomey, J., Meehan, E., Bolger, A., Linehan, C., Murphy, G. (under review). Deepfake Misinformation Effects on Voters in the 2024 US Presidential Election. *Advances in Psychological Science Open.*

Study Motivation

This study was inspired by findings from my scoping review which highlighted the relatively few empirical quantitative papers exploring the effects of deepfake exposure on beliefs, memories, and behaviours, especially in a political environment. I also wanted to provide strong empirical evidence demonstrating how deepfakes may or may not be uniquely persuasive compared to existing forms of misinformation such as text. Answering these questions allows us to contribute productively to early discussions regarding the unique impacts of deepfakes and to challenge speculative claims where they lack empirical evidence.

Author Contribution

Didier Ching, Conor Linehan, and Gillian Murphy conceptualized the study. Didier Ching and Gillian Murphy designed the surveys, Didier Ching collected the data and ran the studies. John Twomey, Aaron Bolger, and Eoghan Meehan contributed to creating the deepfakes used for the studies. Cian O' Mahony created all of the graphs in the paper. Didier Ching analysed the data and wrote the original draft and edited the document based on feedback from John

Twomey, Aaron Bolger, Eoghan Meehan, Conor Linehan, and Gillian Murphy. The final version of the manuscript was approved by every author.

4.1 Abstract

Media coverage and research literature frequently speculate that deepfake misinformation could be a particularly potent means of swaying political preference and voting behaviours due to their audiovisual nature and hyper-realism. Across five pre-registered experiments (total N = 5314) that were conducted before the 2024 USA presidential election, we compared the effects of misinformation about the candidates delivered via deepfakes with the same misinformation presented as text or synthetic audio. Findings suggest that deepfakes were not uniquely persuasive and were in fact more likely to be detected as fake than text or audio. The results did not change when a delay was introduced between exposure and assessment, or with repeated exposures. A follow-up study confirmed that the debriefing process implemented in each study was effective. While deepfakes have the potential to spread political misinformation, the current study does not suggest they are a uniquely potent means of doing so.

Research Transparency Statement

General Disclosures

Conflicts of interest: Didier Ching declares ownership of this manuscript. All other authors declare no conflicts of interest. Funding: This work was supported with the financial support of the Science Foundation Ireland grant 13/RC/2094_2 and co-funded under the European Regional Development Fund through the Southern & Eastern Regional Operational Programme to Lero – the Science Foundation Ireland Research Centre for Software (www.lero.ie). Artificial intelligence: No artificial intelligence assisted technologies were used in this research or the creation of this article. Ethics: This research received approval from the ethics board of University College Cork School of Applied Psychology.

Study one

Preregistration: All hypotheses, analyses were preregistered and can be found here (https://osf.io/g5s8q/?view_only=4ea57c3aca5a45b89b1cba403925353a). Similarly, all materials, data, and analysis scripts can also be found here (https://osf.io/g5s8q/?view_only=4ea57c3aca5a45b89b1cba403925353a). Identifying information such as participant's Prolific ID, email, IP address was removed. Any deviations from the preregistration are mentioned in the manuscript.

Study Two

Preregistration: All hypotheses, analyses were preregistered and can be found here (https://osf.io/g5s8q/?view_only=4ea57c3aca5a45b89b1cba403925353a). Similarly, all materials, data, and analysis scripts can also be found here (https://osf.io/g5s8q/?view_only=4ea57c3aca5a45b89b1cba403925353a). Identifying information such as participant's Prolific ID, email, IP address was removed. Any deviations from the preregistration are mentioned in the manuscript.

Study Three

Preregistration: All hypotheses, analyses were preregistered and can be found here (https://osf.io/g5s8q/?view_only=4ea57c3aca5a45b89b1cba403925353a). Similarly, all materials, data, and analysis scripts can also be found here (https://osf.io/g5s8q/?view_only=4ea57c3aca5a45b89b1cba403925353a). Identifying information such as participant's Prolific ID, email, IP address was removed. Any deviations from the preregistration are mentioned in the manuscript.

Study Four

Preregistration: All hypotheses, analyses were preregistered and can be found here (https://osf.io/g5s8q/?view_only=4ea57c3aca5a45b89b1cba403925353a). Similarly, all materials, data, and analysis scripts can also be found here

(https://osf.io/g5s8q/?view_only=4ea57c3aca5a45b89b1cba403925353a). Identifying information such as participant's Prolific ID, email, IP address was removed. Any deviations from the preregistration are mentioned in the manuscript.

Study Five

Preregistration: All hypotheses, analyses were preregistered and can be found here (https://osf.io/g5s8q/?view_only=4ea57c3aca5a45b89b1cba403925353a). Similarly, all materials, data, and analysis scripts can also be found here (https://osf.io/g5s8q/?view_only=4ea57c3aca5a45b89b1cba403925353a). Identifying information such as participant's Prolific ID, email, IP address was removed. Any deviations from the preregistration are mentioned in the manuscript.

4.2 Introduction

Deepfakes are synthetic audio and video media generated using Generative Adversarial Networks (GAN) that depict individuals saying or doing things they never did (Westerlund, 2019). Both academic and media discourse often frame political deepfakes as uniquely potent, speculating about the technology's potential to spread misinformation, threaten democratic processes, and jeopardize national security (Chesney & Citron, 2019; TBIJ, 2024; Westerlund, 2019). The aim of this study is to understand whether deepfakes are indeed uniquely effective at communicating political misinformation and swaying beliefs, memories and behaviour.

Existing theories suggest that people may update their beliefs due to misinformation when relying on System 1 thinking, an intuitive process using heuristics instead of deliberate reflection (Kahneman, 2011; Pennycook & Rand, 2019). Additionally, individuals are more susceptible to misinformation that aligns with their pre-existing beliefs, a phenomenon known as motivated reasoning (Vegetti & Mancosu, 2020). However, single exposure to misinformation rarely causes significant belief change, though the literature on this topic is dominated by studies assessing once-off exposures and subsequent beliefs (Murphy et al., 2023). Timing and frequency of misinformation exposure also matter, exemplified by the "sleeper effect," where message believability grows over time (Kleinnijenhuis, 2020), and the "illusory truth effect," where repeated exposure increases credibility (Udry & Barber, 2024). These biases are particularly relevant to deepfakes, as efforts to discredit deepfakes may weaken over time, while repeated online sharing inherent in internet culture may reinforce belief.

Misinformation can also influence memory through the misinformation effect (Loftus & Palmer, 1974) where exposure to misleading information can create false memories. For example, research has shown that Brexit voters were more likely to remember fabricated

events that supported their political stance, highlighting how social identity and post-event information can come together to create false memories (Greene et al., 2021). Murphy and Flynn (2022) demonstrated that deepfakes elicited false memories in participants, but not uniquely so compared to text or doctored photographs.

The Theory of Planned Behaviour (Ajzen, 1991) suggests that beliefs influence intentions, which shape behaviour, implying that misinformation can have downstream behavioural effects (Greene & Murphy, 2021; Laney & Loftus, 2016). Yet, belief does not always translate into action. For example, environmental concern doesn't always lead to pro-environmental behaviour (Landry et al., 2018), and belief in vaccine misinformation doesn't necessarily reduce vaccination intentions (de Saint Laurent et al., 2022).

Establishing whether deepfakes represent a special case of misinformation is necessary when prioritising investments in misinformation countermeasures such as in deepfake-specific detection technology or general media literacy campaigns (Ching et al., 2025). Existing studies suggest that deepfakes may reduce the credibility of a depicted speaker, negatively impact their reputation, and induce false memories, but not more so than existing forms of misinformation (Barari et al., 2025; Hameleers, 2024; Murphy et al., 2025; Twomey et al., 2025). Murphy et al., (2025) explored how exposure to deepfakes in a political environment affects voting intentions and found that defamatory deepfakes were capable of reducing intentions to vote for a politician by up to 23%, but this was no different than other forms of misinformation like simple text, further highlighting how deepfakes may not be uniquely persuasive.

There is however theoretical evidence to support the argument that deepfakes could be a uniquely potent form of misinformation compared to text or doctored photographs. Tabbers et al. (2004) proposes a "Modality Effect" whereby information presented multimodally through audio and visual pathways is easier to learn as it reduces cognitive load for the viewer. This theory could potentially explain how individuals may form erroneous beliefs when

viewing deepfakes, as reduced cognitive load can increase processing fluency and encourage more intuitive judgments encouraging System 1 type of processing. As a result, misinformation as deepfakes may be more passively accepted. Similarly for false memories, Dual Coding Theory (Paivio & Clark, 2006) suggests that visual information is encoded through both visual and verbal systems, whereas words are primarily encoded verbally, giving visuals an advantage. As visuals are also processed faster and provide more context, video-based misinformation may be encoded more robustly than other modalities, with its multimodal nature offering a more vivid and tangible representation of reality, potentially exceeding the misinformation effect of static images and narratives. Deepfakes differ in important practical ways from other forms of video-based misinformation as well, such as cheapfakes or shallowfakes, which involve the manipulation or recontextualization of authentic images and videos in ways that distort their original meaning (Hameleers, 2024). Although cheapfakes are generally easier to produce because they rely on basic video-editing tools, they remain constrained by pre-existing visual material. Deepfakes, in contrast, are not restricted to the content of real footage and can be used to fabricate entirely novel scenes, events, or dialogue that never occurred. This makes them a highly flexible and customizable form of misinformation.

The limited persuasive ability of once-off exposure to misinformation may not be surprising, as insights from political science literature indicate that it is exceedingly difficult to influence political attitudes and behaviours such as voting intentions or attitudes towards a politician (Coppock et al., 2020; Kalla & Broockman, 2018). This is because political attitudes and behaviours are informed by a multitude of factors such as identity, values, and demographics (Lazarsfeld et al., 1968) and not hugely swayed by isolated pieces of (mis)information. Despite a history of literature emphasising the complexity of misinformation and persuasion (Altay et al., 2023; Kalla & Broockman, 2018), there is still value in single exposure paradigms in understanding misinformation influences. Single exposure paradigms

demonstrate a mechanism for how specific types of misinformation can affects beliefs, memories, and behaviours, to which more complex studies involving individual factors or prior beliefs can build upon. This is especially crucial for types of misinformation that are not well-established or researched as it provides evidence for what the most basic example of this misinformation can do, such as deepfakes.

Despite the growing body of deepfake literature that study whether deepfakes are a uniquely persuasive form of misinformation, no studies have examined deepfakes on a population who are imminently experiencing a major election, nor have any explored the repeated or long-term implications of deepfake exposure. The emotionally charged and polarized atmosphere commonly seen in elections could influence how people react to political misinformation (Weeks, 2015) and research has shown that voters' choices may be different depending on the proximity to an election (Kalla & Broockman, 2018). A highly ecologically valid and large sample size is notably absent amongst empirical deepfake studies (Hameleers et al., 2022; Hameleers, 2024; Murphy et al., 2025). Moreover, examining the long-term effects of deepfakes is crucial as it broadens our understanding beyond the immediate impact of exposure, an area that political science research has often found to be minimal (Coppock et al., 2020). In the following set of studies, we are motivated by the research question: Are deepfakes uniquely effective at influencing political beliefs, memories, and behaviours of citizens during an ongoing election campaign? We conducted five large, preregistered studies during the 2024 US Presidential Election to investigate this question.

4.3 Experiment 1

We originally planned to utilise deepfakes of Donald Trump and Joe Biden, but Biden dropped out of the race the night before we began data collection. Considering this development, we began data collection with the Trump stimuli first (24th July 2024) and then

conducted the study with the Harris stimuli later, once she was confirmed as the Democratic nominee (6th September 2024). We pre-registered the analyses combining both datasets together and all materials and data are available here: https://osf.io/g5s8q/?view_only=4ea57c3aca5a45b89b1cba403925353a.

4.3.1 Participants

We recruited 1546 valid participants via Prolific (Prolific, 2025) based on an a priori power analysis using G*Power version 3.1.9.4 (Faul et al., 2007). The study was powered at 95% to detect an effect of f = 0.15 at α = .05. Participants ranged from 18 to 95 years of age, (M = 41.74, SD = 13.21), 43% men, 56% women, 1% were non-binary, a third gender, or transgender, and one preferred not to say. 37% of participants considered themselves Republican, 37% considered themselves Democrats, and 26% were neither Republican or Democrats. Half of the participants had completed some form of college education. All participants were residents of the United States at the time of recruitment. 13 participants were excluded due to failed attention checks, as pre-registered.

4.3.2 Materials

Deepfake audio: The generative AI software PlayHT (https://play.ht/) was used to create audio files of Trump and Harris. We uploaded authentic audio of both politicians to the software to create the synthetic voices, and we typed in our fake stories to create the audio used for the deepfake story conditions.

Deepfake video: We used the paid tool “Sync Labs” (https://synclabs.so/) to overlay the newly created deepfake audio onto an existing video of Trump or Harris talking to sync the audio with the politician’s lips.

News Stories: Participants in the experimental condition were shown three news stories: two true stories and one fake story. The true stories featured Ron DeSantis and Nancy Pelosi speaking about issues unrelated to the presidential candidates. The fake story featured either Trump or Harris and was presented as either text, synthetic audio, or a deepfake (see Figure 1.). We created four stories for each candidate covering four different themes: competency, giving credit to the opposing candidate, gun control, and abortion. The stories were designed in a way to match each candidate's vernacular and were piloted to verify that they painted the candidates in a poor light and were plausible (see supplementary materials). Participants in the control condition were only shown the two true stories.

4.3.3 Measures

Political Attitudes: Participant's attitudes towards each politician were measured using a "Feelings Thermometer" ranging from 0 to 100, where positive feelings are rated as "warm" and numerically above 50 and negative feelings are rated "colder" and below 50. This tool is commonly used in political surveys such as American National Election Studies to assess attitudes towards political individuals or groups (Zavala-Rojas, 2014).

Memories: Memories were assessed via the question: "Do you remember this event?" Possible responses were a) I have a clear memory of seeing/hearing about this; b) I have a vague memory of this event occurring; c) I don't remember this; d) I'm not sure. Responses of a) or b) to the fake stories were considered false memories, as in prior research (Greene & Murphy, 2020; Murphy & Flynn, 2022; Murphy et al., 2019).

Behavioural Intentions: After each story, participants were asked: "How likely would you be to share this story with friends and family? This includes social media sharing, sending via text, Whatsapp, or simply telling people about it." on a scale of 1 = Extremely Unlikely to 7 = Extremely Likely. Similarly, we measured voting intentions by asking participants to rate on a scale of 1-7 how likely they were to vote for each candidate in the upcoming presidential election with 1 = extremely unlikely and 7 = extremely likely.

Misinformation Detection: At the end of the study, we asked participants: “Before we wrap up, can you tell us what you think this study is about?” in a multiple-choice question with one choice representing successful detection (“How doctored media affects political views”), one choice matching the ruse given to participants (“How political news stories affect perceptions of candidates”) and seven other filler options (e.g. “How ChatGPT-written news stories are perceived” and “How political parties are perceived by the public”).

Figure 4-1. The screenshots of deepfakes of Kamala Harris and Donald Trump shown to participants in the deepfake condition under the “competency” theme. The stories were

presented as text, audio, or as a deepfake. The text below each deepfake represents what was said in the video and was not shown to participants in the deepfake or audio conditions.

4.3.4 Procedure

Participants first gave their consent to participate in our experiment, pitched as a study assessing people's attitudes towards politicians, before being asked demographic questions. Participants were then shown the news stories in random order. After each news story, participants answered an attention check where they were asked to select one description out of six which most closely described the contents of the story, followed by questions about their memories of the story, and their intentions to share it. Participants then answered questions about their attitudes towards both politicians and their intentions to vote for them. Finally, they were asked about the purpose of the study and were then debriefed.

4.3.5 Ethical Considerations

Extensive consideration was given to the ethical concerns surrounding this experiment. Ethical approval for all five experiments was attained from the University's ethics committee. While there are obvious risks to presenting misinformation to participants, we argue that that risk should be considered in the context of everyday risks (Kopelman, 2004). Specifically, in an election environment where misinformation is rife, brief presentation of very short pieces of misinformation that are then carefully retracted is likely not a significant departure from participants' daily experiences.

The primary ethical concern was that our misinformation stimuli would influence participant's views of the depicted politicians, in a manner that might persist beyond the conclusion of the experiment. This study was part of a larger funded project, where we have

previously found limited evidence of the persuasive nature of deepfakes in non-political settings (Murphy et al., 2023) and in a political context outside of an election (Murphy et al., 2025). Based on the findings of our previous studies, we did not expect significant persuasive effects of deepfakes in the current study. Nevertheless, every participant was extensively debriefed about the purpose of the study, using an evidence-based debriefing procedure that has been shown to successfully remove misinformation effects on beliefs (Murphy & Greene, 2023), false memories (Murphy et al., 2019) and behavioural intentions (Murphy & Greene, 2023), including in a study of political deepfakes (Murphy et al., 2025). Our multi-layered debriefing procedure included precisely highlighting which of the stories the participant viewed was fake, an explanation of deepfakes and our study, and direction towards medical professionals if participants were worried about their memory. We also included an active element to the debriefing process by asking them to identify the fake story from a multiple-choice question, which 63% of participants answered correctly, 12% answered incorrectly, and 25% didn't answer. Participants who answered incorrectly or didn't answer could have done so due to a phenomenon in survey research known as satisficing, where without sufficient motivation or incentives, participants tend towards answering in the least effortful way possible (Barge & Gehlbach, 2012). In this case, as the debriefing question was not an attention check (therefore it wouldn't affect the participants pay) and it was at the very end of the study post debrief, participants may not have had the motivation to engage with the debrief and neglected to answer or clicked through the question as quickly as possible. Despite this possibility, we also conducted a follow-up survey to ensure our debriefing was indeed effective (see Experiment 5)

A secondary ethical concern was that participants may object to the deception used in this study. Previous research suggests that participants generally enjoy misinformation and false memory studies and many report that they found the study even more interesting and

enjoyable after they were debriefed (Murphy et al., 2023; Murphy et al., 2020). However, as well as carefully and respectfully debriefing participants, we also gave them the opportunity to withdraw their data from our analysis post-debriefing (while still being paid for their participation), in line with best practice (Greene et al., 2023) Across all five of the experiments reported here, less than 1% of participants did so.

4.4 Results

All analyses were conducted on SPSS version 29.0.2.0. Our primary question was whether the format in which the misinformation was presented would affect subsequent attitudes, beliefs and behavioural intentions. As shown in Figure 2, a series of one-way ANOVAs were conducted to compare the mean values for the attitudes towards the politician, intentions to share the information, and intentions to vote for the depicted politician. Analyses indicated that there was no significant effect of misinformation format on attitudes, $F (3, 1542) = .70$, $p = .554$, $\eta 2 = .001$, sharing intentions, $F (2, 766.34) = .45$, $p = .640$, $\eta 2 = .001$, or voting intentions, $F (3, 1542) = .82$, $p = .484$, $\eta 2 = .002$. To assess whether there was a general effect of misinformation, we collapsed the misinformation conditions for a series of exploratory t-tests. We found no significant difference between the control condition and the combined text, audio, and deepfake conditions in political attitudes, $p = .225$, $\eta 2 = .001$, or voting intentions, $p = .129$, $\eta 2 = .002$.

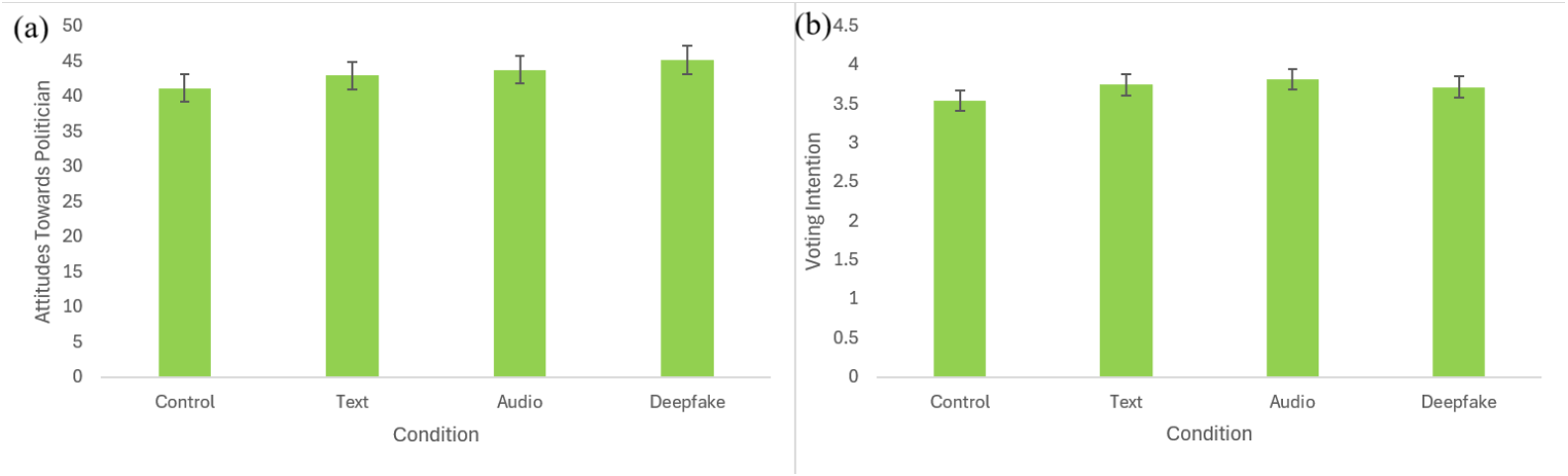


Figure 4-2. Bar charts showing the average (a) feeling thermometer scores towards a politician and (b) voting intention scores across experimental conditions with 95% confidence intervals. Both graphs indicate that there are no significant differences across misinformation formats.

While participants did readily report false memories for the fabricated news stories (text 12%, audio 14%, deepfake 15%), chi-square tests indicated no significant association between misinformation format and false memory rates, X^2 (2, n = 1153) = 3.28, p = .512, phi = .05. Most participants did not suspect the study was investigating misinformation (80%). However, there was an effect of format on detection rates, with participants more likely to detect the nature of the study if the misinformation was presented as audio (24%) or deepfake (24%) compared to text (14%), X^2 (2, n = 1153) = 17.85, p < .001, phi = .12.

Further pre-registered analyses are reported in Supplementary Materials, where we verified that our findings held even if we segregated by political orientation, or congruency between the fake story and political orientation, or if we examined each story individually. In all cases, AI-generated media was no more persuasive than simple text, and in fact was more likely to be detected as fake.

4.5 Experiment 2

Experiment 1 suggested deepfakes are not uniquely persuasive. An unexpected trend was also observed: exposure to the negative misinformation was associated with a small positive shift in attitudes toward the politician, such that participants in the misinformation conditions reported slightly more favourable attitudes and greater voting intentions. However, this effect was not statistically significant (see Figure 2). We hypothesised that in the pre-election environment, participants were on guard for misinformation and may have been particularly distrustful of explicitly negative information. This therefore could have led to a phenomenon known as the "backfire effect" where participants reinforced their position rather than changed their opinion (Nyhan & Reifler, 2010). We therefore replicated Experiment 1 but created "positive deepfakes" designed to promote a positive appraisal of the target candidate. These design decisions were inspired by events on August 18th, 2024, where Donald Trump posted AI-generated photos of Taylor Swift and her fans supporting his campaign (Aljazeera, 2024).

4.5.1 Participants

We recruited 1907 valid participants via Prolific. This study was powered at 95% to detect a small effect of $f = 0.1$ at $\alpha = .05$. Participant were recruited from the 24th - 28th October 2024 ranging from 18 to 88 years of age, ($M = 40.57$, $SD = 13.80$), 41% men, 57% women, 1% were non-binary or a third gender, one participant self-described as a transgender female, and three preferred not to say. 26% of participants considered themselves Republican, 49% were Democrats, and 25% were neither Republican or Democrats. Half of the participants had completed some form of college education. All participants were residents of the United States at the time of recruitment. 20 participants were excluded due to failed attention checks, as pre-

registered https://osf.io/g5s8q. We originally pre-registered a sample size of 1724 participants, but due to a survey design error we under-sampled the control condition and thus recruited an extra 183 participants to balance the number of participants in each condition.

4.5.2 Materials

The control stories and questions remained identical to Experiment 1. We developed three fake stories for each candidate covering three themes: celebrity endorsement, former president endorsement, and competency (all stimuli in supplementary materials and see Figure 2.). The stories were created through an iterative process between members of the research team. The primary investigator initially created the contents of the stories ensuring the statements were persuasive enough to trigger a response from the participant, while simultaneously not being overly offensive for ethical reasons. The research team then reviewed the statements and conducted a brief pilot study to assess the statements for plausibility and damage to reputation before finalizing them for the study. In the celebrity endorsement stories, Taylor Swift and Elon Musk were depicted endorsing Donald Trump and Kamala Harris respectively. Former President endorsement stories showed Barack Obama and George Bush endorsing Donald Trump and Kamala Harris respectively. Finally, the competency stories showed Kamala Harris and Donald Trump solving a complex mathematic riddle.

Figure 4-3. Sample screenshots of deepfakes from the former President endorsement and celebrity endorsement themes of Barack Obama and Elon Musk. The stories were presented as text, audio, or deepfake. The text below each deepfake represents what was said in the video and was not shown in the deepfake and audio conditions.

4.5.3 Procedure

Experiment 2 followed the procedure of Experiment 1 identically.

4.6 Results

Between-subjects ANOVAs indicated that attitudes towards politicians did not significantly differ across misinformation formats, $F(3, 1903) = 2.44$, $p = .063$, $\eta 2 = .004$ (see Figure 4). Mean sharing intentions, $F(2, 1469) = 1.02$, $p = .361$, $\eta 2 = .001$, and voting intentions, $F(3, 1903) = 2.00$, $p = .111$, $\eta 2 = .003$, did not differ across formats, suggesting that similar to Experiment 1, deepfakes were not uniquely persuasive compared to other misinformation

formats as shown in Figure 3. Unlike Experiment 1 however, exploratory t-tests demonstrated that participants in the combined misinformation conditions had significantly better attitudes towards the target politician, ($p = .004$, $\eta 2 = .004$) and higher intentions to vote for them ($p = .014$, $\eta 2 = .003$), than the control.

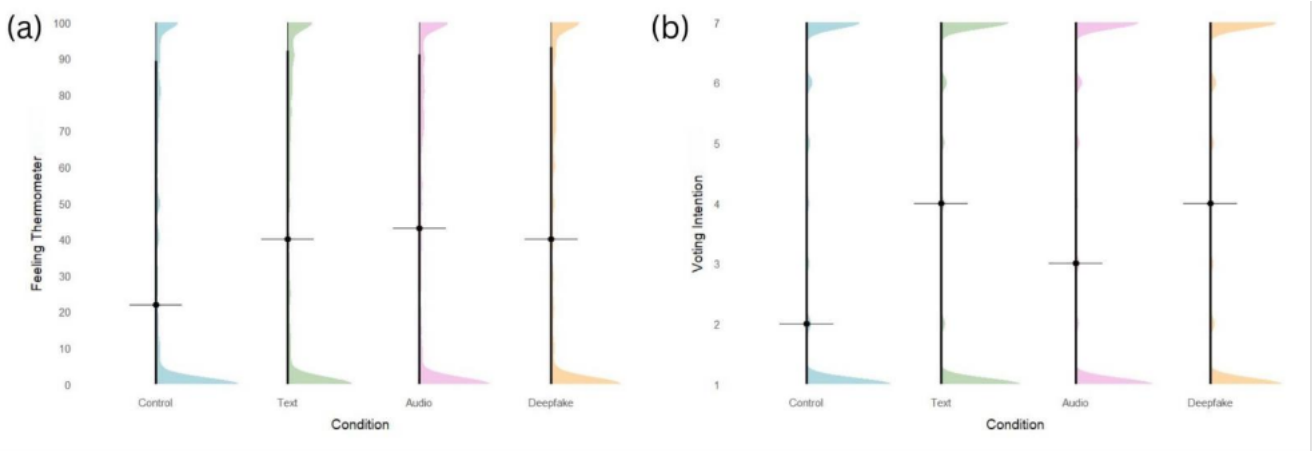


Figure 4-4. Bar charts in Experiment 2 showing no significant differences between (a) attitudes towards a politician and (b) voting intentions among misinformation formats.

Chi square tests showed that there were no significant differences between misinformation format and false memory rates, X^2 (2, n = 1472) = .14, p = .934, phi = .01, (text (12%), audio (12%), and deepfake (11%)). There was a difference between detection rates and misinformation format, X^2 (2, n = 1472) = 18.23, $p < .001$, phi = .11 where people who viewed misinformation in text format were less likely to detect the study's real purpose (10%) than those in the audio (17%) or deepfake (20%) conditions. Similar to Experiment 1, further analyses (in supplementary materials) indicated that our findings remained consistent regardless of political orientation, congruency between fake story and political orientation, and story type.

4.7 Experiment 3

While Experiment 2 found no uniquely persuasive effect of deepfakes, it is possible that viewing deepfakes may result in a "sleeper effect", where the persuasive power of a false message is enhanced over time (Kleinnijenhuis, 2020). Here, we delayed debriefing participants for three days to assess any possible sleeper effects.

4.7.1 Participants

This study was powered at 95% to detect a small effect of $f = 0.1$ at $\alpha = .05$. To achieve sufficient power, we required 520 valid participants across both waves of our study. To account for participant attrition, we recruited an additional 100 participants, making the total participant goal of 620. We recruited 621 participants for Wave 1 via Prolific. The extra participant was due to Prolific erroneously excluding a timed-out participant who fully completed the survey but presumably forgot to click the redirect link. Recruitment ran from the 30th October to the 1st November. After a three-day delay, we invited the 620 participants back for Wave 2, where we measured participants' attitudes towards politicians and voting intentions again. After participant attrition and removing those who failed attention checks, our final sample consisted of 523 participants ranging from 19 to 81 years of age, ($M = 38.31$, $SD = 12.63$), 50% men, 48% women, 1% were non-binary or a third gender, and 1% preferred not to say. 53% of participants had completed some form of college education. 20% of participants considered themselves Republican, 44% considered themselves Democrats, and 36% were neither Republican nor Democrats. All participants were residents of the United States at the time of recruitment but did not live in swing states where the election was expected to be close (see Ethics). Experiment 3 was pre-registered here: https://osf.io/g5s8q.

4.7.2 Materials

This study was identical to Experiment 2, but the fake stories were the ones that were most persuasive in Experiment 2; the former President and competency stories.

4.7.3 Procedure

At Wave 1, participants were shown three stories (2 controls and 1 fake story in either text, audio, or deepfake format) and then asked questions regarding their attitudes towards the politicians in the videos and to what extent did they intend to vote for Donald Trump or Kamala Harris in the upcoming presidential election. At the end of Wave 1, participants were not debriefed and were asked to take a follow-up survey (Wave 2) in three days where we measured attitudes and voting intentions a second time before proper debriefing. Participants did not see the stories again at Wave 2 and we did not measure sharing intentions, false memory rate, or misinformation detection rate, as it was not possible to do so without showing participants the news stories again.

4.7.4 Ethical Considerations

Experiment 3 posed some new ethical concerns, primarily, the delay in the debriefing procedure and the proximity to the election date. This risk was mitigated in a few ways. Firstly, Experiments 1 and 2 found miniscule effects on beliefs, memories, and behaviours when viewing deepfakes in comparison to control conditions. The time delay between time 1 and time 2 was also purposefully kept short (three days) to mitigate the time participants spent with the misinformation without the debrief. Secondly, we sent a debrief to any participant who did not return for Wave 2 to ensure that every participant was indeed debriefed. Thirdly, we took multiple steps to mitigate any potential influence of our misinformation stimuli on the election.

We avoided recruiting participants from seven potential swing states according to US News (Davis Jr., 2024). We also examined the electoral map from the previous 2020 USA Presidential elections to only include participants from states which were decided by at least a 20% margin in favour of either the Republican or Democrat party (CNN, 2020) to further reduce the very slim possibility of our study influencing election outcomes.

4.8 Results

Exploratory t-tests collapsing across misinformation conditions compared to the control condition revealed that there were no significant differences in political attitudes at Wave 1, $p = .370$, $\eta 2 = .002$ or voting intentions, $p = .180$, $\eta 2 = .003$. Mixed between-within subjects ANOVAs showed that for attitudes towards politicians, there were no significant interaction effects between time and format, Wilk's Lambda = .996, $F (3, 520) = .71$, $p = .546$, $\eta p2 = .004$. There were also no main effects of time, Wilk's Lambda = 1, $F (1, 520) = .17$, $p = .679$, $\eta p2 = .000$ or misinformation format, $F (3, 520) = 1.4$, $p = .243$. This suggests that not only were deepfakes not uniquely effective at influencing people's attitudes towards a politician initially, but they also did not have an impact after a time delay.

For voting intentions, there were no interaction effects between time and formatting, Wilk's Lambda = .998, $F (3, 520) = .32$, $p = .814$, $\eta p2 = .002$, no main effect for time, Wilk's Lambda = 1, $F (1, 520) = .01$, $p = .937$, $\eta p2 = .000$, and no main effect of formatting, $F (3, 520) = 1.54$, $p = .202$, as shown in Figure 5.

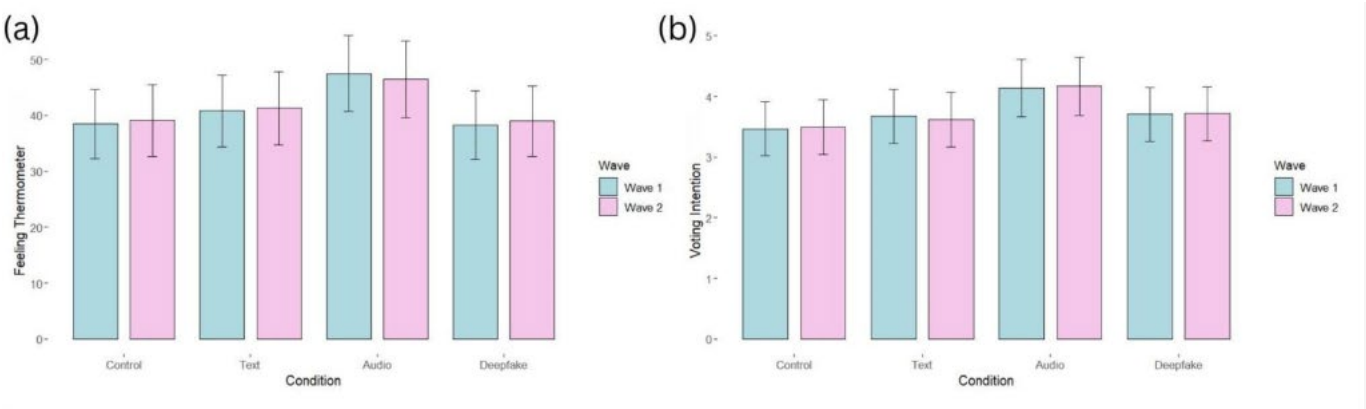


Figure 4-5. Bar charts in Experiment 3 showing no significant differences between (a) attitudes towards a politician and (b)voting intentions across misinformation formats and across a 3-day time gap between wave measurements.

Supplementary analyses dividing the population by political orientation, congruency of the fake stories, and fake story type did not influence the overall results (see supplementary materials). We also pre-registered an analysis to examine whether false memory rates in Wave 1 predict rates in Wave 2, but we dropped this analysis after realizing that we would need to expose participants to the stories again in Wave 2 to measure memory rates again and decided against doing so to prevent exposing participants to the stories a second time.

4.9 Experiment 4

Experiments 1-3 demonstrated minimal persuasive effects of deepfakes, but all involved a once-off exposure to misinformation. In our next experiment, we investigated any potential "illusory truth" effect of deepfake exposure (whereby repeated exposure to a piece of misinformation reinforces its impact) (Hassan & Barber, 2021). Misinformation literature has repeatedly shown that the "illusory truth effect" is a robust phenomenon capable of

manipulating beliefs and sharing intentions, regardless of whether the misinformation was implausible or contradicted pre-existing beliefs (Udry & Barber, 2024).

4.9.1 Participants

Similar to Experiment 3, we aimed to recruit 520 valid participants across Waves 1 and 2. This experiment was powered at 95% to detect an effect of f = 0.1 at α = .05. We recruited an additional 100 participants to account for participant attrition, setting a final target of 620 participants. We recruited 620 participants from Prolific for Wave 1 on the 1st November. After participant attrition and failed attention checks, we were left with 502 valid participants with ages between 18 and 78 years old, 47% men, 51% women, 1% were non-binary or a third gender, one person self-described as bigender and three people preferred not to say. 23% of participants considered themselves Republican, 44% were Democrats, and 33% were neither Republicans nor Democrats. 52% had achieved a college degree. All participants were residents of the United States at the time of recruitment.

4.9.2 Materials

Experiment 4 used identical materials to Experiment 3.

4.9.3 Procedure

Experiment 4 follows a similar design and procedure to Experiment 3, but we showed participants the same news stories in Wave 1 and Wave 2. We also asked participants the false memory questions from Experiments 1-3 and Wave 1 again in Wave 2 to assess the effect of repeated exposure of misinformation on memory. This enabled the inclusion of all of the original outcome measures from Experiments 1 and 2.

4.9.4 Ethical Considerations

This experiment implemented the same ethical procedures as Experiment 3 including a debrief sent through Prolific to every participant who dropped out after Wave 1.

4.10 Results

Exploratory t-tests suggested that there was no difference in political attitudes between the misinformation conditions and the control condition, $p = .993$, $\eta 2 = .000$, and in voting intentions, $p = .949$, $\eta 2 = .000$. Mixed between-within subject's ANOVAs revealed that for attitudes, there was no significant interaction effect between format and time, Wilk's Lambda = .99, $F (3, 498) = .72$, $p = .541$, $\eta p2 = .004$. There also was no main effect of time, Wilk's Lambda = .99, $F (1, 498) = .495$, $p = .482$, $\eta p2 = .001$, or format, $F (3, 498) = .61$, $p = .606$. Similarly, there was no interaction effect, Wilk's Lambda = .99, $F (3, 498) = .532$, $p = .66$, $\eta p2 = .003$, main effect of time, Wilk's Lambda = 1, $F (1, 498) = .11$, $p = .746$, $\eta p2 = .000$, or main effect for format, $F (3, 498) = .22$, $p = .884$, for voting intentions. Finally, for sharing intentions, there was no interaction effect, Wilk's Lambda = .99, $F (2, 370) = .78$, $p = .458$, $\eta p2 = .004$, main effect of time, Wilk's Lambda = .99, $F (1, 370) = .2.63$, $p = .106$, $\eta p2 = .007$, or main effect of formatting, $F (2, 370) = .92$, $p = .398$. Results indicate that deepfake exposure did not significantly affect participant's attitudes, voting intentions, or sharing intentions compared to text or audio, as shown in Figure 6.

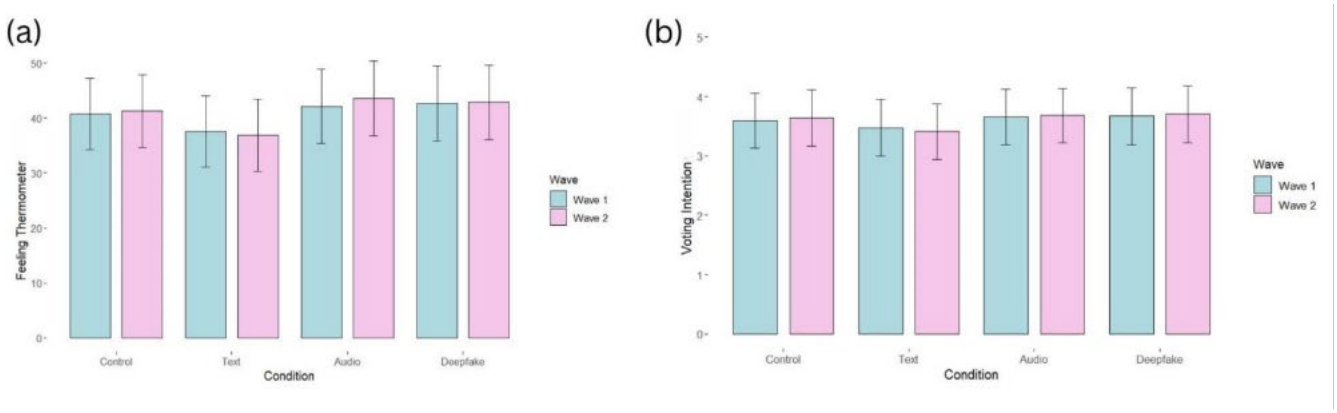


Figure 4-6. Bar charts in Experiment 4 showing no significant differences between attitudes (a) towards a politician and (b) voting intentions across misinformation formats and across a 3-day time gap between wave measurements.

We also conducted McNemar's tests to compare false memory rates within conditions to test whether repeated exposure to misinformation boosts its strength. We found increased false memory rates from Wave 1 to Wave 2 in the text condition (12% vs, 23%, $p = .004$), the audio condition (15% vs 34%, $p < .001$) and the deepfake condition (15% vs 38%, $p < .001$). Supplementary analyses exploring participants' data segregated by political orientation, congruency of the fake story to political orientation, and story type similarly suggested no significant persuasive ability of deepfakes over other misinformation conditions.

4.11 Experiment 5

In the final experiment, we assessed the effectiveness of our debriefing procedures. We asked: Do participants who have previously taken part in a misinformation study report higher false memory rates, lower detection rates, and higher truthfulness ratings for a) repeated vs novel stories and b) relative to newly recruited participants?

4.11.1 Participants

We recruited 835 participants via Prolific. This study was powered at 95% to detect an effect of d = 0.25 at α = .05. Participants were recruited from the 11th – 12th December, 418 were newly recruited participants and 417 were participants from Experiment 2. Participants were aged between 18 and 86 years old (M = 40.07, SD = 13.77), with 47% men, 52% women, 1% non-binary or third gender, and three people preferred not to say. 47% of participants achieved a college degree and all participants were residents of the United States at the time of recruitment.

4.11.2 Materials

Each participant saw four news stories presented as text with headlines. Two stories were true; one story that returning participants had previously seen before (ft. Ron Desantis) and a completely new control about Alexandria Ocasio-Cortez.

Participants also saw two fake stories. For newly-recruited participants, these were a random selection of one story from the former President endorsement theme (either the Barack Obama or George Bush fake story) and one story from the competency theme (either the Donald Trump or Kamala Harris fake story).

Returning participants viewed the same stories, but rather than being randomly assigned, they were bundled, so that participants either saw the Barack Obama and Kamala Harris fake stories together, or they viewed the George Bush and Donald Trump fake stories together depending on which stories the participants originally viewed in Experiment 2. For example, if a participant viewed the Kamala Harris fake story in Experiment 2, they would be allocated to

the Barack Obama/Kamala Harris stories to ensure that they viewed one story they have previously seen before, and one new story.

4.11.3 Procedure

This experiment followed a similar procedure to Experiment 2, except each participant was shown four news stories instead of three. After each news story, participants were only asked whether they remembered the event happening. Furthermore, after viewing all four stories, participants were asked whether they thought any of the four news stories they had just read were fake and to rate each one on a truthfulness Likert scale (where 1 = Extremely Untruthful and 7 = Extremely Truthful).

4.11.4 Ethical Considerations

This experiment followed the same ethical procedures outlined in Experiment 1.

4.12 Results

False memories: McNemar's tests indicated that the rate of false memories in returning participants for the repeated fake story was significantly higher (28%) compared to the false memory rates for the novel fake story (11%), $p < .001$. However, these false memory rates were the same or lower than those reported by the newly-recruited participants who viewed the same stories; 24% for the repeated story ($p = .209$), 24% for the novel story ($p < .001$).

Detection of fake news: For continuing participants, there was no difference in fake story detection rates for the repeated (64%) and novel (69%) fake stories, $p = .135$. There was also no significant difference between continuing participants and new recruits in detecting a

repeated fake story, with new recruits detecting the fake story 60% of the time, X^2 (1, n = 835) = 1.24, p = .255, phi = .04. Conversely, when viewing novel fake stories, newly-recruited participants were significantly less likely to successfully detect the fake story (60%) than continuing participants (69%), X^2 (1, n = 835) = 7, p = .007, phi = .09.

Perceived truthfulness of fake news: Continuing participants did not rate repeated stories as significantly more truthful (M = 3.15, SD = 2.16) than novel fake stories (M = 2.93, SD = 1.91), p = .079, d = .09. However, newly-recruited participants reported higher truthfulness ratings for fake stories (M = 3.8, SD = 1.99), p < .001, d =.32.

Supplementary analyses repeated these analyses on the true stories to examine whether participation in a previous misinformation study negatively impacted participant's belief in real news. Generally, results showed that returning participants had significantly higher recall rates, identification rates, and truthfulness ratings for repeated stories compared to novel ones. When compared to newly recruited participants, in general returning participants had significantly lower recall rates, identification rates, and truthfulness ratings for novel true stories. All analyses are detailed in the supplementary materials.

4.13 Discussion

The current paper advances the literature on misinformation, persuasion, and political theory by demonstrating that deepfakes exhibit persuasive capabilities comparable to those of existing misinformation formats. We found that contrary to the often alarmist literature (e.g., Chesney & Citron, 2019), deepfakes were not uniquely persuasive compared to textual or aural

misinformation, and in fact did not significantly impact political beliefs, memories, and behaviours, even after controlling for political ideation, attitudinal congruency, or individual stories. This finding was consistent even after a delay or repeated exposure. In fact, across five studies, the most consistent finding was that participants were more likely to recognize the study's purpose when exposed to synthetic audio or deepfakes, compared to text.

Our findings across the multiple experiments align with previous studies suggesting that deepfakes are not uniquely powerful (Hameleers et al., 2022; Hameleers et al., 2024; Murphy et al., 2025), even among voters in a polarized election and across multiple exposures. Our results could potentially be explained by findings from existing political science literature showing that it is difficult to influence political attitudes and behaviours (Kalla & Broockman, 2019; Eady et al., 2023). Decades of political science literature highlight the robust influence of party identification (Angus et al., 1960) and social identity (Tajfel & Turner, 1979) on political behaviours such as voting choice (Nickerson & Rogers, 2020). Party identification and social identities dictate political attitudes and behaviours and are resistant to external influences and incidents of misinformation because they are shaped by early socialization (Jennings & Niemi, 1981) and tend to be reinforced over time. Our findings suggest that deepfakes may not be an exception to these longstanding principles.

Our study carries several implications for research, policy, and public education. For researchers, the current study demonstrates that deepfakes function similarly to other forms of misinformation. We have no evidence that they are a “special case” or have outsized political effects. Policymakers may benefit from shifting attention away from counteracting the exaggerated harms of political deepfakes and towards broader misinformation mitigation strategies. In addition, media literacy efforts should not overstate the potential harms of deepfakes, as alarmist narratives risk producing a spillover effect and undermining trust in online information regardless of veracity (Hameleers, 2023; Van Der Meer et al., 2023).

Evidence of deepfake-fuelled scepticism can already be seen on social media (Twomey et al., 2023) and serves to reduce overall trust in news sources (Vaccari & Chadwick, 2020). The alarmist narratives frequently seen in deepfake educational interventions may especially erode trust in video-based information (Twomey et al., 2025). Surveys have shown that consumption of news online through videos is commonplace (Kalogeropoulos, 2017; Reuters Institute, 2024) and undermining trust in video-based information may have significant downstream consequences.

Of course, there are some limitations to taking an experimental approach to studying misinformation. For example, it is possible that the misinformation and deepfake stimuli used in the studies were not convincing enough and alerted participants to our deception thus influencing their responses. We argue that the quality of our stimuli is supported by the overall low average detection rate of the purpose of the experiments (22%) and the mean truthfulness ratings for the deepfakes in Experiment 5 was 3.44 out of 7 suggesting participants rated the videos as neither truthful nor untruthful. While the stories and deepfake quality could be improved, existing political research has indicated that using even less plausible stories and deepfakes could further delegitimize political figures (Hameleers, 2024), therefore, there is reason to believe that even deepfakes of a lower quality may have negative consequences on the perception of the politicians and be useful as stimuli in our experiment. Finally, critiquing the technical quality of the deepfake stimuli in this study risks overlooking a fundamental point about AI technology. AI technology will perpetually be in a state of development with uncertain futures (Dasoriya et al., 2018) and this criticism may apply to virtually all technological prototypes (Salovaara et a., 2017). Insisting that deepfake quality reaches an arbitrary level before testing the impact of exposure may result in missed opportunities to understand current capabilities.

The current study has underscored the importance of analysing the evidence base underpinning alarmist narratives about new forms of misinformation. We urge future researchers in this domain to continue to critically challenge alarmist narratives about technology-based misinformation while maintaining a strong commitment to transparent ethical research practices.

4.14 References

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. Social media+ society, 9(1), 20563051221150412.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Barari, S., Lucas, C., & Munger, K. (2025). Political deepfakes are as credible as other fake media and (sometimes) real media. *The Journal of Politics, 87*(2), 000-000.

Barge, S., & Gehlbach, H. (2012). Using the theory of satisficing to evaluate the quality of survey data. *Research in Higher Education, 53*(2), 182-200.

Broockman, D. E., & Kalla, J. L. (2023). When and why are campaigns' persuasive effects small? Evidence from the 2020 US presidential election. *American Journal of Political Science, 67*(4), 833-849.

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff., 98*, 147.

Coppock, A., Hill, S. J., & Vavreck, L. (2020). The small effects of political advertising are small regardless of context, message, sender, or receiver: Evidence from 59 real-time randomized experiments. *Science advances, 6*(36), eabc4046.

Dasoriya, R., Rajpopat, J., Jamar, R., & Maurya, M. (2018, January). The Uncertain future of artificial intelligence. In *2018 8th International conference on cloud computing, data science & engineering (confluence)* (pp. 458-461). IEEE.

de Saint Laurent, C., Murphy, G., Hegarty, K., & Greene, C. M. (2022). Measuring the effects of misinformation exposure and beliefs on behavioural intentions: a COVID-19 vaccination study. *Cognitive Research: Principles and Implications*, *7*(1), 87.

Eady, G., Paskhalis, T., Zilinsky, J., Bonneau, R., Nagler, J., & Tucker, J. A. (2023). Exposure to the Russian Internet Research Agency foreign influence campaign on Twitter in the 2016 US election and its relationship to attitudes and voting behavior. *Nature communications, 14*(1), 62.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175-191.

Greene, C. M., & Murphy, G. (2020). Individual differences in susceptibility to false memories for COVID-19 fake news. *Cognitive research: principles and implications, 5*(1), 63.

Greene, C. M., & Murphy, G. (2021). Quantifying the effects of fake news on behavior: Evidence from a study of COVID-19 misinformation. Journal of experimental psychology: *Applied, 27*(4), 773.

Greene, C. M., de Saint Laurent, C., Murphy, G., Prike, T., Hegarty, K., & Ecker, U. K. (2023). Best practices for ethical conduct of misinformation research: A scoping review and critical commentary. *European Psychologist, 28*(3), 139.

Greene, C. M., Nash, R. A., & Murphy, G. (2021). Misremembering Brexit: Partisan bias and individual predictors of false memories for fake news stories among Brexit voters. *Memory, 29(*5), 587–604. https://doi.org/10.1080/09658211.2021.1923754

Hameleers, M. (2024). Cheap versus deep manipulation: the effects of cheapfakes versus deepfakes in a political setting. *International Journal of Public Opinion Research, 36*(1), edae004.

Hameleers, M., van der Meer, T. G., & Dobber, T. (2024). Distorting the truth versus blatant lies: The effects of different degrees of deception in domestic and foreign political deepfakes. *Computers in Human Behavior, 152*, 108096.

Hameleers, M. (2023). The (un) intended consequences of emphasizing the threats of mis-and disinformation. *Media and Communication, 11*(2), 5-14.

Hameleers, M., van der Meer, T. G., & Dobber, T. (2022). You won't believe what they just said! The effects of political deepfakes embedded as vox populi on social media. *Social Media+ Society, 8*(3), 20563051221116346.

Hassan, A., & Barber, S. J. (2021). The effects of repetition frequency on the illusory truth effect. *Cognitive research: principles and implications, 6*(1), 38.

Jennings, M. K., & Niemi, R. G. (1981). *Generations and Politics: A Panel Study of Young Adults and Their Parents*. Princeton University Press.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kalogeropoulos, A. (2017). Online News Video Consumption: A comparison of six countries. *Digital Journalism, 6*(5), 651–665. https://doi.org/10.1080/21670811.2017.1320197

Kalla, J. L., & Broockman, D. E. (2018). The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *American Political Science Review, 112*(1), 148-166.

Kleinnijenhuis, J. (2020). Sleeper effect. In *The International Encyclopedia of Media Psychology* (pp. 1-7). Wiley.

Kopelman, L. M. (2004). Minimal risk as an international ethical standard in research. The *Journal of medicine and philosophy, 29*(3), 351-378.

Landry, N., Gifford, R., Milfont, T. L., Weeks, A., & Arnocky, S. (2018). Learned helplessness moderates the relationship between environmental concern and behavior. *Journal of Environmental Psychology, 55*, 18-22.

Laney, C., & Loftus, E. F. (2016). False memories matter: The repercussions that follow the development of false memory. In *False and distorted memories* (pp. 153–165). New York: Psychology Press.

Lazarsfeld, P., Berelson, B. & Gaudet, H. (1968). *The People's Choice: How the Voter Makes Up His Mind in a Presidential Campaign*. Columbia University Press. https://doi.org/10.7312/laza93930

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of Verbal Learning and Verbal Behavior, 13*(5), 585–589. https://doi.org/10.1016/S0022-5371(74)80011-3

Murphy, G., Loftus, E. F., Grady, R. H., Levine, L. J., & Greene, C. M. (2019). False memories for fake news during Ireland's abortion referendum. *Psychological science, 30*(10), 1449-1459.

Murphy, G., & Flynn, E. (2022). Deepfake false memories. In *Memory Online* (pp. 112-124). Routledge.

Murphy, G., Maher, J., Ballantyne, L., Barrett, E., Cowman, C. S., Dawson, C. A., ... & Greene, C. M. (2023). How do participants feel about the ethics of rich false memory studies?. *Memory, 31*(4), 474-481.

Murphy, G., & Greene, C. M. (2023). Conducting ethical misinformation research: Deception, dialogue, and debriefing. *Current Opinion in Psychology, 54*, 101713.

Murphy, G., Ching, D., Twomey, J., & Linehan, C. (2023). Face/Off: Changing the face of movies with deepfakes. *Plos one, 18*(7), e0287503.

Murphy, G., de Saint Laurent, C., Reynolds, M., Aftab, O., Hegarty, K., Sun, Y., & Greene, C. M. (2023). What do we study when we study misinformation? A scoping review of experimental research (2016-2022). *Harvard Kennedy School Misinformation Review*.

Murphy, G., Ching, D., Meehan, E., Twomey, J., Bolger, A., & Linehan, C. (2025). An Average Joe, a Laptop, and a Dream: Assessing the Potency of Homemade Political Deepfakes. *Applied Cognitive Psychology, 39*(2), e70061.

Nickerson, D. W., & Rogers, T. (2020). Campaigns influence election outcomes less than you think. *Science, 369*(6508), 1181-1182.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior, 32*(2), 303–330.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39-50.

Prolific. (2025). Prolific. https://www.prolific.com

Reuters Institute (2024). Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024

Salovaara, A., Oulasvirta, A., & Jacucci, G. (2017). Evaluation of prototypes and the problem of possible futures. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 2064-2077).

Tabbers, H. K., Martens, R. L., & Van Merriënboer, J. J. (2004). Multimedia instructions and cognitive load theory: Effects of modality and cueing. *British journal of educational psychology, 74*(1), 71-81.

TBIJ. Deepfakes and their threat to global democracy. (2024). TBIJ. https://www.thebureauinvestigates.com/stories/2024-06-05/deepfakes-and-the-threat-to-global-democracy

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *Plos one, 18*(10), e0291668.

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2024). What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology. *IEEE Transactions on Technology and Society*.

Twomey, J., Ching, D., O'Mahony, C., Linehan, C., & Murphy, G. (2025, April). I Was Never in the AI. RA: an Experimental Study on the Use of Manipulated Audio for the Censorship of Political Dissidents. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (pp. 1-8).

Twomey, J., O'Mahony, C., Linehan, C., & Murphy, G. (2025). Do No Harm: Designing Interventions for Deepfake-Fueled Conspiracy Theories. *Interactions, 32*(3), 30-33.

Udry, J., & Barber, S. J. (2024). The illusory truth effect: A review of how repetition increases belief in misinformation. *Current Opinion in Psychology, 56*, 101736.

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*(1), 2056305120903408.

Van Der Meer, T. G., Hameleers, M., & Ohme, J. (2023). Can fighting misinformation have a negative spillover effect? How warnings for the threat of misinformation can decrease general news credibility. *Journalism Studies, 24*(6), 803-823.

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

Weeks, B. E. (2015). Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of communication, 65*(4), 699-719.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review, 9*(11).

Chapter 5 Evaluating Methods to Educate the Public about Deepfake Technology

This paper will be submitted for review to *Media Psychology*.

Ching, D., Twomey, J., Linehan., C., Murphy, G. (pending review). Evaluating Methods to Educate the Public about Deepfake Technology. *Media Psychology*

Study Motivation

This multi-experiment study represents the culminating point of my thesis, designed to provide a forward-looking conclusion to my research. In this chapter, I move to evaluate an intervention to educate the public about deepfakes, drawing on my previous work which provided a nuanced understanding of quantitative and qualitative responses to deepfakes, and challenged existing narratives around them. It is both logical and important to explore strategies for mitigating the effects of misinformation in a thesis that explores the potential negative impacts of deepfakes. While results from Chapter 4 have indicated that deepfakes were not a uniquely impactful form of misinformation, studies have shown that deepfake exposure can still manipulate beliefs and memories, especially outside of an impending polarizing political event (Hameleers 2024, Murphy & Flynn, 2022). To this end, I dedicated an entire chapter to examining potential interventions aimed at reducing the negative effects of deepfake exposure in everyday environments. In my prior studies, we found that deepfakes are not uniquely persuasive

compared to existing forms of misinformation. It may be more effective to address issues that are unique to deepfakes and its audio-visual nature, such as their ability to undermine epistemic trust in video-based information. Additionally, my previous work has shed light on the exaggerated alarmist narratives surrounding deepfakes, but we still lack a clear understanding of how these narratives might negatively affect people. To explore this further, we decided to design an intervention and compare it with current educational approaches regarding deepfakes.

Author Contribution

Didier Ching, John Twomey, Conor Linehan, and Gillian Murphy conceptualized the study. Didier Ching, Gillian Murphy and Conor Linehan designed the interventions, and Didier Ching conducted the experiments and analysed the results. Didier Ching wrote the original draft and edited the document based on feedback from John Twomey, Conor Linehan, and Gillian Murphy. The final version of the manuscript was approved by every author.

5.1 Abstract

Deepfakes pose novel challenges to digital literacy and public trust in audiovisual information. Current efforts to educate the public about deepfakes often contain alarmist narratives, which may undermine people's trust in authentic information, or solely focus on detection tips, which may become obsolete over time. In this multi-experiment study (N = 2352), we examined the impact of educational interventions aimed at improving awareness and detection of deepfakes, and maintaining epistemic trust in authentic information. We conducted online, between-subjects experiments to compare the effectiveness of four different intervention approaches. We examined: 1) participants' ability to discern deepfake videos from non-edited videos, 2) self-reported deepfake anxiety, and 3) perceptions of emerging technologies more generally. We found that a deepfake detection-focused intervention improved detection confidence, but not detection skill. An alarmist intervention stressing the need for vigilance increased anxiety about deepfakes and concerns about other emerging technology but did not improve deepfake detection. In contrast, a more tempered evidence-based intervention significantly lowered participants' anxiety about deepfakes and emerging technologies and was also associated with higher rates of detecting authentic videos at follow-up. We propose that offering an evidence-based, measured explanation of deepfake technology can mitigate the negative spillover effects that sometimes result from well-intended educational efforts. Our findings highlight the importance of resisting alarmist narratives.

5.2 Introduction

Both researchers and the general public have in recent years expressed growing concerns about various forms of online misinformation (Allen et al., 2020; Lazer, 2018; Newman et al., 2022). One format of misinformation that has gained attention is the deepfake, which is a manipulated video that can depict people doing or saying something they never did, generated using Generative Adversarial Networks (GAN) (Albahar & Almalki, 2019; Westerlund, 2019). Researchers and policymakers have suggested the importance of developing educational interventions to warn the public and raise awareness about deepfake related harms (Dahir et al., 2024; Deng & Ahmed, 2025; El Mokadem, 2023; GovUK, 2025). However, there is a lack of evidence on how to design these interventions so that they bring about the desired improvements in digital literacy. Existing research on misinformation delivered through formats such as text and images reveal that misinformation interventions may be effective at reducing beliefs in false information, but at the cost of reducing the perceived credibility of information in general, (Guess et al., 2020; Hoes et al., 2024; Hameleers, 2023) and this may also be true for deepfake interventions (Twomey et al., 2025). Indiscriminate scepticism of information can be harmful as individuals may reject potentially valuable information from reliable sources. In this study, we sought to understand how people respond to educational interventions about deepfakes, and what elements contribute to an optimal educational deepfake intervention.

Current deepfake interventions generally fall into two categories. Firstly, there are interventions that specifically aim to boost an individual's deepfake detection ability by providing detection tips (Groh et al., 2022; Somoray & Miller, 2023). These interventions have largely been shown to be ineffective as deepfake detection depends on a complex variety of individual and stimuli-level factors and are not solely on deepfake quality (Somoray et al.,

2025). Appel and Prietzel (2022) demonstrated that higher cognitive ability and analytical thinking skills were associated with better deepfake detection, which aligns with existing misinformation literature documenting the importance of cognitive ability and analytical reasoning in determining the truthfulness of mis/information (Pennycook & Rand, 2019). However, the influence of other individual factors such as ethnicity and age remain unclear due to inconsistent results (Ahmed, 2021; Doss et al., 2023; Tahir et al., 2021). Stimuli-level factors such as the deepfake quality and familiarity with the deepfaked individual also have an impact on deepfake detection success. Higher levels of familiarity and lower quality of deepfakes is associated with better deepfake detection (Groh et al., 2022; Khan et al., 2023), but participants also run the risk of miscategorising low-quality videos with video artifacts as low quality deepfakes (Groh et al., 2022) inadvertently undermining trust in authentic videos. Detection-focused interventions may not be the optimal method of educating the public about deepfakes due to the aforementioned factors that influence susceptibility and may only be a temporary solution as any tips provided during the intervention (such as instructions to closely examine say, the mouth movements and whether they are out of sync with the audio) may become obsolete due to the evolving nature of deepfake quality (Twomey et al., 2025).

The second most prominent form of deepfake education intervention are general media literacy efforts that seek to raise awareness and warn people about deepfakes (AIforeducation, 2024; Kobis et al., 2021). This approach has been well validated in the general field of misinformation and proven methods include prebunking, where people are made aware of misinformation strategies before they are exposed (Vivion et al., 2022), media literacy interventions, where a person is exposed to strategies that encourage critical thinking and verification of information (Droog et al., 2024), and debunking, where people are shown corrective information discrediting misinformation they may have seen (Chan et al., 2017). These educational interventions have frequently been found to be effective at reducing beliefs

in misinformation (Hoes et al., 2024; Lewandowsky & Van der Linden, 2021). However, few empirical studies have explicitly tested these intervention strategies to counteract deepfake-based misinformation. Hwang et al. (2021) demonstrated that a media literacy intervention in the form of an educational video about deepfakes reduced participants' credibility ratings of deepfake messages, while Somoray & Miller (2023) conversely found that raising awareness of deepfakes did not improve people's ability to detect deepfakes. No studies have examined other intervention strategies such as prebunking or debunking methods with deepfakes leading to a gap between general misinformation intervention theories and our understanding of how they might apply to deepfakes.

Despite the efficacy and widespread implementation of counter-misinformation strategies, there is a growing concern that educational misinformation interventions can produce a negative spillover effect in which people become sceptical of all information regardless of veracity after being taught about misinformation (Hameleers, 2023; van Doorn, 2023). Even simple media discourse about misinformation can foster distrust in authentic information (van Duyn & Collier, 2019). This concern has been reinforced by empirical evidence demonstrating that misinformation interventions can sometimes reduce credibility assessments of both truth and false information (Hameleers, 2023; Hoes et al., 2024). Hoes et al. (2024) tested ecologically valid and widely used misinformation intervention strategies such as fact-checking, media literacy tips, and raising awareness of misinformation in the context of participant's credibility ratings of social media posts and found that these strategies negatively impacted the credibility of factual information.

Not all research supports the claim that misinformation interventions lead to harmful scepticism. For instance, Leder et al. (2024) investigated participants' responses to accurate information after playing the "Bad News" misinformation inoculation game and found only a marginal decrease in trust toward news headlines, with overall trust remaining high. They argue

that concerns about spillover effects are overstated, as the reduction in trust is both minimal and insignificant (e.g., ratings on a 7-point scale dropped from 7 to 6). Furthermore, they contend that a decline in perceived credibility following an intervention does not indicate a generalized distrust of all accurate information, but rather a context-specific response. This may be true, because typical misinformation strategies do not question all information but specifically target misinformation narratives such as in vaccines (Whitehead et al., 2023) or climate change (Cook, 2022) or misinformation techniques such as emotionally manipulative tactics or impersonation (Harjani et al., 2022). In contrast, deepfake-based interventions may foster a more generalized distrust of authentic information than traditional targeted misinformation strategies, as they inherently challenge the authenticity of visual evidence broadly (Fallis, 2021) rather than undermining trust in a specific topic or source. An intervention to reduce susceptibility to deepfakes must, by design, challenge the authenticity of video content in general, as deepfakes represent a *format* of misinformation rather than a specific, recognisable topic or technique. Designing an educational intervention against deepfakes presents a particular challenge because it risks undermining epistemic trust by encouraging scepticism toward the authenticity of video content as a whole (Twomey et al., 2025). However, avoiding raising awareness on deepfakes is equally problematic, as ignorance of their existence may lead individuals to either generate inaccurate alternative explanations for the videos they encounter or accept them uncritically (Ching et al., 2025).

Little research has been conducted on the potential spillover effects of deepfake-specific educational interventions. Lewis et al. (2023) explained to participants what deepfakes were, and that one of the five videos they were about to see was a deepfake, before asking participants to identify the deepfake from five videos. They found that those who received this deepfake warning were more likely to rate real videos as deepfakes, relative to a control condition who saw no warning. Ternovski et al. (2022) found a similar result after warning

participants about the dangers of deepfakes and asking them to distinguish between real and deepfaked videos afterwards. However, the interventions used in Lewis et al. (2023) and Ternovski et al. (2022) do not reflect current recommendations for intervention design as they were overly brief and alarmist (Green et al., 2023), thus reducing their ecological validity and preventing conclusions from being drawn about the potential spillover effects of modern interventions. Somoray & Miller (2023) and Groh et al. (2022) tested the effectiveness of detection and awareness interventions on deepfake detection and found mixed results. Somoray & Miller (2023) showed that raising awareness about deepfakes did not improve deepfake detection, while Groh et al. (2022) found a modest but significant improvement in detection rates after exposure to a deepfake literacy intervention. These studies did not examine whether the effect of the intervention on detection rates maintained after the experiment and measured detection through percentage of videos correctly identified. This method of measuring detection accuracy omits information about the strategies participants used such as blindly labelling every video as a deepfake for fear of being deceived (O'Mahony, 2024). As in general misinformation research, measuring discernment is key to evaluating the effectiveness of interventions (Guay et al., 2023; Murphy et al., 2023). Current empirical research on deepfake interventions suggests that enhancing individuals' detection accuracy and discernment remains challenging. Moreover, methodological limitations persist, including a lack of ecologically valid interventions, insufficient attention to the detection strategies employed by participants, and no long-term follow-up of intervention efficacy.

Deepfake intervention design should also consider negative downstream effects outside of the undermining of epistemic trust in authentic information. Current misinformation literacy campaigns often highlight the prevalence and potency of the misinformation they are trying to warn people about (Hameleers et al., 2023) which could inadvertently foster feelings of anxiety and discomfort about deepfakes, similar to AI anxiety (Kim et al., 2025). Researchers and

policymakers often adopt an alarmist attitude when discussing deepfakes, speculating on the harmful and widespread application of deepfakes, particularly in politics, without sufficient empirical evidence to support their claims (Ching et al., 2025) further exaggerating the harms of deepfakes. These feelings of anxiety may also affect how individuals view future technologies. Research has shown that fear and moral panic surrounding existing technologies is associated with a reduced acceptance of future technologies (Khasawneh, 2018, 2023). Alarmist discourse that generates panic surrounding deepfakes serves to spread stress and discomfort, potentially hinders perceptions of emerging technology that may improve people's lives, and may also undermine epistemic trust (Jungherr & Rauchfleisch, 2024). In addition, deepfake interventions also need to consider how they impact participants' confidence in detecting deepfakes. Research suggests that individuals are not proficient at detecting deepfakes (Somoray & Miller, 2023), even when they are aware that they may be seeing deepfakes or given a financial incentive (Kobis et al., 2021). Despite the inability to detect deepfakes, people are generally confident that they can, leading people to overestimate their abilities and underestimate deepfakes (Kobis et al., 2021). Thus, an intervention may inadvertently boost detection confidence, without improving skill, leading to an unintentional increase in susceptibility. Optimal deepfake interventions should consider participants' beliefs and skills in the round, rather than wholly focusing on improving detection.

With video-based news consumption steadily rising (Kalogeropoulos, 2017) and reports showing 66% of respondents from a 95,000 globally representative survey reported viewing at least one short news video a week (Reuters Institute, 2024), an intervention that undermines the authenticity and credibility of online videos may have downstream effects on how people obtain information and who they trust. It is reasonable to believe that people may label authentic videos as deepfakes due to elements inherent to audio-visual information such as video quality and definition, or even visual artifacts. Twomey et al. (2023) found that

individuals labelled a video of Vladimir Putin as a deepfake due to a visual artifact which resulted when the video was compressed, making it appear as though Putin's hand passed through a microphone. By promoting well-intentioned but poorly evidenced educational interventions for deepfakes, we risk polluting people's trust in video information. In this paper, we report a multi-experiment study to establish how deepfake educational interventions affect people's perception of video-based information and how a more optimal intervention should be designed.

5.3 Experiment 1

In Experiment 1, we tested a deepfake prebunking intervention, which was designed using current best practices for misinformation intervention design, on participant's perceptions of true political videos. We hypothesised that participants who viewed the intervention would perceive the authentic videos as significantly less truthful than those who did not view the intervention. Additionally, we hypothesised that poor video quality may render participants more likely to detect "false alarms" and declare authentic political videos as deepfakes. We pre-registered the analyses and all materials and data are available here (https://osf.io/xvke5/overview?view_only=e827e5ff49da43f1a09b37ce58df92de).

5.3.1 Participants

Data collection began on the 12th March (2024) through convenience sampling. This project was designed as part of an undergraduate module at a University in the Republic of Ireland, and undergraduate students recruited friends and relatives to initially take the survey. We also recruited participants from Prolific (Prolific, 2025) to reach the necessary sample size according to our power analysis. We recruited 478 participants in total, with 258 participants recruited by students, and 221 via Prolific based on an *a priori* power analysis using G*Power

version 3.1.9.4 (Faul et al., 2007). One participant was excluded due to failed attention checks, and therefore we recruited one more participant leading to a final sample of 478 valid participants. The study was powered at 90% to detect an effect of $p = 0.15$ at $\alpha = .05$. Participants were between the ages of 18 and 80 years of age, (M = 35.76, SD = 14.73), 37% men, 60% women, 2% were non-binary or a third gender, and 1% preferred not to say. Over half of the participants had completed a college/university degree. All participants recruited from Prolific were residents of the Republic of Ireland at the time of recruitment. We did not record the country of residence of participants recruited by students.

5.3.2 Materials

Video Stimuli: The video clips used for this study were downloaded from YouTube at 144p for the low definition condition, and at 720p for the high-definition condition. The videos were of authentic speeches made by Kamala Harris, Donald Trump, and Joe Biden discussing vaccines (see Figure 1 for an example, all videos can be found in the supplementary materials). Using a highly politicized topic such as vaccines served to ensure ecological validity and generalisability of our results. The topic also served as a believable ruse for participants as we framed the survey as one about attitudes towards vaccines.

Deepfake Intervention: The intervention was designed to inoculate participants from deepfakes in the manner outlined in "A Practical Guide to Pre-bunking Misinformation" (Harjani et al., 2022). This guide was created as a collaborative work between the University of Cambridge, Jigsaw (Google), and BBC Media Action, collating empirical evidence to provide instruction on designing effective misinformation interventions. The goal of the guide was to demonstrate how to educate the public about a technique or format of misinformation by first exposing them to a "microdose" of the misinformation first. We closely followed the document, including the

three elements of a successful prebunking message: 1) We warned participants that they were about to see one deepfake video (as our microdose of misinformation), 2) We pre-emptively refuted the misinformation by explaining deepfakes and how they are manipulative, and finally 3) we provided a microdose of the misinformation, a deepfake video showing Ron Desantis conceding to Donald Trump, as a weakened, practical example of the misinformation technique (see supplementary materials).

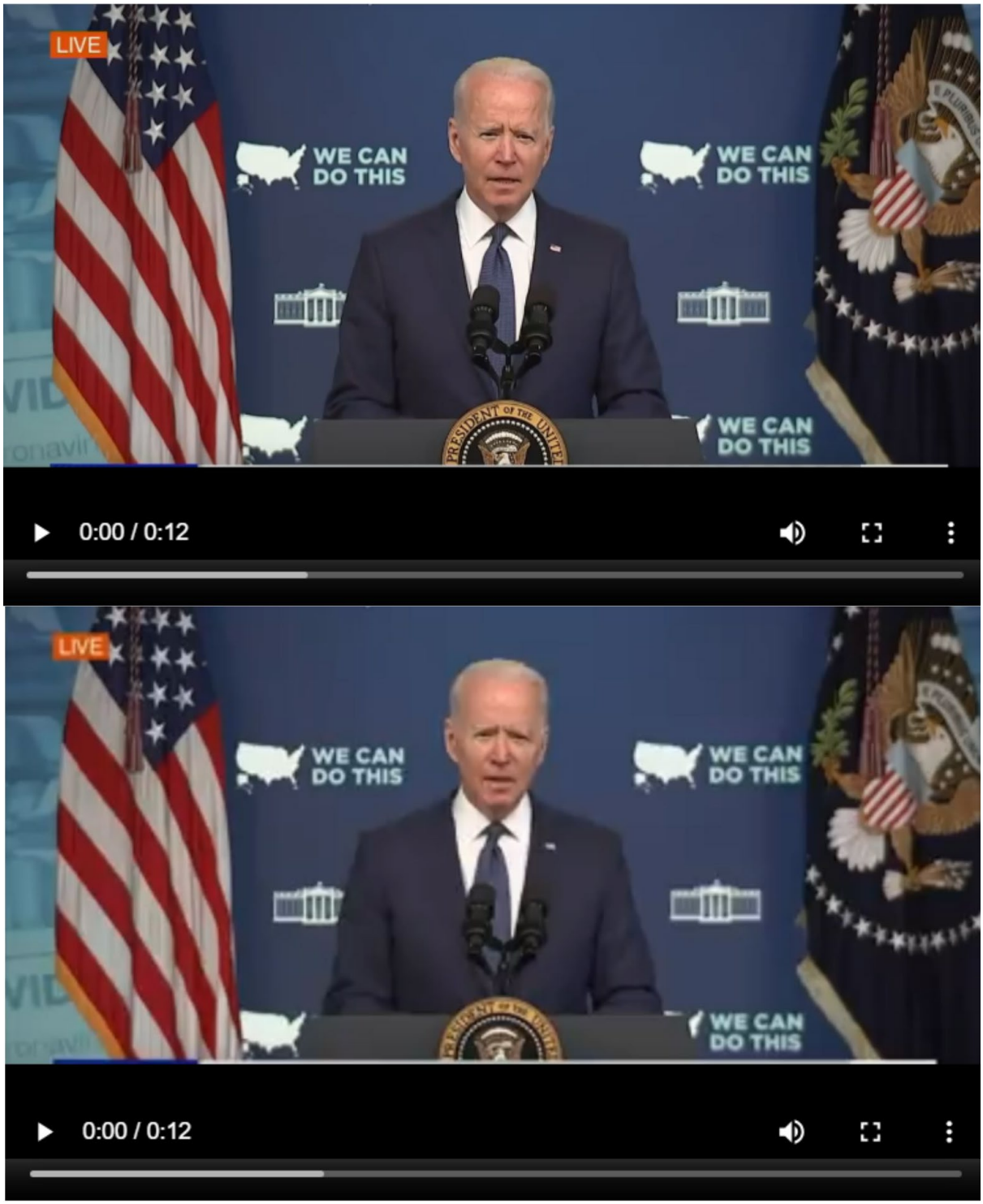


Figure 5-1. The screenshots of the high-definition Joe Biden video (top video) and low-definition Joe Biden video (bottom video) shown to participants.

5.3.3 Measures

Belief in the video: Participants' belief in the video was measured via their agreement with the following statement: "I believe that Politician X made that statement in their video" on a scale of 1 = Strongly disagree, and 7 = Strongly agree.

Sharing Intentions: After each video, participants were asked about the likelihood of them sharing the video with the following statement: "How likely would you be to share this video with friends and family? This includes social media sharing, sending via text, Whatsapp, or simply telling people about it" and answers were recorded on a scale of 1 = Extremely Unlikely, 7 = Extremely Likely.

Trust in news and social media: This is assessed via the question: "How much do you trust the news and information about affairs that you see on social media?" and responses were recorded on a scale of 1 = Completely trust to 7 = Completely distrust.

5.3.4 Procedure

Participants were first asked to give their consent to participate in the study, pitched as a study about statements made by politicians on vaccines. Participants answered demographic questions about their age, gender, education level, political ideation, and their support of vaccines. The vaccine question was to maintain the ruse of the study. Participants completed an audio check to ensure they could detect sound and then were assigned to one of four experimental conditions in a 2 (intervention: present vs absent) x 2 (video quality: high-definition vs low-definition) design. If participants saw the intervention, they were also asked an attention check about the intervention to ensure that participants read the intervention. Afterwards, participants were asked to view three videos, one each of Harris, Trump, and Biden (all three in high definition or low definition) and answered an attention check where they selected a general description of what was said in the video from a multiple-choice list, to ensure they had watched the video. They were then asked whether they believed the politician

made their statement and whether they would share each video. Participants were then debriefed (see supplementary materials for debrief).

5.3.5 Ethical Considerations

Ethical approval for all three experiments reported in this paper was attained from the University's ethics committee. We carefully considered the ethical concerns surrounding Experiment 1, and we identified and addressed each of them. Firstly, we considered the potential effects of showing participants the deepfake of Republican senator Ron Desantis capitulating the Republican presidential nomination to Donald Trump in the interventions condition. Harm from the deepfake exposure was limited as we sampled from Irish participants who were unlikely to be influential in the election of the Republican party presidential nominee. Also, Ron Desantis withdrew his name from the Republican presidential nomination in January 2024, before the experiment began.

Secondly, we considered the possibility of increasing participants' skepticism of authentic information after introducing deepfake knowledge. We ensured that our debriefing procedures included explaining the concept of deepfakes and urging participants to not immediately doubt video information based on deepfake knowledge. We included an active element of the debriefing process as well by asking participants to answer a multiple-choice question regarding the contents of the debrief. Despite not misinforming participants in the study, we emulated elements of empirically-tested debriefing procedures used in misinformation studies (Murphy & Greene, 2023)

Finally, we considered that some participants may have felt discomfort at the thought of being deceived by our ruse. Research suggests that participants generally do not object to studies involving deception such as is necessary in misinformation or false memory studies (Murphy et al., 2023). Also, at the end of the study, we gave participants the chance to withdraw

from the study if they felt uncomfortable and still be paid on Prolific, in accordance with established ethical recommendations (Greene et al., 2023). No participants elected to withdraw from the study.

5.4 Results

There was a statistically significant main effect of viewing the deepfake intervention on beliefs about the video, F (1, 474) = 21.54, $p < .001$, $\eta^2 = .043$, but deepfake definition did not significantly influence this effect, F (1, 474) = 4.65, $p = .032$, $\eta^2 = .01$. Individuals who received the deepfake warning intervention (M = 4.71, SD =1.13) rated the videos they saw as significantly less believable compared to individuals who did not receive the intervention (M = 5.21, SD = 1.26), regardless of video definition, as shown in Figure 2. There was also no significant interaction between video definition and viewing the deepfake intervention on beliefs about the video, F (1, 474) = 1.86, $p = .173$, $\eta^2 = .004$. The homogeneity of variances for this test was violated, therefore we only considered significance values of $p < .01$ as statistically significant (Pallant, 2020).

Participants did not significantly differ in sharing intentions across the four conditions with no significant interaction effects between the experimental condition and video definition, F (1, 474) = .54, $p = .463$, $\eta^2 = .001$ (see Figure 2) and no significant main effects of the experimental condition, F (1, 474) = 1.22, $p = .270$, $\eta^2 = .003$ or video definition, F (1, 474) = .30, $p = .586$, $\eta^2 = .001$. However, mean sharing intention values were below the midpoint across all conditions (M = 2.44, SD = 1.43), suggesting low intentions to share the videos regardless of quality or experimental condition.

Finally, there was no significant interaction effect between video definition and the experimental condition on mean trust in news on social media scores (see Figure 2), F (1, 474) = .09, $p = .765$, $\eta^2 = .000$. Similarly, there was no main effect of video definition, F (1, 474) =

1.41, p = .235, η^2 = .003 or experimental condition, F (1, 474) = .00, p = .965, η^2 = .000. Participants' mean trust in news on social media scores and standard deviations did not suggest a floor or ceiling effect (M = 3.51, SD = 1.43).

Further pre-registered analyses (see supplementary materials) revealed that these results held even if we repeated the analyses for each politician, except for the beliefs about the Joe Biden video where participants who viewed the low-definition video (M = 5.20, SD = 1.43) had a significantly lower belief in video score than the high-definition videos, though with a small effect (M = 5.47, SD = 1.25), p = .021, η^2 = .011.

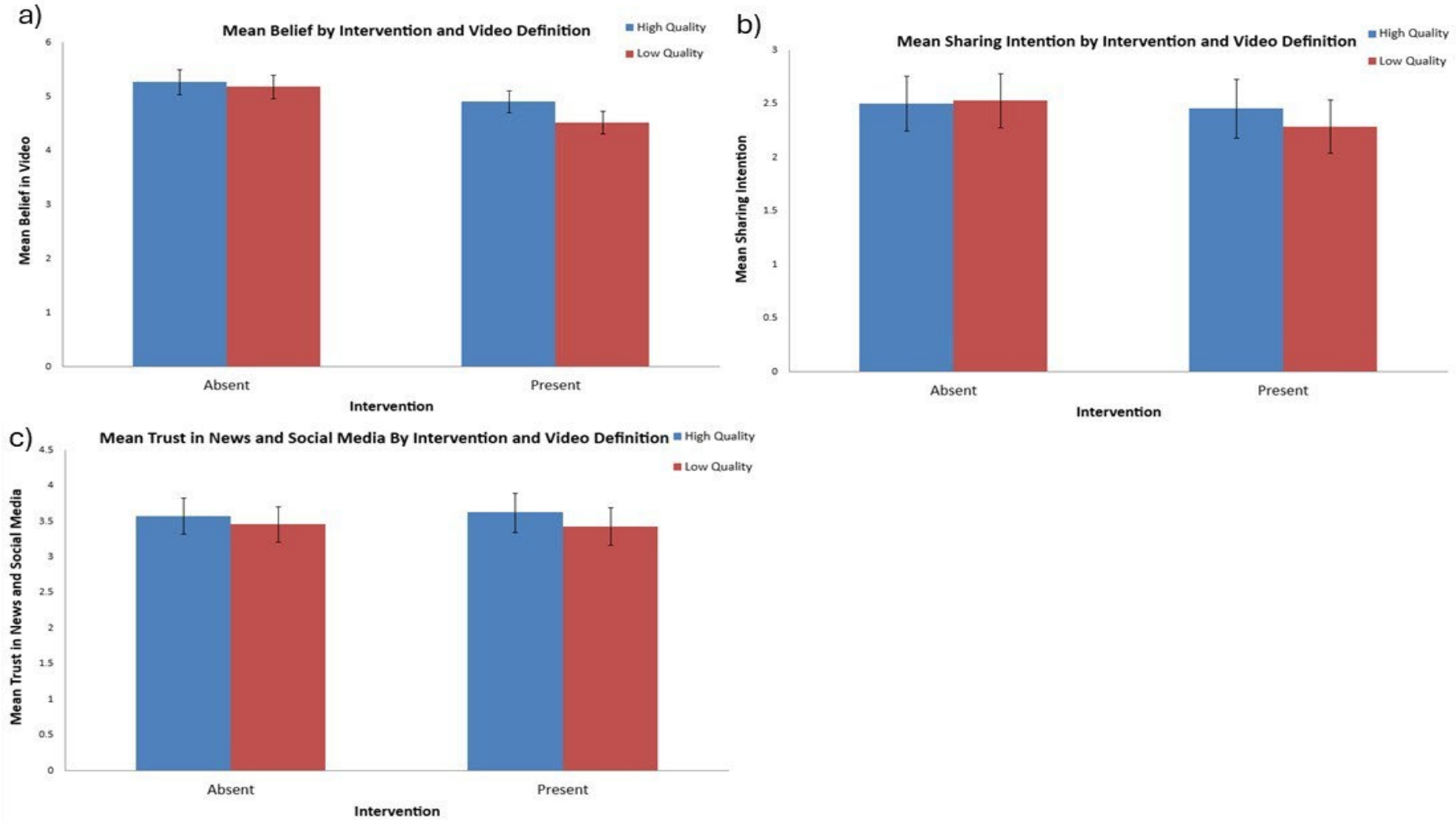


Figure 5-2. Bar charts for Experiment 1 showing no significant differences in (a) video belief, (b) sharing intentions, or (c) trust in news and social media across video definition with the error bars representing 95% confidence intervals, but the experimental condition significantly reduced mean belief in video scores.

5.5 Experiment 2

Experiment 1 established that an educational deepfake intervention following current best practices for misinformation interventions can negatively impact people's beliefs about authentic videos. In Experiment 2, we aimed to design an intervention to educate people about deepfakes without undermining their epistemic trust. We also considered other important recommendations for effective misinformation intervention design such as reducing alarmist discourse (Jungherr & Rauchfleisch, 2024), improving discernment abilities (Guay et al., 2023; Murphy et al., 2023), and limiting fear and panic surrounding deepfakes and emerging technologies (Hameleers, 2023). In this experiment, we aimed to design an evidence-based deepfake educational intervention to enhance deepfake discernment ability, while simultaneously reducing anxieties about deepfakes and future technologies. We compare the efficacy of this evidence-based intervention to existing deepfake literacy interventions with an alarmist framing comparable to modern deepfake discourse, and a detection-focused intervention comparable to existing deepfake interventions. All analyses were pre-registered and materials and data are available here (https://osf.io/xvke5/overview?view_only=e827e5ff49da43f1a09b37ce58df92de).

5.5.1 Participants

We began data collection on the 3rd March, 2025 via Prolific. The power analysis indicated we needed a sample of 1096 participants to detect an effect of 0.15 at $\alpha = .05$ if the study was powered at 80%. After recruiting 1096 participants, 37 participants were removed due to failed attention checks as pre-registered, and therefore an additional 37 participants were recruited leaving a total of 1096 valid participants. Participants' ages ranged from 18 to 80 years of age, ($M = 41.05$, $SD = 14.09$), 41% men, 57% women, 1% non-binary or a third gender,

and 1% preferred not to say. 68% of participants had obtained a college/university degree or higher, and only 5% of participants had never heard of the term “deepfake technology”. All participants were residents of the United States at the time of the experiment. US participants were used for experiment 2 because there were a limited amount of Irish prolific participants who had not already taken part in experiment 1. 37 participants were removed due to failed attention checks as pre-registered.

5.5.2 Materials

Interventions: We created three infographic interventions (see Figure 3). We chose infographics as our medium for the interventions for several reasons. Infographics have been shown to capture attention more effectively, are easier to process and recall information learned than basic text, and are more persuasive at shaping attitudes (Kessler & Bachmann, 2022). They have been effectively used in misinformation studies to reduce harmful beliefs such as in conspiracy theories (O’Mahony et al., 2024) and vaccines (Domgaard & Park, 2021).

Alarmist: This infographic contained information about deepfakes which emulated modern literacy campaigns and alarmist narratives such as deepfakes becoming widespread online in the near future, eroding trust in online video-based information (Duboust, 2024) or that they will have a severe impact in political environments (Chesney & Citron, 2019), which are unsubstantiated by empirical evidence (Ching et al., 2025).

Detection: This infographic provided detection tips for deepfakes and discussed deepfakes in a neutral tone similar to existing methods of raising awareness for this technology (Groh et al., 2023).

Evidence-based: This infographic involved an evidence-based approach to raising awareness of deepfakes. We specifically addressed unfounded claims about deepfakes such as their hypothesized harms in politics (Albahar & Almalki, 2019) and their uniquely persuasive

nature (Chesney & Citron, 2019) and instead focused on directing attention towards practical implications of deepfakes such as in non-consensual pornographic material generation (Karasavva & Noorbhai, 2021). The infographic also served to pre-emptively assure participants that real-world prevalence rates of misinformation were low, online information they viewed is generally authentic, and to be critical, but not overly sceptical of video-based information. The evidence-based intervention is designed to reduce anxiety around deepfakes and by extension future technologies, as well as urge viewers to not be overly sceptical of video-based information to improve discernment ability by maintaining trust in authentic videos.

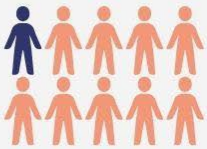
Deepfake Videos: The deepfake videos used in this study to assess discernment ability were generated in a two-stage process. The audio for the Kamala Harris and Donald Trump deepfakes were synthesized using generative AI software PlayHT (https://play.ht/). The audio files were then overlayed onto existing videos of Trump and Harris to sync to their lips using the paid site "Sync Labs" (https://synclabs.so/) (see supplementary materials for all videos shown).

Authentic Videos: The authentic videos used as part of assessing discernment ability were downloaded from YouTube and edited to match the length of the deepfake videos. These videos were of Kamala Harris discussing vaccines, Ron Desantis discussing the issue of homelessness, and Joe Biden discussing how his job is akin to taking a daily cognitive test (see supplementary materials).

Deepfakes

Deepfakes are a form of audiovisual misinformation generated using AI technology. In this graphic, we will explain what deepfakes are, how they can be harmful, and what to look out for if you think you are seeing one.

WHAT ARE DEEPFAKES?



Deepfakes are videos of a person in which their face has been digitally altered so that they appear to be someone else, typically used maliciously to spread false information.

HARMS OF DEEPFAKES

96% of deepfakes on the internet are pornographic in nature, mostly depicting the faces of celebrity women on the bodies of pornographic actors without their consent.

Deepfakes have been used to mimic the voices of friends, family, bosses, CEOs to defraud millions of dollars from people and companies.

Although deepfakes have the potential to be used in politics to defame candidates, there is little evidence to suggest that they can influence political beliefs

There is no scientific consensus on whether deepfakes are uniquely effective at persuading the viewer than edited photos or fake news articles.

One of the primary under-discussed harm of deepfakes is generating non-consensual pornographic content of females in order to harass, blackmail, or defame them.

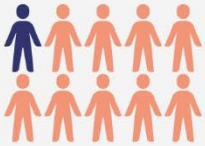
Don't be alarmed!

While deepfakes may sound scary, only a very small portion of videos online are deepfakes. Deepfakes aren't suddenly around every corner. Just remember to be vigilant and check your sources!

Deepfakes

Deepfakes are a form of audiovisual misinformation generated using AI technology. In this graphic, we will explain what deepfakes are, how they can be harmful, and what to look out for if you think you are seeing one.

WHAT ARE DEEPFAKES?



Deepfakes are commonly videos of a person in which their face has been digitally altered so that they appear to be someone else, typically used maliciously to spread false information.

HOW TO DETECT DEEPFAKES?

While deepfakes may be difficult to spot, we have compiled a list of things to look out for when watching a video to see if it is a deepfake.

Pay attention to the face. Look out for inconsistencies in facial features, especially in the cheeks and lips. They may look distorted.

Pay attention to objects on the head/face like hats or glasses. They may appear distorted or the glare in it might look unnatural when the head moves.

Pay attention to their speech patterns. Are you familiar with how the speaker talks and do they sound robotic or out of rhythm?

Consider the context of the video. Is it plausible that the person depicted is saying those things?

Be Vigilant!

If in doubt, check other sources to confirm whether the video is real. Keep these tips in mind when you view a video to see if it is a deepfake.

Figure 5-3. The intervention infographics showing (from top to bottom) the evidence-based intervention, detection-based intervention, and alarmist intervention.

5.5.3 Measures

Trust in news on social media: Participants were asked “How much do you trust the news and information about politics and public affairs that you see on social media?” and answers were given on a Likert scale with 1 = “Completely distrust”, and 7 = “Completely trust”. This measure was inspired by Vaccari & Chadwick (2020) and used as a general measure of uncertainty and scepticism of information on social media.

Deepfake Anxiety: We adapted the AI anxiety scale by Wang & Wang, (2022) to create a novel deepfake anxiety scale. Anxiety about deepfakes was measured with 8 items answered on a five point scale from 1 = Very slightly or not at all, 2 = A little, 3 = Moderately, 4 = Quite a bit, 5 = Extremely. The statements were: "I find deepfakes scary", "I am afraid that deepfakes will be misused", "I am afraid of various problems potentially associated with deepfakes", "I find deepfakes intimidating", "Being unable to keep up with the advances associated with deepfakes makes me anxious", "I feel distressed when I think about deepfakes ", "I feel nervous when I think about deepfakes ", and "I feel afraid when I think about deepfakes ". Cronbach’s alpha for the adapted deepfake anxiety scale was .94. Although the items were adapted from the AI Anxiety Scale (Wang & Wang, 2022), the factor structure observed in the present sample did not replicate the original multidimensional structure (see supplementary materials Appendix F: https://osf.io/xvke5/overview?view_only=e827e5ff49da43f1a09b37ce58df92de). Instead, the adapted items loaded onto a single component, suggesting that in the context of deepfakes they functioned as a unidimensional measure. Despite this difference in structure, the scale demonstrated excellent internal consistency, indicating that it was an acceptable adapted measure of deepfake anxiety for use in the present study.

Perceptions of Dystopian Technology: This was examined using a "design fiction" approach, in which participants evaluated a fictional article about an imagined technology (see supplementary materials for the full article), rating it based on their perceived level of danger and the speed of development of the hypothetical technology. Design fiction has been shown to be an effective tool in eliciting information from participants about fictional technologies to gain insights into their perceptions (Rapp, 2020). The article was intentionally designed to describe an exaggerated fictional technology that clearly has the potential to cause harm and poses a threat to privacy and safety, but is underdeveloped and unlikely to be realised for some time. The passage used in Experiment 2 detailed a microchip implanted in the brain that uses AI technology to record and upload one's dreams to the internet (see supplemental materials for the full article). After reading about the fictional technology, participants rated how much they agreed with the following statements on a scale of 1 = Strongly disagree to 7 = Strongly agree: "This technology is likely to cause a great deal of harm to society", "This technology is likely to develop into a big threat to society in the next 5 years", "I am excited to try this technology", "I would be happy to invest my money into this technology", and "I think this technology has a lot of beneficial uses to society". As pre-registered, only statements 1 and 2 were used to determine perceptions of future technologies, as we were concerned about perceptions of the speed of development and perceived harm of the technology. Statements 3-5 were used as red herrings to distract participants from the nature of the question.

Deepfake Discernment: Participants were shown 5 videos (2 deepfakes, 3 authentic videos) and asked "Do you think this is an authentic video?" with two possible answers: Yes and No. Binary answers were considered appropriate for this question to conduct an effective Signal Detection Analysis.

Discernment Confidence: After each video, participants were asked "What is your confidence that you guessed correctly?" (measured using an unnumbered graphic rating scale anchored by “50 = As confident as flipping a coin” and “100 = 100% sure”, as used in Somoray & Miller (2023)). Participants were also asked how many videos they thought they correctly identified out of 5.

5.5.4 Procedure

Participants first provided informed consent to take part in the study, which was introduced as a study about individuals' attitudes toward deepfake technology. They then completed a brief demographic questionnaire, including items on age, gender, and educational background. All participants were then presented with a brief and neutral explanation of deepfake technology, followed by a short example featuring a deepfake video of actor Jim Carrey portraying Jack Nicholson in a scene from “The Shining”. Following this, participants reported their prior familiarity with deepfake technology.

Participants were then randomly assigned to one of four conditions: a control group (no intervention), an alarmist intervention, a detection-based intervention, or an evidence-based intervention. After viewing the assigned intervention, participants completed an attention check involving a multiple-choice question with six options and were asked to select the option that most closely described the intervention. The options were intentionally designed so that the incorrect options included topics that were not mentioned in the infographic. For example, in the Evidence-based infographic, participants were given the choice between the correct option “The graphic explained that deepfakes make up only a very small portion of videos online. We should not become overly skeptical of real media and we should not be overly alarmed at this technology.” and incorrect options such as “The graphic explained how

deepfakes can be used to resurrect deceased family members". Finally, participants answered questions regarding our outcome variables (in a randomized order) before being debriefed.

5.5.5 Ethical Considerations

Ethical procedures for Experiment 2 closely followed Experiment 1. Our debriefing procedures were further updated by indicating to participants the videos that were indeed deepfakes in the five that they viewed. They were urged to be critical when viewing online video-based information, but to not be overly sceptical due to the real-world prevalence rates of deepfakes being extremely low. The evidence-based infographic was also shown to all participants who had not seen it during the survey and they were asked an attention check question to verify they had read the infographic.

5.6 Results

The primary aim of Experiment 2 was to compare the efficacy of different educational deepfake interventions and identify an optimal method for raising awareness of deepfakes.

5.6.1 Deepfake Anxiety and Perceptions of Dystopian Technology

A series of between-subjects ANOVAs demonstrated that there was a significant difference between the effects of the interventions on perceptions of dystopian technologies, F (3, 1092) = 4.11, p = .007, $\eta 2$ = .011. The evidence-based intervention had significantly lower scores when rating perceptions of dystopian technology (participants were less worried about the harm and speed of development of the fictional technology) (M = 9.37, SD = 2.86) than the alarmist (M = 10.2, SD = 2.85), p = .005, d = 0.29, and detection (M = 10.01, SD = 2.87), p = .045, d = 0.22 interventions, suggesting that while the evidence-based intervention did not

reduce negative perceptions about dystopian technologies compared to the control condition (M = 9.9, SD = 2.99), it tempered the negative perceptions of the alarmist and detection interventions (see Figure 4). Since the perceptions of dystopian technologies was measured using two 7-point Likert scale items so the minimum score was 2 and maximum score was 14. This suggests that the evidence-based intervention resulted in a score that was still slightly above the midpoint.

There was also a significant difference across interventions in mean deepfake anxiety scores, F (3, 1092) = 16.32, $p < .001$, $\eta 2 = .043$. Post-hoc tests reveal that the evidence-based intervention had significantly lower deepfake anxiety scores (M = 3.02, SD = 1.05) than the control (M = 3.33, SD = 1.02), $p < .001$, $d = .25$, alarmist (M = 3.59, SD = .97), $p < .001$, $d = .53$, and detection conditions (M = 3.46, SD = .922) $p < .001$, $d = .42$,. The alarmist condition also had significantly higher deepfake anxiety scores than the control, $p = .017$, $d = .25$. This suggests that the evidence-based intervention significantly reduced deepfake anxiety compared to the other interventions and the control, and the alarmist intervention significantly increased deepfake anxiety in participants.

Finally, there were no differences in Trust in News and Social media scores between the four conditions, F (3, 1092) = 1.01, $p = .386$, $\eta 2 = .003$ (M = 3.3, SD = 1.34).

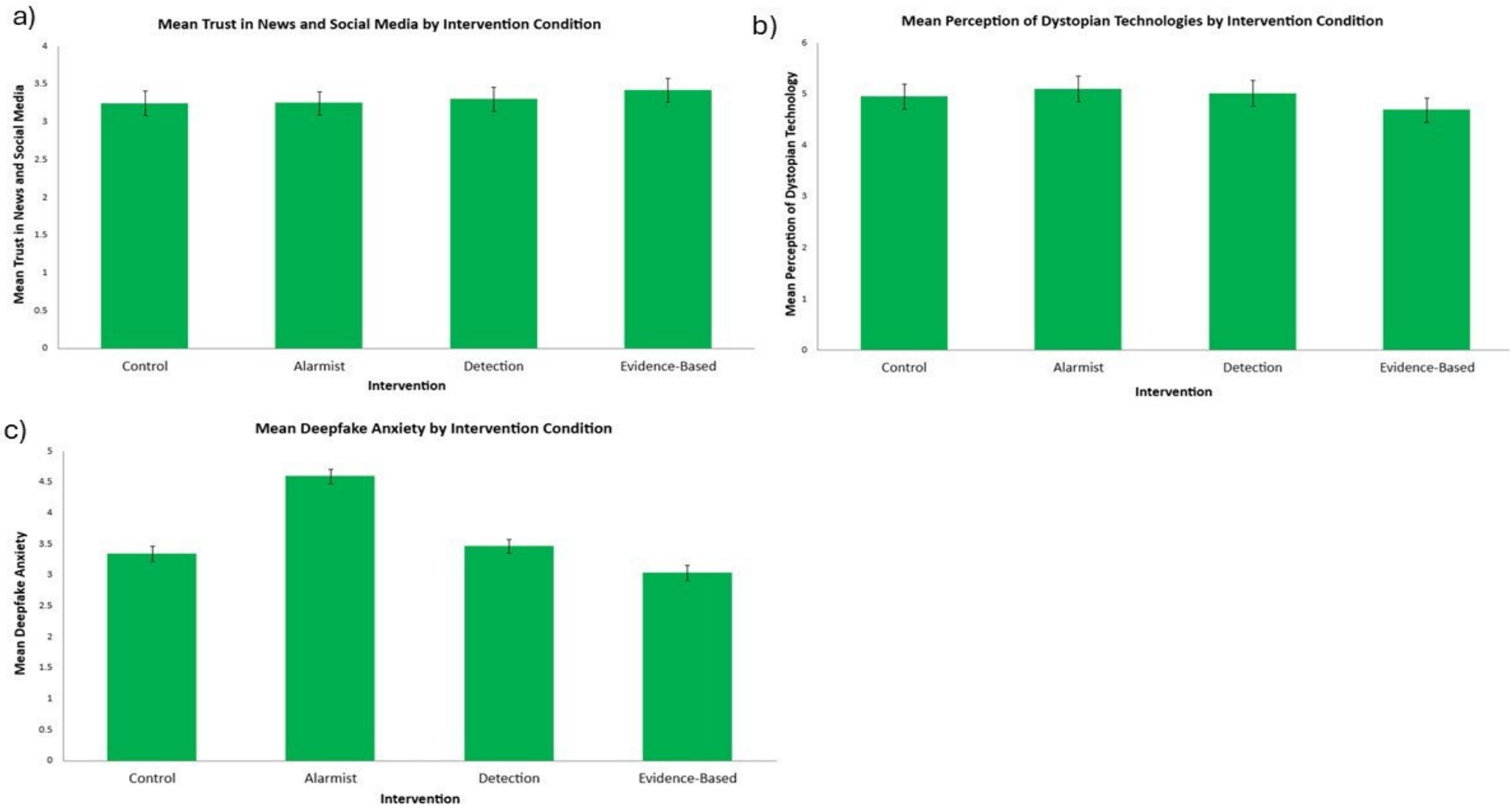


Figure 5-4. Bar charts showing the evidence-based intervention leading to (a) no effect on trust in news and social media and reduced (b) negative perceptions of dystopian technologies and (c) deepfake anxiety scores with the error bars representing 95% confidence intervals.

5.6.2 Deepfake Discernment

We drew on signal detection theory to measure ability to discern between deepfakes and authentic videos (known as "d′" or "d prime") and to identify what strategies participants tend to use when discerning between real and fake videos (knows as "c") (MacMillan & Creelman, 2005). ANOVAs revealed that none of the interventions significantly influenced d′ values, F (3, 1092) = 1.14, p = .348, $\eta 2$ = .003 suggesting that none of the interventions improved discernment ability. Similarly, the conditions did not significantly differ in c values, F (3, 1092) = .22, p = .867, $\eta 2$ = .001 and thus participants did not alter their discernment strategy regardless of intervention exposure. Average d' and c values of all participants was 1.2 and -.2 respectively, suggesting that overall, participants were moderately capable of discerning between deepfakes and authentic videos, and they slightly leaned towards declaring videos as deepfakes. One-sample t-tests indicated that both d′ and c were significantly different from zero, d′: $p < .001$, $d = 1.56$; c: $p < .001$, $d = - 0.51$.

In small samples (such as the current study in which there are only two deepfakes and three authentic videos), Signal Detection Analysis (SDA) can be unstable. This is because SDA relies on stable estimates of hit and miss rates which require a sufficient number of signal and noise trials to produce meaningful results (Stanislaw & Todorov, 1999). With only two opportunities to register a hit (two deepfakes out of five videos), the hit rate is limited to just being 0%, 50%, or 100% making it sensitive to random variation preventing reliable estimates of d' and c values and difficulty detecting smaller effect sizes. Stanislaw and Todorov (1999) suggests performing a log linear correction to the data to prevent extreme values due to the low number of trials. Therefore, while we have not pre-registered the use of log linear corrected values, we used the corrected values for this experiment.

To further understand participants' responses to the deepfake and authentic videos, we conducted a set of exploratory Kruskal Wallis and Mann Whitney tests on the different intervention conditions and their respective detection rates for authentic videos and deepfakes. These tests were considered appropriate as the successful detection rate data was interval data which was not normally distributed (Pallant, 2020). Results show that there was no significant difference in the successful detection rates of authentic videos, H (3) = 3.28, p = .351 or deepfakes, H (3) = 2.88, p = .411 across the intervention conditions and thus the interventions did not influence detection rates for either the authentic or deepfake videos.

5.6.3 Detection Confidence

Across both individual and overall measures of detection confidence, only the detection-based intervention significantly increased confidence compared to the alarmist condition. ANOVAs indicated that there was a significant difference between interventions in individual video confidence, F (3, 1092) = 3.75, p = .011, $\eta 2$ = .010. The detection intervention (M = 430.29, SD = 40.79) had significantly higher individual video confidence scores than the alarmist intervention (M = 418.69, SD = 40.83), p = .008, d = .28, while the evidence-based (M = 426.26, SD = 42.29) and control (M = 422.46, SD = 44.25) interventions did not significantly differ from other interventions.

Similarly, ANOVAs indicated a significant difference between interventions on overall confidence, F (3, 1092) = 3.69, p = .012, $\eta 2$ = .01 with the detection intervention (M = 3.65, SD = .91) leading to significantly higher overall confidence scores than the alarmist (M = 3.41, SD = .89), p = .012, d = .27. The control (M = 3.45, SD = .99) and evidence-based (M = 3.54, SD = .92) conditions did not differ significantly.

Correlational analyses were conducted to examine the relationship between video confidence (both overall and individual) and detection accuracy. All three experimental

conditions demonstrated a weak to medium positive correlation between overall confidence and detection accuracy (Alarmist, ρ (259) = .18, p = .004, Detection, ρ (271) = .13, p = .039, and Evidence-based, ρ (271) = .13, p = .039) suggesting that participants who were more confident were also slightly better at differentiating real from fake videos.

5.7 Experiment 3

Previous research on educational misinformation interventions has often lacked follow-up assessments to evaluate their sustained impact (Rau & Premo, 2025). As a result, there remains limited understanding of the long-term effectiveness of these interventions, raising concerns about their overall efficacy. Similar concerns have been expressed regarding deepfake interventions, with recent calls emphasizing the need for delayed post-test evaluations (Twomey et al., 2025). In Experiment 3, we address this gap by examining the effectiveness of our evidence-based interventions following a nine-day delay. Data collection began on the 12th March 2025, nine days after Experiment 2. An error was made when pre-registering the analyses here (https://osf.io/xvke5/overview?view_only=e827e5ff49da43f1a09b37ce58df92de). We conducted between-subjects ANOVAs alongside repeated-measures ANOVAs. We could not use Experiment 2's control group as the control for Experiment 3 because all participants at the end of Experiment 2 viewed the evidence-based infographic as part of the debriefing process, therefore we had to recruit new participants for the control group making it an independent-samples analysis. We still conducted repeated-measures ANOVAs comparing results from Experiment 2 to 3, but without the controls. All data and materials can be found here (https://osf.io/xvke5/overview?view_only=e827e5ff49da43f1a09b37ce58df92de).

5.7.1 Participants

We recruited 787 participants from Prolific, 255 were newly recruited participants who had not taken part in Experiment 2 who were recruited as a control group, and the remaining 532 were participants who had been part of the alarmist, detection-based, and evidence-based conditions in Experiment 2. Nine participants were removed due to failed attention checks, as pre-registered, leaving a total of 778 valid participants for the analysis. We recruited as many participants as we could from each of the intervention conditions from Experiment 2 for a period of 48 hours, as pre-registered. We chose to recruit 250 participants for the control as that was similar to the number of participants in each of the intervention conditions in Experiment 2. We could not recruit participants from the control condition of Experiment 2 because we debriefed those participants using the evidence-based intervention and therefore they would have been previously exposed to our intervention and could not serve as a control group going forward. Participants' ages ranged from 18 to 78 years of age, (M = 42.47, SD = 14.11). 41% of participants were men, 58% were women, 0.4% were non-binary or a third gender, 0.6% preferred not to say. 70% of participants had at least achieved a college/university degree and only 2% of all participants had never heard of deepfakes.

5.7.2 Materials

Deepfake Videos: The deepfake videos used were created in a similar fashion to Experiment 2. We created deepfakes of Kamala Harris and Joe Biden for this follow-up survey (see Figure 5).
Authentic Videos: Authentic videos of Kamala Harris, Ron Desantis, and Donald Trump were downloaded from YouTube as in Experiment 2 (see supplementary materials for all videos).

"It seems every week, there's a school shooting. We have to stop these once and for all. They say the only way to stop a bad guy with a gun, is a good guy with a gun. Well I believe the United States of America is full of good guys and we should loosen gun control and arm them. With more guns in the hands of fellow Americans, we can all protect our neighbours in times of trouble."

"You know the reversal of Roe V Wade changed my position on abortion and like many Americans I have changed my position on abortion. I've been lucky to have beautiful children and grandchildren and I know what a blessing they are. I think we can further protect the unborn by criminalising abortion after a heartbeat is detected."

Figure 5-5. Screenshots showing the deepfakes used for the deepfake discernment outcome measure with the transcript of each video underneath (the text was not shown to participants).

5.7.3 Measures

All measures were identical to Experiment 2. A new fictional technology article was created to assess perceptions of emerging technology measure (see supplementary materials). The new passage described how a new AI-powered technology could track one's thoughts and emotions and share them online with others to find other like-minded people to communicate with.

5.7.4 Procedure

The procedure was identical to Experiment 2, except participants did not view an intervention prior to answering the outcome measures.

5.7.5 Statistical Design

The original design of the experiment involved re-recruiting participants from Wave 1 (Experiment 2) and administering the same dependent measures in Wave 2, namely deepfake anxiety, perceptions of dystopian technologies, trust in news and social media, d' and c values, and overall and individual detection confidence. However, because participants in all Wave 1 conditions were debriefed using the evidence-based intervention at the end of that wave, the original Wave 1 control group could no longer function as an appropriate control condition in Wave 2. Consequently, a new control group was recruited for Wave 2. This meant that the primary analyses for Experiment 3 consisted of between-subjects ANOVAs comparing the newly recruited Wave 2 control group with participants from the experimental conditions established in Wave 1. In addition, repeated-measures ANOVAs were conducted to examine changes over time within the experimental conditions excluding the controls across Wave 1 and Wave 2.

5.8 Results

The primary research question for Experiment 3 was to determine whether the interventions from Experiment 2 continued to have an effect on deepfake anxiety, perceptions of future technologies, trust in news on social media, discernment ability, and detection confidence.

5.8.1 Deepfake Anxiety, Trust in News and Social Media, and Perceptions of Dystopian Technology

A series of between-subjects ANOVAs indicated that there was a significant difference in Deepfake Anxiety among the control and intervention conditions, $F(3, 774) = 5.02$, $p = .002$,

η^2 = .02. Participants who had previously been in the evidence-based intervention condition in Experiment 2 were significantly less anxious about deepfakes (M = 25.79, SD = 8.17) than the control condition (M= 28.7, SD = 8.1), $p < .001$. There were also significant differences in perceptions of dystopian technologies scores between conditions, $F(3, 774) = 6.01$, $p < .001$, η^2 = .02. The Evidence-based intervention condition significantly lowered negative perceptions of dystopian technologies (M = 9.32, SD = 2.78) compared to the control (M = 10.18, SD = 2.58), $p < .001$ (see Figure 6). Finally, there were significant differences in Trust in News and Social Media scores across conditions, $F(3, 774) = 3.14$, $p = .025$, $\eta^2 = .01$. Participants from the evidence-based intervention condition also had significantly lower trust in news and social media scores (M = 2.99, SD = 1.31) than participants who were in the detection-based condition (M = 3.39, SD = 1.33), $p = .026$.

Repeated-measures ANOVA revealed a significant interaction between condition and time on deepfake anxiety, $F(2, 525) = 20.94$, $p < .001$, $\eta^2 = .074$. Follow-up paired-samples *t*-tests comparing Wave 1 and Wave 2 indicated that participants in the Alarmist condition exhibited a significant reduction in deepfake anxiety at follow-up (Wave 1: $M = 29.18$, $SD = 7.7$; Wave 2: $M = 27.59$, $SD = 8.17$), $p < .001$, $d = .276$. In contrast, no significant change was observed in the Detection condition between Wave 1 ($M = 27.17$, $SD = 7.46$) and Wave 2 ($M = 27$, $SD = 7.51$), $p = .113$, $d = .126$. Notably, participants exposed to the Evidence-based intervention showed a significant increase in anxiety from Wave 1 ($M = 23.64$, $SD = 8.66$) to Wave 2 ($M = 25.79$, $SD = 8.17$), $p < .001$, $d = -.359$, although their overall anxiety levels remained lower than those observed in the other conditions. Results suggest that the Evidence-based intervention over time helped reduce anxiety around deepfakes for those who viewed the Alarmist intervention, but over time participants who viewed the Evidence-based intervention increased in anxiety over deepfakes although their deepfake anxiety levels remain below participants who viewed other interventions. The increase in deepfake anxiety for the evidence-

based intervention may have been due to significant events concerning deepfake legislation occurring in the US at the time of the experiments, which could have influenced how participants experienced deepfakes outside of the studies. On March 3rd and 4th, Donald Trump and his wife Melania publicly addressed and supported the "Take It Down Act" aimed at combatting deepfakes and online abuse (Oremus et al., 2025). It is possible that between Waves 1 and 2 of the experiment, concerns around deepfakes were heightened due to these public addresses.

For Trust in News and Social Media scores, there was no interaction effect between Wave 1 and 2 (time) and the intervention conditions, $F(2, 525) = .574$, $p = .563$, $\eta 2 = .002$. The main effect of Time was also not significant, $F(2, 525) = 1.42$, $p = .233$, $\eta 2 = .003$. Finally, there was a main effect of intervention type, $F(2, 525) = 5.283$, $p = .005$, $\eta 2 = .02$ with the Detection condition (M = 3.41, SD = 1.33) having significantly higher Trust in News and Social Media scores than the Evidence-based condition (M = 3.09, SD = 1.25), $p = .005$.

Finally, repeated-measures ANOVAs for Perceptions of Dystopian Technologies scores found that there was no interaction effect between time and intervention conditions, $F(2, 525) = 3.63$, $p = .027$, $\eta 2 = .014$ when Bonferroni corrected for three repeated-measures ANOVAs ($p = .0167$). There was also no main effect of time, $F(2, 525) = .208$, $p = .649$, $\eta 2 < .0001$. There was a main effect of intervention type, $F(2, 525) = 10.77$, $p < .001$, $\eta 2 = .039$. Pairwise comparisons reveal that the Evidence-based intervention resulted in significantly less negative perceptions of Dystopian Technologies (M = 9.13, SD = 2.82) compared to Alarmist (M = 10.15, SD = 2.7) and Detection-based (M = 10.08, SD = 2.73) interventions, $p < .001$. These results indicate that the Evidence-based condition significantly reduced participant's negative perceptions of Dystopian technologies.

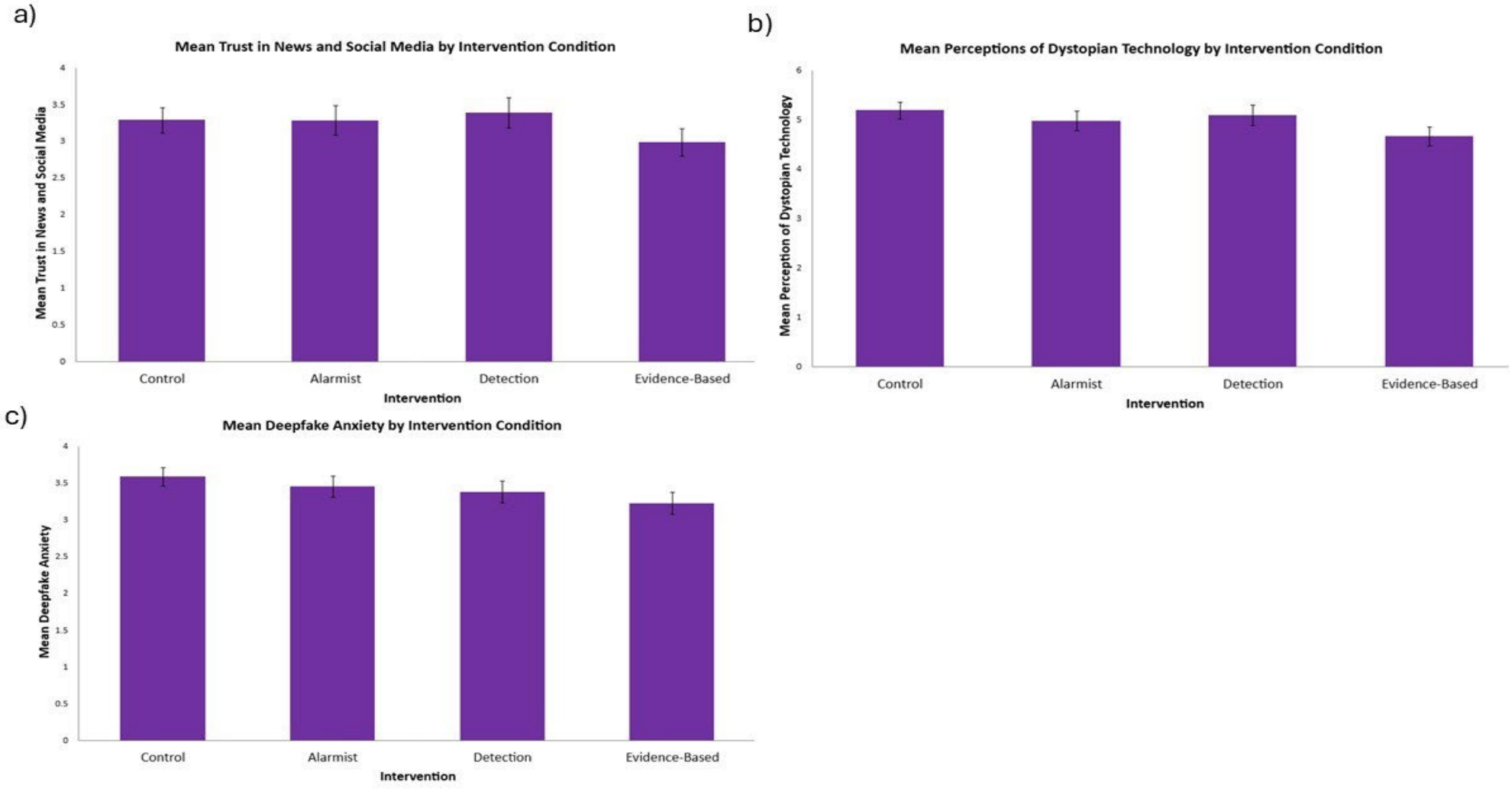


Figure 5-6. Bar charts showing participants who had previously been in the evidence-based condition having (a) trust in news and social media, (b) negative perceptions of dystopian technologies, and (c) deepfake anxiety scores compared to other conditions with the error bars representing 95% confidence intervals.

5.8.2 Deepfake Discernment

Two between-subjects ANOVAs were conducted to explore whether the interventions differed in terms of participant's ability to discern from real and fake videos and what strategies were employed (d` and c values respectively). Results revealed that there were no significant differences in d` values, F (3, 774) = 2.48, p = .06, $\eta 2$ = .01 and c values between the four conditions, F (3, 774) = 1.26, p = .289, $\eta 2$ = .005. Similar to Experiment 2, viewing any of the deepfake interventions did not influence participant's discernment ability or strategy. Overall, participants demonstrated moderate sensitivity in distinguishing deepfake from authentic videos (mean d′ = 1.44) and showed only a slight bias toward judging videos as deepfakes (mean c = - 0.05). One-sample t-tests indicated that both d′ and c were significantly different from zero, d′: $p < .001$, d = 1.86; c: $p < .001$, d = - 0.16.

For d' values, repeated-measures ANOVAs reveal that there were no interaction effects between time and intervention conditions, F(2, 525) = .621, p = .538, $\eta 2$ = .002. There was no main effect of intervention condition, F(2, 525) = .573, p = .564, $\eta 2$ = .002 but there was a significant main effect of time, F(2, 525) = 50.95, $p < .001$, $\eta 2$ = .088. Follow-up t-tests comparing d' values of the Alarmist, Detection, and Evidence-based interventions from Wave 1 to Wave 2 suggest participants in the Alarmist condition had significantly higher d' values at follow-up in Wave 2 (Wave 1 M = 1.19, SD = .72; Wave 2 M = 1.45, SD = .83), $p < .001$, d = -.281. Participants d' value in the Detection condition also increased significantly from Wave 1 (M = 1.27, SD = .85) to Wave 2 (M = 1.51, SD = .7), $p < .001$, d = -.266. d' values for the Evidence-based intervention also increased from Wave 1 (M = 1.16, SD = .74) to Wave 2 (M = 1.51, SD = .75), $p < .001$, d = -.39. These results suggest that on follow-up, participants were better at discerning between deepfakes and authentic videos.

For c values, repeated measures ANOVAs indicate that there were no interaction effects between time and intervention conditions, $F(2, 525) = 1.2$, $p = .303$, $\eta 2 = .005$. There was no main effect of intervention condition, $F(2, 525) = .442$, $p = .643$, $\eta 2 = .002$ but there was a significant main effect of time, $F(2, 525) = 7.63$, $p < .001$, $\eta 2 = .112$. Follow-up t-tests comparing c values of the Alarmist, Detection, and Evidence-based interventions from Wave 1 to Wave 2 suggest participants in the Alarmist condition had significantly higher c values at follow-up in Wave 2 (Wave 1 M = -.21, SD = .37; Wave 2 M = -.04, SD = .37), $p < .001$, $d = -.28$. Participants c values in the Detection condition also increased significantly from Wave 1 (M = -.17, SD = .38) to Wave 2 (M = -.04, SD = .34), $p < .001$, $d = -.27$. c values for the Evidence-based intervention also increased from Wave 1 (M = -.24, SD = .39) to Wave 2 (M = -.03, SD = .3), $p < .001$, $d = -.39$. These results suggest that over the 9-day delay between Wave 1 and 2, participants who had received the Evidence-based intervention were less likely to label the videos they saw in Wave 2 as deepfakes.

5.8.3 Detection Confidence

Two between-subjects ANOVAs were conducted to examine whether intervention type influenced the individual and overall video confidence when identifying real videos from fake ones. Results indicated that there was a significant effect of intervention type on individual video confidence, $F(3, 774) = 2.84$, $p = .037$, $\eta 2 = .011$. Participants who had previously been in the evidence-based condition were more confident when discerning deepfakes (M = 434.63, SD = 47.15) than the control (M = 423.11, SD = 47.98), $p = .041$, $d = .24$ while the alarmist (M = 432.58, SD = 39.77) and detection (M = 431.74, SD = 45.46) conditions did not differ from other conditions.

Conversely, results indicated that there were no significant differences among the interventions in terms of overall video confidence, $F(3, 774) = 1.70$, $p = .167$, $\eta 2 = .007$ with

control (M = 3.49, SD = .99), alarmist (M = 3.51, SD = 1), detection (M = 3.65, SD = 1.06), and evidence-based (M = 3.68, SD = 1.07) conditions showing similar values.

We compared the individual video and overall detection confidence of each of the four conditions with their corresponding detection accuracy values and found that individual video and overall detection confidence was positively correlated with detection accuracy and had a medium strength positive significant relationship indicating that increased confidence correlated with better detection performance (see supplementary materials for the full analyses).

5.9 Discussion

Through three studies we aimed to establish whether educational deepfake interventions might undermine the credibility of authentic videos, and to develop an approach that minimizes such unintended effects. Prior research (Hoes et al., 2024; Ternovski et al., 2022) has shown that deepfake interventions can reduce the perceived credibility of authentic videos. We demonstrated in Experiment 1 that applying current guidelines for educational misinformation interventions to deepfakes can undermine peoples trust in authentic information. We have also shown that an intervention designed with an alarmist framing, similar to widespread media coverage related to deepfake technology, significantly increases people's anxiety around deepfakes without improving detection. The detection-based intervention appeared to increase participants' confidence in their ability to identify deepfakes, as well as their trust in news and social media. However, it did not improve actual discernment performance. This pattern may be concerning, as it suggests participants may have developed an inflated sense of their ability to detect deepfakes despite showing no corresponding improvement in detection accuracy. Initially our evidence-based intervention did not immediately impact deepfake discernment and only lowered participant's anxiety about

deepfakes and emerging dystopian technologies. However, follow-up data showed that participants exposed to the evidence-based intervention maintained lower anxiety about deepfakes and future technologies, were better at discerning between deepfakes and authentic videos, and also were less inclined to call a video a deepfake, suggesting a small, delayed benefit to viewing the evidence-based intervention.

Findings from Experiment 1 on reduced trust following exposure to an educational intervention may not entirely be unexpected given prior evidence from both deepfake-based studies (Lewis et al., 2023; Ternovski et al., 2022) and misinformation intervention studies (Hoes et al., 2024; Hameleers, 2023). Our results have established that these effects happen specifically with deepfake-based educational interventions which were designed using current counter-misinformation guidelines. This suggests that some of the ways in which we currently conceptualize educational interventions could generate problems regarding the undermining of trust in authentic information if applied to deepfakes. Importantly, in line with Leder et al. (2024), we acknowledge that these effects may be context-dependent. Our study focused on political content, and we caution against generalizing these findings to signify that deepfake interventions reduce trust in all forms of online information without sufficient empirical evidence.

Our evidence-based intervention approach did not influence discernment ability immediately, but effectively boosted deepfake discernment after follow-up. However, the different control group recruited at follow-up introduces a potential confound as differences observed between conditions over time may partly reflect sample variation rather than effects of the intervention. This confound affects our interpretation of the effectiveness of our intervention at boosting discernment, which is arguably one of the most desirable outcomes for an educational intervention about a misinformation format. Our inability to boost discernment immediately closely resembles previous failed attempts to improve deepfake detection ability

(Groh et al., 2022; Kobis et al., 2021; Somoray & Miller, 2023). That deepfake detection ability is difficult to train in a brief intervention is understandable considering the complex factors that predict deepfake susceptibility, such as cognitive ability (Appel & Prietzel, 2022), ethnicity and age (Ahmed, 2021; Doss et al., 2023; Tahir et al., 2023), familiarity with the deepfaked subject (Khan et al., 2023), and deepfake quality (Groh et al., 2022). Participants in the evidence-based intervention after a nine-day follow-up also had lower trust in news and social media compared to the detection-based intervention. We hypothesise that the detection-based intervention increased participants' self-efficacy around deepfake detection and provided a sense of control over the technology, subsequently maintaining participants' trust in news and social media (Hocevar et al., 2014). This is further evidenced by the detection-based intervention leading to significantly higher individual and overall detection confidence scores compared to the alarmist condition. These results imply that promoting the public's confidence and self-efficacy in detecting deepfakes may be crucial for maintaining trust in news and social media. This possibly suggests that future interventions should comprise of a combination of evidence-based and detection-based methods.

Encouragingly, after a nine-day delay, participants who had previously been exposed to the evidence-based intervention demonstrated greater discernment ability and detection accuracy in identifying true videos compared to those in the control group, suggesting that the intervention was effective in mitigating the extent in which educational efforts may inadvertently undermine trust in legitimate political videos. Our evidence-based intervention also proved to be effective at reducing anxiety surrounding deepfakes and emerging technologies, even across a nine-day delay. This finding shows that it is possible to counteract the alarmist and unfounded narratives about deepfakes that are widespread in the media, which is crucial for the public to maintain trust in authentic news sources and information (Hameleers, 2023) and maintain a tempered perception of emerging technologies. We encourage future

deepfake interventions to include information about the prevalence rate and real-world implications of deepfakes to reassure participants about the capabilities of deepfakes and reduce spillover effects of educational interventions.

We explain the efficacy of the evidence-based intervention over the control and other interventions through its emphasis on discernment and critical thinking. We theorize that the evidence-based infographic improved true video detection accuracy in authentic information compared to the other infographics and the control because we redirected their focus towards discernment, rather than encouraging blind scepticism (O'Mahony et al. 2024). The importance of this approach is underscored by studies that highlight the harmful consequences of encouraging blind scepticism (Hameleers, 2023; Hoes et al., 2024). Our approach also emphasises the need to maintain critical vigilance when discerning between real videos and deepfakes, an approach which has been supported by general misinformation literature (Machete & Turpin, 2020). Our study supports this notion, as we encouraged critical thinking and vigilance at the end of the evidence-based intervention, which we believe contributed to the observed improvement in participants' ability to accurately identify genuine videos. This finding is increasingly important for researchers when designing and testing educational misinformation interventions considering the evidence demonstrating that educational interventions may undermine epistemic trust through negative spillover effects (Hameleers, 2023; Hoes et al., 2024).

Our work has important implications for public educators creating educational content about political deepfakes, and for misinformation educators in general. We argue that it is counterproductive to include alarmist narratives and empirically unfounded fears around political deepfakes in educational deepfake interventions as they serve to increase anxiety around deepfakes as shown in our findings, negatively affect perceptions of emerging technologies and undermine epistemic trust (Jungherr & Rauchfleisch, 2024). It is important

to refer to empirical and factual evidence when creating educational interventions about deepfakes to ensure the public are accurately informed about the real-world harms deepfakes pose such as in non-consensual pornography generation (Karasavva & Noorbhai, 2021) and financial fraud (Ke et al., 2025). We do recommend including a brief note at the end of deepfake interventions and misinformation interventions in general reassuring viewers that misinformation isn't as prevalent as some may think and that maintaining vigilance to information we encounter is key to countering misinformation.

There are several limitations to this study. Firstly, in Experiments 2 and 3, the interventions might have been more effective as videos instead of infographics, to more accurately represent deepfakes. Teaching about an audio-visual form of misinformation such as deepfakes in a two-dimensional format may risk a disconnect between the misinformation we are teaching about and the reality of the technology even though we separately provided a visual example of a deepfake to participants. Another limitation is our use of only five videos to determine deepfake discernment. We did not use enough videos to conduct a signal-detection analyses that could reliably detect smaller effects, leading us unable to determine participant strategy when discerning deepfakes from authentic videos (Stanislaw & Todorov, 1999). While higher quality deepfakes may have produced differing results across Experiments 2 and 3, our results reflect how participants respond to deepfakes created with current capabilities after exposure to educational deepfake interventions. Regardless of the technological advancement of deepfakes and technological prototypes in general (Salovaara et al., 2017), there remains value in understanding perceptions of deepfakes today and this includes understanding how deepfakes of various different qualities are perceived.

The current study has highlighted the unintended consequences of educational efforts surrounding deepfakes. By designing an evidence-based educational intervention about deepfakes that does not overstate the harms of deepfakes, we show that it is possible to reduce

technophobia around deepfakes and maintain trust in authentic political videos. We urge future public educators to design interventions that presents an accurate picture deepfakes and prevent the use of alarmist narratives.

5.10 References

Ahmed, S. (2021). Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. *Personality and Individual Differences, 182*, 111074.

AI for Education, (2024) Uncovering Deepfakes: Classroom Guide. AI for Education. https://www.aiforeducation.io/ai-resources/uncovering-deepfakes

Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology, 97*(22), 3242-3250.

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.

Altay, S., Berriche, M., & Acerbi, A. (2023). Misinformation on misinformation: Conceptual and methodological challenges. *Social media+ society, 9*(1), 20563051221150412.

Asadi, O. (2025). Exploring Current and Potential Solutions: The Rise of Deepfakes in Legislative, Legal, and Technological Arenas. *Berkeley Undergraduate Journal, 39*(1).

Appel, M., & Prietzel, F. (2022). The detection of political deepfakes. *Journal of Computer-Mediated Communication, 27*(4), zmac008.

Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science, 28*(11), 1531-1546.

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war: The coming age of post-truth geopolitics. *Foreign Aff., 98*, 147.

Ching, D., Twomey, J., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2025). Can deepfakes manipulate us? Assessing the evidence via a critical scoping review. *PLoS One, 20*(5), e0320124.

Ching, D., Twomey, J., Linehan, C., & Murphy, G. (2025). Encountering Deepfakes: A Thematic Analysis of Comments on a Music Video featuring Deepfake Content. *Proceedings of the ACM on Human-Computer Interaction*, *9*(7), 1-23.

Cook, J. (2022). Understanding and countering misinformation about climate change. *Research anthology on environmental and societal impacts of climate change*, 1633-1658.

Deng, R., & Ahmed, S. (2025). Breaking the illusion: impact of news literacy on deepfake identification and sharing. *Online Information Review*.

Dhahir, D. F., Kenda, N., Dirgahayu, D., Syarifuddin, S., Djaffar, R., Widihastuti, R., ... & Pala, R. (2024). The Relationship of Digital Literacy, Exposure to AI-Generated Deepfake Videos, and the Ability to Identify Deepfakes in Generation X. *Jurnal Pekommas, 9*(2), 357-368.

Domgaard, S., & Park, M. (2021). Combating misinformation: The effects of infographics in verifying false vaccine news. *Health Education Journal, 80*(8), 974-986.

Doss, C., Mondschein, J., Shu, D., Wolfson, T., Kopecky, D., Fitton-Kane, V. A., ... & Tucker, C. (2023). Deepfakes and scientific knowledge dissemination. *Scientific reports, 13*(1), 13429.

Droog, E., Vermeulen, I., van Huijstee, D., Harutyunyan, D., Tejedor, S., & Pulido, C. (2024). Combatting the misinformation crisis: A systematic review of the literature on characteristics and effectiveness of media literacy interventions. *Communication Research*, 00936502251363705.

Duboust, O. (2024, June 13). "Faster than expected": Experts warn deepfakes are spreading. Euronews; Euronews.com. https://www.euronews.com/next/2024/06/13/deepfakes-are-spreading-in-scams-and-on-social-media-faster-than-expected-experts-warn

El Mokadem, S. S. (2023). The Effect of Media Literacy on Misinformation and Deep Fake Video Detection. *Journal of Arab Media & Society*, 53-78.

Fallis, D. (2021). The epistemic threat of deepfakes. *Philosophy & Technology, 34*(4), 623-643.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175-191.

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.

Green, Y., Gully, A., Roth, Y., Roy, A., Tucker, J. A., & Wanless, A. (2023). Evidence-based misinformation interventions: Challenges and opportunities for measurement and collaboration. *Cargengie Endowment for International Peace. January, 9.*

Greene, C. M., de Saint Laurent, C., Murphy, G., Prike, T., Hegarty, K., & Ecker, U. K. (2023). Best practices for ethical conduct of misinformation research: A scoping review and critical commentary. *European Psychologist*, *28*(3), 139.

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences, 119*(1), e2110013119.

Groh, M. (2023). Detect DeepFakes: How to Counteract Misinformation Created by AI. *MIT Media Lab*. https://www.media.mit.edu/projects/detect-fakes/overview/

GovUK. (2025, May 27). Deepfakes and media literacy. *GOV.UK*. https://www.gov.uk/government/publications/deepfakes-and-media-literacy?utm

Guay, B., Berinsky, A. J., Pennycook, G., & Rand, D. (2023). How to think about whether misinformation interventions work. *Nature Human Behaviour, 7*(8), 1231-1233.

Hameleers, M. (2023). The (un) intended consequences of emphasizing the threats of mis-and disinformation. *Media and Communication, 11*(2), 5-14.

Harjani, T., Roozenbeek, J., Biddlestone, M., van der Linden, S., Stuart, A., Iwahara, M., Piri, B., Xu, R., Goldberg, B., & Graham, M. (2022). *A Practical Guide to Prebunking Misinformation*.

Hocevar, K. P., Flanagin, A. J., & Metzger, M. J. (2014). Social media self-efficacy and information evaluation online. *Computers in Human Behavior, 39*, 254-262.

Hoes, E., Aitken, B., Zhang, J., Gackowski, T., & Wojcieszak, M. (2024). Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour, 8*(8), 1545-1553.

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.

Jungherr, A., & Rauchfleisch, A. (2024). Negative downstream effects of alarmist disinformation discourse: Evidence from the United States. *Political Behavior, 46*(4), 2123-2143.

Kalogeropoulos, A. (2017). Online News Video Consumption: A comparison of six countries. *Digital Journalism, 6*(5), 651–665. https://doi.org/10.1080/21670811.2017.1320197

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.

Ke, Z., Zhou, S., Zhou, Y., Chang, C. H., & Zhang, R. (2025, March). Detection of ai deepfake and fraud in online payments using gan-based models. In *2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE)* (pp. 1786-1790). IEEE.

Kessler, S. H., & Bachmann, E. (2022). Debunking health myths on the internet: the persuasive effect of (visual) online communication. *Journal of Public Health, 30*(8), 1823-1835.

Khan, M. R., Naeem, S., Tariq, U., Dhall, A., Khan, M. N. A., Al Shargie, F., & Al-Nashash, H. (2023, October). Exploring neurophysiological responses to cross-cultural deepfake videos. In *Companion Publication of the 25th International Conference on Multimodal Interaction* (pp. 41-45).

Khasawneh, O. Y. (2018). Technophobia without boarders: The influence of technophobia and emotional intelligence on technology acceptance and the moderating influence of organizational climate. *Computers in Human Behavior, 88*, 210-218.

Khasawneh, O. Y. (2023). Technophobia: how students' technophobia impacts their technology acceptance in an online class. *International Journal of Human–Computer Interaction, 39*(13), 2714-2723.

Kim, J. J., Soh, J., Kadkol, S., Solomon, I., Yeh, H., Srivatsa, A. V., ... & Ajilore, O. (2025). AI anxiety: A comprehensive analysis of psychological factors and interventions. *AI and Ethics*, 1-17.

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *Iscience, 24*(11).

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). *The science of fake news. Science, 359*(6380), 1094-1096.

Levine, T. R. (2014). Truth-default theory (TDT) a theory of human deception and deception detection. *Journal of Language and Social Psychology, 33*(4), 378-392.

Levine, T. R. (2022). Truth-default theory and the psychology of lying and deception detection. *Current Opinion in Psychology, 47*, 101380.

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.

Lewis, A., Vu, P., Duch, R. M., & Chowdhury, A. (2023). Deepfake detection with and without content warnings. *Royal Society Open Science, 10*(11), 231214.

Machete, P., & Turpin, M. (2020). The use of critical thinking to identify fake news: A systematic literature review. In *Conference on e-Business, e-Services and e-Society* (pp. 235-246). Springer, Cham.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum Associates.

Murphy, G., & Greene, C. M. (2023). Conducting ethical misinformation research: Deception, dialogue, and debriefing. *Current Opinion in Psychology, 54*, 101713.

Murphy, G., de Saint Laurent, C., Reynolds, M., Aftab, O., Hegarty, K., Sun, Y., & Greene, C. M. (2023). What do we study when we study misinformation? A scoping review of experimental research (2016-2022). *Harvard Kennedy School Misinformation Review*.

Newman, N., Fletcher, R., Schulz, A., Andi, S., Robertson, C. T., & Nielsen, R. K. (2022). Reuters Institute digital news report 2022. *Reuters Institute*. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/ 2022-06/Digital_News-Report_2022.pdf

O'Mahony, C., Murphy, G., & Linehan, C. (2024). True discernment or blind scepticism? Comparing the effectiveness of four conspiracy belief interventions. *advances. in/psychology, 2*, e215691.

Oremus, W., Jiménez, A., & Schaffer, A. (2025, March 4). *Melania Trump joins push for law on deepfakes, nonconsensual porn*. The Washington Post. https://www.washingtonpost.com/politics/2025/03/04/melania-trump-be-best-deepfakes-law

Pallant, J. (2020). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS.* Routledge.

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39-50.

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.

Pennycook, G., & Rand, D. G. (2022). Nudging social media toward accuracy. *The Annals of the American Academy of Political and Social Science, 700*(1), 152-164.

Prolific. (2025). Prolific. https://www.prolific.com

Rapp, A. (2020). Design fictions for learning: A method for supporting students in reflecting on technology in Human-Computer Interaction courses. *Computers & Education, 145*, 103725.

Rau, M. A., & Premo, A. E. (2025). Systematic Review of Educational Approaches to Misinformation. *Educational Psychology Review, 37*(2), 1-41.

Reuters Institute (2024). Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2024

Roozenbeek, J., Remshard, M., & Kyrychenko, Y. (2024). Beyond the headlines: On the efficacy and effectiveness of misinformation interventions. *advances.in/psychology*, *2*, e24569. https://doi.org/10.56296/aip00019

Salovaara, A., Oulasvirta, A., & Jacucci, G. (2017). Evaluation of prototypes and the problem of possible futures. In *Proceedings of the 2017 CHI conference on human factors in computing systems* (pp. 2064-2077).

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior, 149*, 107917.

Somoray, K., Miller, D. J., & Holmes, M. (2025). Human Performance in Deepfake Detection: A Systematic Review. *Human Behavior and Emerging Technologies, 2025*(1), 1833228.

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, *31*(1), 137-149.

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., ... & Zaffar, M. F. (2021, May). Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (pp. 1-16).

Ternovski, J., Kalla, J., & Aronow, P. (2022). The negative consequences of informing voters about deepfakes: evidence from two survey experiments. *Journal of Online Trust and Safety, 1*(2).

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *Plos one, 18*(10), e0291668.

Twomey, J., O'Mahony, C., Linehan, C., & Murphy, G. (2025). Do No Harm: Designing Interventions for Deepfake-Fueled Conspiracy Theories. *Interactions, 32*(3), 30-33.

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*(1), 2056305120903408.

van Doorn, M. (2023). Advancing the debate on the consequences of misinformation: clarifying why it's not (just) about false beliefs. *Inquiry*, 1-27.

Van Duyn, E., & Collier, J. (2019). Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass communication and society, 22*(1), 29-48.

Vivion, M., Anassour Laouan Sidi, E., Betsch, C., Dionne, M., Dubé, E., Driedger, S. M., ... & Canadian Immunization Research Network (CIRN). (2022). Prebunking messaging to inoculate against COVID-19 vaccine misinformation: an effective strategy for public health. *Journal of Communication in Healthcare, 15*(3), 232-242.

Wang, Y. Y., & Wang, Y. S. (2022). Development and validation of an artificial intelligence anxiety scale: An initial application in predicting motivated learning behavior. *Interactive Learning Environments, 30*(4), 619-634.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review, 9(*11).

Whitehead, H. S., French, C. E., Caldwell, D. M., Letley, L., & Mounier-Jack, S. (2023). A systematic review of communication interventions for countering vaccine misinformation. *Vaccine, 41*(5), 1018-1034.

Xu, Z., Wen, X., Zhong, G., & Fang, Q. (2025). Public perception towards deepfake through topic modelling and sentiment analysis of social media data. *Social Network Analysis and Mining, 15*(1), 16.

Zimmerman, T., Njeri, M., Khader, M., & Allen, J. (2022). Default to truth in information behavior: a proposed framework for understanding vulnerability to deceptive information. *Information and Learning Sciences, 123*(1/2), 111-126.

Chapter 6 Discussion

6.1 Thesis Overview

The overall research question for this thesis was: how do deepfakes affect beliefs, memories, and behaviours and how can the effects of misleading deepfakes be mitigated? By answering this question, the thesis contributes to the field by providing empirical evidence on the real-world impacts of deepfakes and by developing actionable guidance for reducing misinformation-related harms. The four studies in this thesis served to address different elements of this research question. Chapters 2, 3, and 4 contributed to our understanding of deepfake effects on beliefs, memories, and behaviours through quantitative and qualitative means. Chapter 5 examined the best practices for mitigating negative effects of misleading deepfake exposure.

Chapter 2 presented a scoping review that established the context for the remainder of the thesis by collating and criticising existing empirical deepfake research. I asked foundational questions about deepfakes such as 1) How does deepfake exposure affect beliefs, memories, and behaviours? 2) Are deepfakes uniquely persuasive compared to existing forms of misinformation? 3) What factors affect susceptibility to deepfake-based misinformation? and 4) What kind of deepfake technology is being used in modern empirical research? The scoping review yielded several important findings, such as quantifying deepfake effects on beliefs, memories, and behaviours, and revealing the ambiguity in evidence supporting the existence of unique deepfake effects. The review in Chapter 2 served to formally de-sensationalize the deepfake debate. It strategically moved the field past conjecture by demonstrating that the alarmist narratives driving much of the early attention were unsupported by empirical evidence. Critically, this foundational work did not just define the extant research, but it advanced the discipline's understanding by clearly identifying problems with current empirical deepfake

research, thereby providing methodological guidelines for the subsequent studies in this thesis. Notably, the scoping review in Chapter 2 did not find any qualitative studies or studies that examined the lived everyday experience of deepfakes.

Findings from Chapter 2 directly led to Chapter 3 where I gained an understanding of real-world experiences of deepfakes using thematic analysis to examine the response of commenters to a popular YouTube music video featuring deepfakes. This study was essential to providing real-world ecological context to our understanding of deepfake effects which were unattainable through laboratory paradigms. I found that in a naturalistic entertainment environment, people generally respond positively to deepfakes citing their creative and meaningful usage. However, alarmist narratives around deepfakes can be seen among everyday audiences despite the lack of empirical support as shown in the thematic analysis. The study also reveals that many viewers misinterpret or rationalize and minimalize deepfake content. These findings critically inform intervention design, suggesting that educational strategies must account for naturalistic audience responses and counteract alarmist myths about deepfakes, in order to accurately support discernment, reduce unnecessary anxiety, and preserve trust in authentic information.

Chapter 4 directly addresses the findings from the scoping review in Chapter 2 using quantitative methods and furthers our understanding of deepfakes by rigorously testing deepfake impacts on beliefs, memories, and behaviours in an ecologically valid political environment. By testing participants in realistic settings, such as during a significant election, this study provides reliable evidence that advances the broader academic understanding of deepfake impacts. The multi-experiment online survey study in Chapter 4 ultimately challenges the prevalent alarmist narratives around deepfake effects. Chapter 4 was guided by the research question: Are deepfakes uniquely effective at influencing political beliefs, memories, and behaviours of citizens during an ongoing election campaign? Based on the findings from the

scoping review, I conducted this study with comparator conditions to identify unique effects of deepfakes, I used custom-made deepfakes to avoid limiting my study design, I conducted the experiments with participants who were imminently experiencing an important election to examine responses to deepfakes under real-world conditions, and I examined potential delayed effects of deepfake exposure such as the "sleeper effect" and the "illusory truth" effect. Results indicate that deepfakes could affect beliefs, memories, and behaviours in a political environment, but they were not uniquely persuasive compared to existing formats of misinformation such as text and audio and did not lead to lasting delayed effects, further challenging the alarmist narratives around deepfakes. This finding provides evidence that directly contradicts years of discussion projecting the negative consequences of deepfake exposure in politics (Albahar & Almalki, 2019, Islam et al., 2024; Pantserev, 2020; Westerlund, 2019).

Finally, Chapter 5 concludes the thesis by addressing the second part of the overarching research question concerning strategies for mitigating the harms caused by deepfakes. It is possible that the prevailing alarmist narratives around deepfakes as shown in Chapters 2, 3, and 4, could undermine epistemic trust in individuals who may feel that they cannot trust video information for fear of being deceived by a deepfake. To balance these two positions, I designed an educational deepfake intervention aimed at educating people about deepfakes while mitigating the negative effects of alarmist narratives around deepfake such as fearmongering and epistemic pollution. As established in Chapter 4, deepfakes are not inherently more harmful than other forms of misinformation, which suggests that we do not need to reinvent educational strategies when addressing deepfakes. Rather, we can adapt existing media literacy frameworks to incorporate the unique characteristics of deepfakes such as their audio-visual nature and potential to undermine video evidence. In an online survey experiment, I demonstrated that existing best practices for educating people about

misinformation may lead to the undermining of trust in authentic information. I then designed an educational deepfake intervention and compared its efficacy to interventions emulating existing educational efforts such as interventions with alarmist narratives or detection tips. My findings show that an evidence-based intervention can reduce anxiety around deepfakes and perceptions of emerging technologies, as well as improve discernment of authentic videos after a delayed follow-up thus preserving epistemic trust in real videos. The difficulty in improving deepfake discernment suggests that the ability to detect deepfakes may be influenced by a range of factors beyond the visual quality of the deepfake itself. These factors could include individual differences such as prior beliefs, cognitive ability, and analytical reasoning, which I plan to explore in future research. In relation to this thesis, I have addressed all of the research questions posed in the introduction regarding the understanding of deepfake effects and developing strategies for mitigating the negative consequences of deepfake exposure.

6.2 Contributions

6.2.1 Situating Deepfake Effects in Real-World Contexts

The first contribution of this thesis is providing a more accurate understanding of deepfake exposure effects in real-world contexts. Much of existing empirical misinformation research takes place in artificial, controlled, and highly simplified conditions which do not resemble how people truly encounter misinformation in the world (Murphy et al., 2023). Situating the effects of deepfakes within applied settings allows us to understand deepfake effects more precisely as the research findings from this thesis reflect the experience of deepfakes in everyday life. By examining deepfake exposure in naturalistic environments such as in YouTube comments sections in Chapter 3 or during an ongoing election in Chapter 4, this thesis extends its findings beyond laboratory settings, making them more relevant and

applicable to everyday experiences and living. Current empirical deepfake research often faces limitations to ecological validity. For example, many experimental deepfake research use deepfakes of local actors or anonymous individuals from pre-made deepfake datasets as experimental stimuli to see how deepfake exposure affects beliefs, memories, and behaviours (Huang & Hu, 2025; Sharma et al., 2023; Ternovski et al., 2021). Participants would be tasked with discerning between real videos and deepfakes of anonymous individuals or asked questions about their perception of fictional persons depicted as a deepfake. These studies may be using fictional characters for deepfake stimuli to avoid ethical issues or to control for the prior beliefs of individuals. However, using fictional characters that participants do not recognize may compromise the ecological validity of these studies, as real-world deepfakes are typically made using well-known public figures or celebrities to maximize their emotional impact and evocativeness (Chesney & Citron, 2019) and therefore participant's prior beliefs are a crucial element to incorporate in deepfake research. Studying decontextualized and unfamiliar deepfakes limits the generalizability and real-world applicability of the experimental findings potentially leading to an inaccurate understanding of the effects of deepfakes.

Furthermore, ecologically valid research, such as the thematic analysis in Chapter 3, helped inform the design of experimental research. The study revealed that alarmist and dystopian narratives about deepfakes extend beyond academic and policy discourse to everyday audiences, indicating that public concern is widespread yet often unfounded. This finding was crucial as it provides us with an ecologically valid understanding of how the everyday person experiences deepfakes, which informed intervention-design in Chapter 5 where I directly attempted to design an educational intervention to counteract anxieties around deepfake capabilities and perceptions of dystopian technologies. Moreover, many viewers were unaware of the technology and responded by rationalizing deepfake visuals, doubting their own

perceptions, or minimizing facial alterations. While these reactions occurred in an entertainment context where one would not expect misinformation, they highlight potential vulnerabilities to deception and the importance of fostering accurate understanding. These insights directly inform the design of educational interventions. Rather than relying on alarmist framing, interventions should focus on enhancing deepfake literacy, critical thinking, and discernment. By aligning intervention strategies with naturalistic audience responses, the ecologically valid findings from Chapter 3 not only informed the design of the experimental studies but also provided a more accurate understanding of how people naturally respond to deepfakes. These insights proved instrumental in identifying necessary areas to address when designing educational interventions.

Studies in this thesis were designed to ensure that the findings could be generalised to real-world settings. The qualitative study in Chapter 3 helped us understand people's lived experiences of deepfakes, the multi-experiment study in Chapter 4 specifically sampled participants who were imminently undergoing an important election, and the study in Chapter 5 examined how guidelines that were currently being used to design misinformation interventions could have negative downstream effects on epistemic trust. This thesis contributes valuable insights to misinformation researchers by deepening their understanding of deepfake effects and multi-modal misinformation in an applied context. The emphasis on ecological validity throughout this thesis enhances the credibility of our findings, ensuring that the results offer a realistic and everyday perspective on the effects of deepfakes, which researchers can use as a foundation for future studies. This thesis should also prove useful for policymakers when designing legislation for deepfakes. For example, without accurate portrayals of deepfake effects, policymakers may run the risk of excessively or insufficiently regulating deepfakes, which could result in unintended problems such as unjustly limiting free speech. By demonstrating how deepfakes realistically influence beliefs, memories, and

behaviours, this thesis equips policymakers with an empirically grounded framework for crafting balanced, evidence-based interventions.

6.2.2 Demonstrating the Ethical Conduct of Deepfake Misinformation Research

My thesis makes a significant contribution to the growing body of empirical research on deepfakes by demonstrating that it is possible to conduct ethical experimental research on deepfakes in line with recent advancements in ethical meta-science. As outlined in Chapter 2, many deepfake experiments rely on using pre-existing popular deepfakes or deepfakes generated from custom-made datasets. This practice is often motivated by ethical considerations, such as the desire to avoid disseminating new deepfakes on public platforms, the potential to influence individuals' beliefs or behaviours through deepfake-driven misinformation or concerns about defaming the person depicted in the deepfake.

While these ethical concerns are important, I argue that they introduce limitations to the design of deepfake experiments. Specifically, they can restrict research to controlled laboratory environments, leading to findings that may not be easily generalizable to real-world situations. Additionally, the hesitancy to create custom deepfakes for experimental purposes can prevent researchers from exploring specific research questions they would otherwise wish to investigate. Consequently, ethical challenges can act as a barrier to addressing more meaningful questions about deepfakes, such as their relevance in real-world contexts or how individuals' pre-existing beliefs about a person might influence their perception of a deepfake.

In this context, the present thesis makes a meaningful contribution to future research by demonstrating that important questions regarding deepfakes and misinformation can be investigated rigorously while maintaining adherence to ethical standards. This is exemplified through the multi-experiment studies presented in Chapters 4 and 5, where I adhered to ethical

meta-science principles for misinformation studies (Greene & Murphy, 2023; Murphy & Greene, 2023), including proper debriefing of participants and conducting follow-up studies to ensure that participants were not adversely affected by their involvement in the experiments. I show that with careful planning, including stringent debriefing procedures and follow-up studies, researchers can explore the impact of deepfakes in ways that minimize potential harm to participants and the individuals depicted. This approach allows for more comprehensive, real-world relevant studies of deepfakes without compromising the ethical integrity of the research process.

6.2.3 Challenging Alarmist Narratives around Emerging Technologies through empirical study

Identifying, understanding, and challenging the prevailing alarmist narratives around deepfakes is a consistent contribution throughout all studies in this thesis. Narratives become alarmist when they present the capabilities and spread of deepfakes in an exaggerated negative light, unsupported by empirical evidence. These accounts distort public perception by fuelling unwarranted fear and panic about the prevalence and potential harms of the technology. This moral panic and growing technophobia may also deter individuals from embracing future innovations that could offer significant benefits (Khasawneh, 2023). Additionally, alarmist portrayals often emphasise specific concerns, such as the role of deepfakes in hyperbolic harms like political manipulation or threats to national security (Chesney & Citron, 2019; Tang et al., 2023). By focusing exclusively on one aspect of the issue, these narratives divert attention from other, possibly more realistic concerns, like the harmful use of deepfakes in pornography (Karasavva & Noorbhai, 2021). Moreover, exaggerating the spread and impact of the technology erodes public trust in credible information despite the narratives lacking in evidence to substantiate their claims. It is crucial to challenge these exaggerated accounts, as they hinder

the adoption of beneficial technologies, misdirect attention from more pertinent threats, and misinform the public.

My thesis challenges the widespread narratives around deepfakes and shows that they are unsubstantiated by empirical evidence. At the start of my thesis in October 2021, I perceived deepfakes as being the next frontier of misinformation due to its hyper-realistic audio-visual portrayal of reality. Certainly, this was reflected in the literature at the time with many papers discussing and predicting the dangerous consequences of deepfake technology (Chesney & Citron, 2019; Westerlund, 2019). However, the foundation of this narrative was challenged early in my thesis during the scoping review in Chapter 2, which revealed that the prevailing narratives at the time were largely speculative and lacked empirical support. Existing research was divided on whether deepfakes were uniquely harmful and dangerous compared to existing formats of misinformation (Hameleers, 2024; Lee & Shin, 2022; Murphy & Flynn, 2022). This became the inspiration for the study in Chapter 4 where I decided to directly challenge the claims around deepfakes and found that deepfakes indeed did not differ from pre-existing formats of misinformation, changing my perspective from questioning the claims about deepfakes to concretely identifying them as alarmist and unfounded. Discovering that the discourse surrounding deepfakes was deeply exaggerated marked a significant shift in this thesis. Rather than focusing exclusively on mitigating the direct harms of misleading deepfakes, the interventions in Chapter 5 also needed to address the indirect harms arising from alarmist discourse surrounding deepfakes. Ultimately, my thesis unearthed the speculative foundation of narratives around deepfakes, confirmed that the narratives are alarmist and dystopian, and mitigated the indirect consequences of alarmist narratives.

Challenging alarmist discourse around deepfakes is a contribution that is especially relevant for several stakeholders. For misinformation researchers, it encourages a shift from viewing deepfakes solely as a dangerous consequence in politics and national security. Rather

than focusing primarily on the political risks, researchers might consider exploring other, more prevalent harms, such as those associated with deepfake pornography and online bullying or harassment (Jackson et al., 2025; Karasavva & Noorbhai, 2021). For public educators, this work highlights the importance of avoiding the spread of alarmist narratives about deepfakes, which could undermine trust in authentic content (Hameleers, 2023). Instead, educators should aim to provide a more accurate understanding of the actual capabilities and concerns of this technology. Finally, media outlets stand to benefit by refraining from sensationalizing deepfakes and instead focusing on reporting the technology's real capabilities, thereby reducing the fear and anxiety that currently surrounds this technology.

6.2.4 Identifying Important Elements of Educational Deepfake Interventions

This thesis serves as a foundational investigation of educational deepfake interventions. It is necessary to educate the public about deepfakes to better equip them with knowledge to identify deepfakes in their everyday lives. Findings from Chapter 3 evidently show how lack of awareness and literacy around deepfakes may lead to erroneous interpretations of deepfakes such as assuming the visuals are truthful or rationalizing the deepfakes with familiar concepts. Despite the importance in exploring optimal methods of educating the public about deepfakes, to date, only one other study has explored elements of effective educational deepfake interventions (Huang & Hu, 2025). Huang & Hu (2025) examined how media literacy interventions varying by formatting (text, infographic or video) and type (detection tips or warnings) about deepfakes affected outcome measures such as discernment, perceived AI literacy, and discernment self-efficacy. While they did find that literacy interventions focusing on detection tips boosted deepfake discernment, the videos used to measure discernment were

of anonymous deepfaked individuals from publicly available datasets, whereas most notable examples of deepfakes circulating in real-world contexts depict well-known or public figures. Deepfakes of familiar individuals are likely to elicit stronger emotional responses, activate pre-existing beliefs, and engage recognition-based cognitive processes based on familiarity of the individual alone (Sugiura, 2014), all of which can influence susceptibility to misinformation and therefore efficacy of the intervention. The limitations to ecological validity and general lack of research investigating deepfake intervention design necessitates further contributions in this area.

Chapter 5 of this thesis directly contributes to addressing these issues as it compares and situates the efficacy of different deepfake interventions in real-world contexts by using custom-made deepfakes of prominent political figures as discernment stimuli, deepfakes which could reasonably be seen on everyday online platforms (Walker et al., 2024). Chapter 5 informs the first steps to designing an educational intervention that specifically addresses deepfake harms by finding that an evidence-based intervention can improve discernment of authentic information and reduce the anxiety around deepfakes and negative perceptions of future technologies. These findings should prove valuable for researchers and public educators alike. This research provides future scholars with a deeper understanding of the theoretical foundations for educating individuals about deepfakes. For instance, findings from Chapter 5 indicate that discernment is influenced not only by the technical quality of a deepfake but also by individuals' prior beliefs, which is in line with existing misinformation research (Ludwig & Sommer, 2024; Rio & Waldzus, 2023; Roozenbeek et al., 2022). For public educators, these results offer guidance for designing effective interventions, including the recommendation to avoid alarmist messaging and to reassure audiences that deepfakes are neither ubiquitous nor uniquely potent or threatening, a recommendation that is also echoed by general misinformation literature (Hameleers, 2023). Despite being unable to boost deepfake

discernment, this thesis lays the groundwork for understanding effective strategies to maintain the public's trust in real videos while simultaneously educating and reassuring the public about deepfakes.

6.3 Strengths

The simplistic and direct approach I take to investigating deepfake exposure effects represents a particular strength of this thesis. Chapter 4 establishes fundamental causal relationships between deepfake exposure and subsequent beliefs, memories, and behavioural intentions. Existing empirical work contrasts with my approach by emphasising different outcome measures such as perceived accuracy, credibility, and realism of the deepfake stimuli (Ching et al., 2025). My approach explores the direct cognitive consequences of deepfake exposure rather than capturing the subjective impressions of deepfakes by participants, allowing me to make inferences about how deepfake exposure affects real-world beliefs, memories, and behaviours. However, there is evidence to suggest that once-off misinformation exposure is unlikely to shift attitudes and behaviours due to deeply entrenched prior beliefs, lack of relevancy of the misinformation, and small shifts to beliefs does not lead to behavioural changes (Greene et al., 2025; Lazer et al., 2018) but this research has not been explored in the context of deepfakes. It remains a worthwhile endeavour to measure simple deepfake exposure effects despite evidence suggesting that it might not influence attitudes or behaviours because it is crucial to establish a baseline for how people would react to deepfakes first before future researchers can explore more complex research designs and outcome measures.

Another key strength of this thesis was its emphasis on assessing deepfake-related harms beyond the direct effects on beliefs, memory, and behaviour. In addition to investigating the core cognitive consequences of deepfake exposure, I also exposed the indirect harms surrounding deepfakes such as the undermining of epistemic trust, stress and anxiety around

deepfakes, and negative perceptions of future emerging technologies as demonstrated in Chapter 5. These dimensions of deepfake impact have been relatively understudied and remain poorly understood, yet they constitute meaningful forms of harm associated with this technology (De Ruiter, 2021; Twomey et al., 2023). By incorporating these elements into the thesis, I offer a multi-faceted account of the various harms deepfakes may produce and demonstrate that deepfakes can exert negative effects even when individuals are not directly exposed to them.

Finally, a notable strength of the thesis lies in the methodological rigour throughout each experiment. Each of the experimental studies in Chapter 4 and 5 were sufficiently powered to detect small effects to ensure greater confidence in null results (Gaskill & Garner, 2020). Across eight experiments, a total of 7,666 participants were recruited via Prolific, supplemented with a small amount of convenience sampling, providing a robust sample that supports the credibility and reliability of the findings of this thesis. The experiments reported in Chapters 4 and 5 were carefully controlled and incorporated follow-up assessments to examine delayed effects of deepfake or intervention exposure, an aspect often overlooked in prior misinformation research (Murphy et al., 2023). Every experiment concluded with a rigorous, evidence-informed debriefing procedure (Greene & Murphy, 2023; Murphy & Greene, 2023) to ensure that participants that took part in my studies were not adversely affected by deepfakes and/or misinformation, including a follow-up experiment to assess potential negative impacts of taking part in my studies. Collectively, these methodological strengths reflect careful attention to experimental design, ethical conduct, and the rigour of the thesis overall.

6.4 Limitations

Methodological Limitations

A methodological limitation of this research relates to the culture of the samples used throughout the experiments in this thesis. Much existing misinformation research, including deepfake studies, disproportionately relies on American or Western, White populations (Dobber et al., 2021; Hameleers, 2024; Murphy et al., 2023), which is problematic because it limits the generalizability of misinformation research to western societies (Henrich et al., 2010). I also sampled participants from the US and Ireland for practical reasons such as the U.S. having a sufficiently large number of participants on Prolific for the well-powered studies in this thesis, people in the U.S. undergoing an important election, and the primary language used in the U.S. and Ireland being English. However, I recognize that this choice limits the ability to generalize the results across diverse cultural contexts. This limitation is particularly relevant for deepfakes and AI-based misinformation, as global survey data indicate that views toward these technologies differ markedly between Western and Asian populations. For example, a 29-country IPSOS survey revealed that respondents from different regions differ in how seriously they regard the threat posed by AI-generated misinformation, as well as in their self-assessed ability to discern manipulated content (IPSOS, 2023). On average, 40% of Japanese and South Korean citizens are not confident that they can discern between real and fake news compared to an average of 60-65% in the United States and European countries (IPSOS, 2023). Some studies indicate that individuals' confidence in their ability to distinguish between true and false information can impact their susceptibility to misinformation. Specifically, overconfidence in this ability may actually increase vulnerability to misinformation (Lyons et al., 2021). As such, the predominantly western sample from the US and Ireland in this study may respond differently to the deepfake stimuli than participants from other cultural backgrounds, such as those from Asian countries, therefore limiting the broader applicability of the findings.

Theoretical Limitations

This thesis also faces theoretical limitations. Throughout the thesis, I rely on several theories to understand and possibly explain how deepfake exposure affects beliefs, memories, and behaviours such as motivated reasoning and Dual Process Theory (Kahan, 2013; Kahneman, 2011), the misinformation effect (Loftus & Palmer, 1974), and the Theory of Planned Behaviour (Ajzen, 1991). While some of these theories were not originally applied to misinformation contexts, many studies have repeatedly and successfully tested these theories in misinformation environments to understand how textual or pictorial information affects beliefs, memories, and behaviours (Murphy et al., 2019; Pennycook & Rand, 2021; Pundir et al., 2021). Notably, these theories and concepts were not designed for highly realistic multimodal misinformation and have also not been extensively tested within this context. As such, I made assumptions about how these theories would apply to deepfakes throughout the thesis. For example, I use established literature on political persuasion and research highlighting the resilience of attitudes to change due to deeply ingrained social identities, values, and political motivations as a potential explanation for the null results from Chapter 4. However, it is possible that there were mechanism-level differences despite the same outcome. Deepfakes may have generated greater sensory engagement, but this did not translate to attitudinal or behavioural changes due to limiting factors related to the audio-visual element of deepfakes (such as the lack of immersion or the uncanny valley effect where that near-perfect but imperfect human likeness can provoke negative affective responses (Weisman & Pena, 2021)) thus counteracting potential persuasive advantages. Therefore, while the application of established theories in this thesis is justified and consistent with my results, it is important to acknowledge that these theories were not built with deepfakes in mind and definitive conclusions require more extensive research.

Another limitation is that the theoretical lens used throughout this thesis to understand deepfake effects emphasises individual cognition (e.g. motivated reasoning, dual processing theory, misinformation effect), which under-represents the multitude of individual, metacognitive, and socio-technical factors that influence misinformation susceptibility (Lazer, 2018). In the context of deepfakes, individual factors such as one's familiarity with the deepfaked subject could impact their responses to viewing deepfakes (Ahmed & Chua, 2023). For example, Ahmed & Chua (2023) found that increased familiarity with the deepfaked subject elicits fluency rendering participants more likely to believe the deepfake as true in a sort of "truthiness" effect. Metacognitive factors such as one's perceived ability to detect misinformation or confidence in discerning between real and false information can similarly affect susceptibility to misinformation with overconfidence in one's discernment ability associated with greater susceptibility to misinformation (Rapp et al., 2024). Finally, socio-technical factors like how a person responds to counter-misinformation precautions on online platforms such as fact-checking or community notes, or responses to platform metrics such as follower count, number of view, and the number of likes a video receives may also affect a person's credibility judgments of deepfakes. Jin et al. (2025) demonstrated that participants perceived deepfakes as more credible if the deepfake video was presented alongside positive video cues such as high number of views and likes. Evidently, these factors lay outside the scope of my thesis but are important factors to consider when explaining deepfake susceptibility which I notably was unable to comprehensively incorporate into my studies.

6.5 Future Directions

Future research should look beyond the prevalent alarmist discourse surrounding deepfake-related threats in politics and national security and instead focus on the substantiated and realistic harms associated with deepfakes. Numerous documented cases demonstrate the

significant impact of deepfakes on individuals, including their use in non-consensual deepfake pornography (Twomey et al., 2025) and in acts of bullying and harassment (Jackson et al., 2025). Yet these well-evidenced harms have been overshadowed by a disproportionate research focus on speculative political consequences that have not been shown to affect electoral outcomes (Dennehy, 2024). Building on the everyday lived-experience approach to understanding deepfakes outlined in Chapter 3, additional qualitative research is recommended to examine how individuals affected by deepfake pornography or by bullying and harassment perceive and navigate these experiences. Research has started to examine the experiences of individuals targeted by non-consensual deepfake pornography (Flynn et al., 2025; Twomey et al., 2025), however, comparatively little attention has been given to understanding experiences of cyberbullying facilitated by deepfakes. Such work has the potential to offer policymakers more nuanced, evidence-informed insights to guide the development of effective policies and legislation addressing deepfakes.

Research should also be extended in the area of educational deepfake interventions. As outlined in Chapter 5, there is a notable scarcity of studies that clearly identify the components of effective deepfake interventions. Without such evidence, it remains unclear how to design interventions that enhance deepfake literacy, preserve epistemic trust in authentic information, and strengthen individuals' ability to discern manipulated content, which is arguably the most critical outcome of any intervention (Huang et al., 2024). Findings from Chapter 5 indicate that recognising visual artifacts alone does not fully account for susceptibility to or detection of deepfakes, suggesting that additional factors shape how people engage with these media. These may include individual differences such as prior beliefs, social identities, cognitive abilities, and digital literacy, as well as demographic variables like age and gender. Future research is recommended to examine how such individual and demographic factors influence

responsiveness to educational deepfake interventions, with the goal of optimising strategies for effectively educating the public about deepfakes.

6.6 Conclusion

This thesis provides a comprehensive examination of how deepfakes affect beliefs, memories, and behaviours, and evaluates strategies to mitigate potential harms. Across four chapters, I demonstrated that while deepfakes can influence cognition and behaviour, they are not uniquely persuasive compared to other forms of misinformation, such as text or audio, challenging prevailing alarmist narratives. By integrating qualitative and quantitative approaches, I situate deepfake effects in ecologically valid, real-world contexts highlighting the nuanced ways individuals engage with deepfake content.

The thesis also establishes that deepfake research can be conducted ethically without compromising methodological rigor, using carefully controlled experiments, proper debriefing, and follow-up assessments to minimize harm to participants. Importantly, these studies show that educational interventions can effectively enhance discernment of authentic content, reduce anxiety, and preserve epistemic trust, particularly when interventions avoid alarmist messaging and acknowledge the multi-modal nature of deepfakes.

By challenging exaggerated narratives, this work offers a more accurate and evidence-based understanding of deepfake effects, providing insights for researchers, educators, policymakers, and media practitioners. It highlights the need for further research on individual and demographic factors that shape responsiveness to interventions. Overall, this thesis contributes to a realistic, applied, and ethically grounded understanding of deepfakes, paving the way for informed strategies that protect individuals from harm while fostering trust in authentic media.

6.7 References

Ahmed, S., & Chua, H. W. (2023). Perception and deception: Exploring individual responses to deepfakes across different modalities. *Heliyon*, *9*(10).

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, *50*(2), 179-211.

Albahar, M., & Almalki, J. (2019). Deepfakes: Threats and countermeasures systematic review. *Journal of Theoretical and Applied Information Technology*, *97*(22), 3242-3250.

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, *107*, 1753.

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

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & De Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes?. *The International Journal of Press/Politics, 26*(1), 69-91.

Flynn, A., Powell, A., Eaton, A., & Scott, A. J. (2025). Sexualized deepfake abuse: Perpetrator and victim perspectives on the motivations and forms of non-consensually created and shared sexualized deepfake imagery. *Journal of Interpersonal Violence*, 08862605251368834.

Gaskill, B. N., & Garner, J. P. (2020). Power to the people: power, negative results and sample size. *Journal of the American Association for Laboratory Animal Science*, *59*(1), 9-16.

Greene, C. M., & Murphy, G. (2023). Debriefing works: Successful retraction of misinformation following a fake news study. *PloS one, 18*(1), e0280295.

Greene, C. M., Brassil, M., Bryan, E., Glass, S. B., O'Connor, R., O'Keeffe, V., ... & Murphy, G. (2025). Evaluating real-world effects of one-off fake news exposure. *Scientific Reports, 15*(1), 1-19.

Hameleers, M. (2023). The (un) intended consequences of emphasizing the threats of mis-and disinformation. *Media and Communication, 11*(2), 5-14.

Hameleers, M. (2024). Cheap versus deep manipulation: The effects of cheapfakes versus deepfakes in a political setting. *International Journal of Public Opinion Research, 36*(1), edae004.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and brain sciences, 33*(2-3), 61-83.

Huang, G., Jia, W., & Yu, W. (2024). Media literacy interventions improve resilience to misinformation: a meta-analytic investigation of overall effect and moderating factors. *Communication Research*, 00936502241288103.

Huang, G., & Hu, B. (2025). "A Warning is Not Enough. Teach Me How to Spot Deepfakes.": Testing Media Literacy Interventions for Combating Deepfakes. *Science Communication*, 10755470251382889.

IPSOS. (2023) A 29-country global survey. https://www.ipsos.com/sites/default/files/ct/news/documents/2023-11/Ipsos_Global_Views_on_AI_and_Disinformation_full_report.pdf

Islam, M. B. E., Haseeb, M., Batool, H., Ahtasham, N., & Muhammad, Z. (2024). AI threats to politics, elections, and democracy: a blockchain-based deepfake authenticity verification framework. *Blockchains*, *2*(4), 458-481.

Jackson, B. A., Diliberti, M. K., & Moore, P. (2025, September 24). Artificially Intelligent Bullies: Dealing with Deepfakes in K–12 Schools. Rand.org; *RAND Corporation*. https://www.rand.org/pubs/research_reports/RRA3930-5.html

Jin, X., Zhang, Z., Gao, B., Gao, S., Zhou, W., Yu, N., & Wang, G. (2025). Assessing the perceived credibility of deepfakes: The impact of system-generated cues and video characteristics. *new media & society*, *27*(3), 1651-1672.

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

Kahneman, D. (2011). Thinking, fast and slow. *Farrar, Straus and Giroux*.

Khasawneh, O. Y. (2023). Technophobia: how students' technophobia impacts their technology acceptance in an online class. *International Journal of Human–Computer Interaction, 39*(13), 2714-2723.

Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., ... & Zittrain, J. L. (2018). The science of fake news. *Science, 359*(6380), 1094-1096.

Lee, J., & Shin, S. Y. (2022). Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news. *Media Psychology, 25(*4), 531-546.

Loftus, E. F., & Palmer, J. C. (1974). Reconstruction of automobile destruction: An example of the interaction between language and memory. *Journal of verbal learning and verbal behavior, 13*(5), 585-589.

Ludwig, J., & Sommer, J. (2024). Mindsets and politically motivated reasoning about fake news. *Motivation and Emotion*, *48*(3), 249-263.

Lyons, B. A., Montgomery, J. M., Guess, A. M., Nyhan, B., & Reifler, J. (2021). Overconfidence in news judgments is associated with false news susceptibility. *Proceedings of the National Academy of Sciences, 118*(23), e2019527118.

Murphy, G., Loftus, E. F., Grady, R. H., Levine, L. J., & Greene, C. M. (2019). False memories for fake news during Ireland's abortion referendum. *Psychological science, 30*(10), 1449-1459.

Murphy, G., & Flynn, E. (2022). Deepfake false memories. In *Memory Online* (pp. 112-124). Routledge.

Murphy, G., & Greene, C. M. (2023). Conducting ethical misinformation research: Deception, dialogue, and debriefing. *Current Opinion in Psychology, 54*, 101713.

Murphy, G., de Saint Laurent, C., Reynolds, M., Aftab, O., Hegarty, K., Sun, Y., & Greene, C. M. (2023). What do we study when we study misinformation? A scoping review of experimental research (2016-2022). *Harvard Kennedy School Misinformation Review*.

Pantserev, K. A. (2020). The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. In *Cyber defence in the age of AI, smart societies and augmented humanity* (pp. 37-55). Cham: Springer International Publishing.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in cognitive sciences, 25*(5), 388-402.

Pundir, V., Devi, E. B., & Nath, V. (2021). Arresting fake news sharing on social media: a theory of planned behavior approach. *Management Research Review, 44*(8), 1108-1138.

Rapp, D. N., & Withall, M. M. (2024). Confidence as a metacognitive contributor to and consequence of misinformation experiences. *Current opinion in psychology, 55*, 101735.

Rijo, A., & Waldzus, S. (2023). That's interesting! The role of epistemic emotions and perceived credibility in the relation between prior beliefs and susceptibility to fake-news. *Computers in Human Behavior*, *141*, 107619.

Roozenbeek, J., Maertens, R., Herzog, S. M., Geers, M., Kurvers, R., Sultan, M., & Van Der Linden, S. (2022). Susceptibility to misinformation is consistent across question framings and response modes and better explained by myside bias and partisanship than analytical thinking. *Judgment and Decision Making*, *17*(3), 547-573.

Sharma, I., Jain, K., Behl, A., Baabdullah, A., Giannakis, M., & Dwivedi, Y. (2023). Examining the motivations of sharing political deepfake videos: the role of political brand hate and moral consciousness. *Internet Research, 33*(5), 1727-1749.

Stockwell, S. (2024). AI-Enabled Influence Operations: Threat Analysis of the 2024 UK and European Elections. Centre for Emerging Technology and Security. https://cetas.turing.ac.uk/publications/ai-enabled-influence-operations-threat-analysis-2024-uk-and-european-elections

Sugiura, M. (2014). Neuroimaging studies on recognition of personally familiar people. *Frontiers in Bioscience, 19*(4), 672.

Tang, Z., Yin, S. X., & Goh, D. H. L. (2023, July). Understanding major topics and attitudes toward deepfakes: An analysis of news articles. In *International Conference on Human-Computer Interaction* (pp. 337-355). Cham: Springer Nature Switzerland.

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*.

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *Plos one*, *18*(10), e0291668.

Twomey, J., Foley, S., Robinson, S., Quayle, M., Aylett, M. P., Linehan, C., & Murphy, G. (2025). " What do you expect? You're part of the internet": Analyzing Celebrities' Experiences as Usees of Deepfake Technology. *arXiv preprint arXiv:2507.13065*.

Walker, C. P., Schiff, D. S., & Schiff, K. J. (2024, March). Merging AI incidents research with political misinformation research: introducing the political Deepfakes incidents database. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 21, pp. 23053-23058).

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review, 9*(11).