CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	Socially (de)constructing deepfakes: understanding socio-technical interests and concerns in deepfake media
Authors	Twomey, John
Publication date	2025-11-30
Original Citation	Twomey, J. 2025. Socially (de)constructing deepfakes: understanding socio-technical interests and concerns in deepfake media. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2025, John Twomey. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-14 16:13:01
Item downloaded from	https://hdl.handle.net/10468/18856



UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork



Socially (De)constructing Deepfakes: Understanding socio-technical interests and concerns in deepfake media.

Thesis presented by

John Twomey

[ORCID: 0000-0002-5493-8518]

for the degree of

Doctor of Philosophy in Applied Psychology

University College Cork

School of Applied Psychology

College of Arts, Celtic Studies, and Social Sciences

Head of School: Professor Christopher McCusker

Supervisors: Prof. Gillian Murphy, Dr. Conor Linehan

November 2025

Table of Contents

LIST OF FIGURES II

LIST OF TABLES III

ACKNOWLEDGEMENTS IV

DECLARATION OF ORIGINALITY VI

PUBLICATIONS ARISING FROM THIS THESIS VII

ABSTRACT IX

CHAPTER 1: INTRODUCTION 1

1.1 A TECHNICAL AND GENERAL INTRODUCTION TO DEEPFAKE TECHNOLOGY 1

1.2 THEORETICAL APPROACH: THE SOCIAL CONSTRUCTION OF DEEPFAKE TECHNOLOGY .. 4

1.3 THE MORAL WRONG OF DEEPFAKES 7

1.4 A PSYCHOLOGICAL APPROACH TO THE STUDY OF DEEPFAKE MEDIA 10

1.5 THESIS OVERVIEW 13

1.6 THESIS OUTLINE AND RESEARCH QUESTIONS 17

CHAPTER 2: WHAT IS SO DEEP ABOUT DEEPFAKES? A MULTI-DISCIPLINARY THEMATIC ANALYSIS OF ACADEMIC NARRATIVES ABOUT DEEPFAKE TECHNOLOGY 20

2.1 ABSTRACT.....22

2.2 INTRODUCTION 23

2.2.1 CRITICAL QUALITATIVE WORK ON DEEPFAKE NARRATIVES 23

2.2.2 OUR SCOPING REVIEW APPROACH.....24

2.3 BACKGROUND 24

2.3.1 THE ORIGINS OF DEEPFAKE TECHNOLOGY AND DEFINITIONAL CONFUSION 24

2.3.2 FURTHER REVIEW WORK ON DEEPFAKE DEFINITIONS....25
2.3.3 DEFINING DEEPFAKES IN RELATION TO OTHER TECHNOLOGIES....26
2.3.4 HOW DO DEEPFAKES COMPARE TO OTHER FORMS OF EDITED MEDIA?....26
2.3.5 NARRATIVES AROUND THE CURRENT AND FUTURE USES OF DEEPFAKES....27
2.3.6 CRITICAL REVIEW WORK ON THE USES OF DEEPFAKE TECHNOLOGY....28
2.3.7 AIMS AND RESEARCH QUESTIONS....30
2.4 METHODS....31
2.4.1 SCREENING PROCESS....32
2.4.2 DATA EXTRACTION....33
2.4.3 DATA ANALYSIS....33
2.5 HOW ARE DEEPFAKES DEFINED IN THE LITERATURE?....34
2.5.1 DEEPFAKE DEFINITIONS ACROSS THE DATASET....34
2.5.2 COMMON FACTORS OF DEFINITIONS....34
2.5.3 POINTS OF CONFUSION AROUND DEFINITIONS....36
2.6 THE NOVELTY OF DEEPFAKES....36
2.6.1 ACCESSIBILITY AND BELIEVABILITY: THE TWO “MORES” OF DEEPFAKES....36
2.6.2 BROADER IMPACTS OF ACCESSIBILITY AND BELIEVABILITY....37
2.6.3 COMPARISONS WITH OTHER TECHNOLOGIES....38
2.7 THE EXISTING AND FUTURE USES OF DEEPFAKES....40
2.7.1 POTENTIAL USES OF DEEPFAKES....40
2.7.2 PERSONAL HARMS OF DEFAMATORY DEEPFAKES....43
2.7.3 SOCIALLY MANIPULATION USING DEFAMATORY DEEPFAKES....44
2.7.4 SOCIALLY MANIPULATIVE DEEPFAKES IN CONFLICT....45

2.7.5 EPISTEMIC HARMS OF DEEPFAKE VIDEOS 46

2.7.6 EPISTEMIC HARMS OF DEEPFAKE NARRATIVES 46

2.7.7 ARTISTIC, EDUCATIONAL AND SELF-EXPRESSIVE DEEPFAKES 47

2.7.8 ECONOMIC BENEFITS OF EDUCATIONAL DEEPFAKES 48

2.7.9 DEEPFAKES AS A "DOUBLE-EDGED SWORD" 49

2.8 DISCUSSION 50

2.8.1 DEEPFAKES ARE INCONSISTENTLY DEFINED 51

2.8.2 THE IMPORTANCE OF MIMICKED IDENTITY TO DEEPFAKES 51

2.8.3 DEEPFAKES ARE A NEW WAY OF CAUSING OLD PROBLEMS 52

2.8.4 THE NOVEL EPISTEMIC HARMS OF DEEPFAKES 53

2.8.5 DEEPFAKE ACCESSIBILITY AND BELIEVABILITY: CAUSE FOR CONCERN, STRATEGIES FOR SOLUTION AND NEED FOR EMPIRICAL EVIDENCE 54

2.8.6 ACADEMIC NARRATIVES AROUND THE POSITIVE AND NEGATIVE USES OF DEEPFAKES 55

2.8.7 RECONSIDERING THE CATEGORIZATION OF DEEPFAKES USES 56

2.8.8 FURTHER IMPLICATIONS FOR A HOLISTIC APPROACH TO DEEPFAKES 56

2.9 LIMITATIONS 57

2.10 CONCLUSION 58

CHAPTER 3: DO DEEPFAKE VIDEOS UNDERMINE OUR EPISTEMIC TRUST? A THEMATIC ANALYSIS OF TWEETS THAT DISCUSS DEEPFAKES IN THE RUSSIAN INVASION OF UKRAINE 60

3.1 ABSTRACT 61

3.2 INTRODUCTION 62

3.2.1 BACKGROUND: DEEPFAKES AND MISINFORMATION 63

3.2.2 DEEPFAKES, DISPUTED DEEPFAKES AND OTHER SYNTHETIC MEDIA DURING THE RUSSIAN INVASION OF UKRAINE 65

3.2.3 SYNTHETIC VIDEO FOOTAGE OF A UKRAINIAN FIGHTER JET. 67

3.2.4 DEEPFAKE OF PUTIN ANNOUNCING PEACE WITH UKRAINE. 67

3.2.5 DEEPFAKE OF ZELENSKYY SURRENDERING TO RUSSIA 68

3.2.6 EDUCATIONAL DEEPFAKES FROM THE UKRAINIAN GOVERNMENT AND NEWS MEDIA OUTLETS. 68

3.2.7 EUROPEAN MAYORS RECEIVE POSSIBLE DEEPFAKE VIDEO CALL FROM THE MAYOR OF KYIV. 69

3.2.8 THE CURRENT STUDY 70

3.3 METHODS 71

3.3.1 DATA COLLECTION 71

3.3.2 PREVALENCE OF DEEPFAKE DISCOURSE DURING THE START OF THE WAR 73

3.3.3 ETHICS 74

3.3.4 DATA ANALYSIS 74

3.4 RESULTS 76

3.4.1 DEEPFAKES AND MISINFORMATION (INTERACTIONS WITH NEWS ARTICLES, DEEPFAKES IN WARFARE, SOLICITATION OF DEEPFAKES) 76

3.4.2 DEEPFAKE FUELLED SCEPTICISM 79

3.4.3 NON-MISINFORMATION RELATED DEEPFAKE DISCOURSES 83

3.5 DISCUSSION 84

3.5.1 REACTIONS TO DEEPFAKES IN THE RUSSO-UKRAINIAN WAR 85

3.5.2 DEEPFAKES AND DISTRUST 86

3.6 LIMITATIONS 88

3.7 IMPLICATIONS 89

3.8 CONCLUSIONS 91

CHAPTER 4: DO NO HARM: DESIGNING INTERVENTIONS FOR DEEPFAKE-FUELED CONSPIRACY THEORIES 92

4.1 ABSTRACT 93

4.2 INSIGHTS 94

4.3 THE USE OF DEEPFAKES IN CONSPIRACY THEORIES 94

4.4 CONSPIRACY INTERVENTIONS: AVOIDING A BACKFIRE 95

4.5 DEEPFAKE MISINFORMATION INTERVENTIONS 96

4.6 WHAT SHOULD A DEEPFAKE CONSPIRACY INTERVENTION LOOK LIKE? 98

4.7 CONCLUSION 99

CHAPTER 5: "WHAT DO YOU EXPECT? YOU'RE PART OF THE INTERNET": ANALYZING CELEBRITIES' EXPERIENCES AS USEES OF DEEPFAKE TECHNOLOGY 100

5.1 ABSTRACT 102

5.2 INTRODUCTION 103

5.2.1 NOTE ON TERMINOLOGY USED 105

5.3 RELATED WORK 106

5.3.1 CELEBRITY WOMEN AS DEEPFAKE USEES 107

5.3.2 THE CONSEQUENCES OF BEING A DEEPFAKE USEE 109

5.3.3 THE INFRASTRUCTURAL JUSTICE APPROACH TO NSII 112

5.4 METHODS 115

5.4.1 DATA COLLECTION 115

5.4.2 ANALYSIS 117

5.4.3 POSITIONALITY 119

5.5 ANALYSIS 119

5.5.1 HOW CELEBRITY VICTIMS OF DEEPFAKE ABUSE CONSTRUCT BEING USEES OF DEEPFAKE TECHNOLOGY 120

5.5.2 HOW DO VICTIMS OF DEEPFAKES NEGOTIATE INFRASTRUCTURAL AND SOCIAL OBSTACLES AGAINST SEEKING RECOURSE FOR DEEPFAKE ABUSE? 125

5.6 DISCUSSION 133

5.6.1 IMPLICATIONS FOR DESIGN, POLICY AND PRACTICE 133

5.6.2 IMPLICATIONS FOR HCI: RECONSIDERING DEEPFAKES THROUGH THE LENS OF TRAUMA INFORMED COMPUTING 137

5.6.3 IMPLICATIONS FOR IMPLEMENTING POLICY AND INTERVENTIONS 139

5.7 LIMITATIONS 141

5.8 CONCLUSION 142

CHAPTER 6: THE DEVELOPMENT AND VALIDATION OF THE DEEPFAKE MYTH ACCEPTANCE SCALE (DMAS) 143

6.1 ABSTRACT 145

6.2 INTRODUCTION 146

6.2.1 MYTHS UNDERLYING IMAGE-BASED SEXUAL ABUSE 148

6.2.2 DEEPFAKES: IMAGE-BASED SEXUAL ABUSE UTILISING AI 149

6.3 STUDY 1: EXPLORATORY FACTOR ANALYSIS 153

6.3.1 METHOD 153

6.3.2 RESULTS 155

6.4 STUDY 2: CONFIRMATORY FACTOR ANALYSES AND PERPETRATION BEHAVIOURS 156

6.4.1 METHOD 156

6.4.2 RESULTS 158

6.5 STUDY 3: CONFIRMATORY FACTOR ANALYSIS IN A HIGHER PERPETRATING SAMPLE 165

6.5.1 METHODS 166

6.5.2 RESULTS 167

6.6 DISCUSSION 168

CHAPTER 7: DESIGNING AN INTERVENTION TO REDUCE ATTITUDES AND BEHAVIOURS RELATED TO DEEPFAKE NON-CONSENSUAL INTIMATE IMAGERY 173

7.1 ABSTRACT 174

7.2 INTRODUCTION 175

7.2.1 HOW CAN WE REDUCE THE PERPETRATION OF DEEPFAKE ABUSE? 176

7.2.2 DEEPFAKES/REAL HARMS INTERVENTION 180

7.3 PILOT STUDY 183

7.3.1 INTERVENTION DESIGN 183

7.4 STUDY 1: INTERVENTION EVALUATION 187

7.4.1 METHOD 188

7.4.2 RESULTS 191

7.5 STUDY 2: REVISED INTERVENTION EVALUATION 195

7.5.1 METHOD 196

7.5.2 RESULTS 197

7.6 STUDY 3: EVALUATING A FURTHER-REVISED INTERVENTION IN HIGHER PERPETRATING POPULATIONS 204

7.6.1 METHODS 204

7.6.2 RESULTS 206

7.7 DISCUSSION 215

7.7.1 IMPLICATIONS 219

7.8 CONCLUSION 221

CHAPTER 8: DISCUSSION 222

8.1 CONTRIBUTION OF THE RESEARCH 222

8.1.1 SCOPING CURRENT ACADEMIC CONSENSUS ON DEEPFAKES 222

8.1.2 REAL WORLD CASE STUDIES OF THE HARMS OF DEEPFAKES 223

8.1.3 IMPROVING PSYCHOLOGICAL SOLUTIONS AND MEASURES FOR TECHNOLOGY FACILITATED GENDERED VIOLENCE 225

8.2 CONTRIBUTION OF THE THESIS TO THEORY 226

8.2.1 ONLINE IDENTITY AND MORAL INFRACTION 227

8.2.2 AUGMENTING TECHNOLOGICAL SOLUTIONS WITH PSYCHOEDUCATIONAL INTERVENTIONS 228

8.2.3 RE-EVALUATING THE FOCUS ON “NOVEL” TECHNOLOGIES VERSUS ESTABLISHED HARMS 230

8.3 FUTURE DIRECTIONS 231

8.4 LIMITATIONS 234

8.5 CONCLUSION 236

REFERENCES 237

APPENDIX I: I WAS NEVER IN THE AI.R.A: AN EXPERIMENTAL STUDY ON THE USE OF MANIPULATED AUDIO FOR THE CENSORSHIP OF POLITICAL DISSIDENTS 271

APPENDIX II: FACE/OFF: CHANGING THE FACE OF MOVIES WITH DEEPFAKES 279

APPENDIX III: ENCOUNTERING DEEPFAKES: A THEMATIC ANALYSIS OF COMMENTS ON A MUSIC VIDEO FEATURING DEEPFAKE CONTENT 298

APPENDIX IV: QUALITATIVE CODING AND ANALYSIS EXAMPLE IN CHAPTER 5 321

APPENDIX V: EXAMPLE PARAGRAPH OF WRITING UP THE ANALYSIS FOR STUDY 4.....322

APPENDIX VI: ONE WEEK FOLLOW UP RESULTS FOR STUDY 4. 324

List of Figures

Figure 1.1: The three stages of deepfake production as outlined by (Perov et al., 2020), illustrated with an image of the author (John Twomey) face-swapping with the actor Jim Carey. 3

Figure 1.2: Structure of the papers in the thesis by aim 15

Figure 2.1: Prisma flow diagram for search results 32

Figure 2.2: Themes drawn from deepfake definitions. 35

Figure 2.3: Themes and subthemes relating to the accessibility and believability of deepfakes 37

Figure 3.1: Timeline of the use of deepfakes/synthetic media in the war. 67

Figure 3.2: Prevalence of tweets over time in the dataset. 74

Figure 5.1: Summary of Robinson's and colleagues (2024) work on the infrastructure behind the creation of deepfake NSII 113

Figure 6.1: Second order factor analysis where the first order factors are deepfakes: i) having no consequences, ii) being deserved and iii) not being harmful. Deepfake myths are the second order factor. 164

Figure 7.1: Deepfake Myth acceptance by group in study 1 192

Figure 7.2: NSII watching intentions by group in study 1 193

Figure 7.3: DMAS scores across the control and intervention group at times 1 and 2 for study 1. 194

Figure 7.4: Deepfake Myth acceptance by group in study 2 199

Figure 7.5: NSII watching intentions by group in study 2. 199

Figure 7.6: DMAS scores across the control and intervention group at times 1 and 2 for study 2. 201

Figure 7.7: Intended NSII watching perpetration by victim group in study 2 comparing the original control group with the intervention group at follow up one month afterwards 202

Figure 7.8: Deepfake Myth acceptance by group in study 3 208

Figure 7.9: DMAS scores between control group at initial data collection nand intervention group at the follow-up one month afterwards. 210

List of Tables

Table 2.1 The key beneficial/ harmful uses of deepfake technology according to Chesney and Citron (2019) 28

Table 2.2: Approaches to organizing deepfakes in comparison to other media 40

Table 2.3 Wordcounts of deepfake-related words in the impacts section of our dataset. 43

Table 5.1: Deepfake statements in our corpus 117

Table 5.2: Various stakeholders identified as holding responsibility for deepfake abuse 135

Table 5.3: Implications of our work for Trauma Informed Computing targeting deepfake NSII 138

Table 5.4 Deepfake NSII myths from our corpus and how they were challenged 141

Table 6.1: Deepfake NSII myths expressed by celebrity victim-survivors (Twomey, Foley, et al., 2025) 151

Table 6.2: Means, standard deviation, and component loadings of the initial draft of the DMAS 155

Table 6.3: Correlations between different deepfake NSII behaviours 159

Table 6.4: Perpetration intentions for different types of victim-survivors (1 = Strongly Disagree and 5 = Strongly Agree) 160

Table 6.5: Component Matrix of the final six-item scale 162

Table 6.6: Pearson correlations of deepfake myths (DMAS), Acceptance of Modern Myths about Sexual Aggression Scale (AMMSA), and Misogyny with deepfake NSII intentions by victim type 165

Table 6.7: Pearson correlations between DMAS scores (as well as the AMMSA and misogyny scales) and intended perpetration behaviours in a higher-perpetration sample 168

Table 7.1: Means and standard deviations for self-reported perpetration behaviours and intentions in Study 1 191

Table 7.2: Exploratory analysis of differences between groups at follow up 193

Table 7.3: Self -reported perpetration behaviours and intentions 198

Table 7.4: Comparisons between experimental and control group across measures 200

Table 7.5: Means and standard deviations across perpetration measures 207

Table 7.6: Difference between the control group and intervention by victim-type 209

Table 7.7: Comparison of the means in the month long follow up versus the original control group 211

Acknowledgements

First of all, I must thank my supervisors. To Gillian Murphy, you have set the standard for all of us lucky enough to be under your supervision. Every time the work felt like too much or I felt lost, you were the one who helped bring it all back into focus. I am grateful to be ending the PhD with a wealth of lecturing and research experience, and I owe this to the opportunities and guidance you have given me. To Conor Linehan, you have been a constant source of insight. Your ethos towards the writing process and the PhD itself has been invaluable and I count myself incredibly lucky to fall under your tutelage. To Matthew Aylett and Mike Quayle, you have both consistently offered useful perspectives on this work which have shaped the form of this thesis.

This thesis has benefitted massively from a wealth of talented colleagues who have also contributed to the work as coauthors. Sarah Robinson and Sarah Foley, you have both been constant supporters and provided a wealth of knowledge and insight around qualitative research. I would also like to thank the two undergraduate psychology students who contributed to the last research study: Emily Lavoie and Amy Louise Geary. Didier, I have been so lucky to have another PhD student or work so closely with. I have learned so much from you and I hope you feel the same. Likewise, Cian, comrade-in-arms, there's so much I could write here about how valuable your friendship, insight and hot-tub have been to me.

I must also note how blessed I have been in terms of my work environment, both in the department generally and also in the PhD office – there really is too many people to mention here. Dan, Stephanie and Ella, you have made the office a place I really look forward to going to every day. You are extremely talented academics, and I am so glad to call you, my friends. There are so many other people who have contributed to making this office such a pleasant place to work over the past four years. Jen, Orla, Billy, Hongru and Tadgh, you have provided constant support, kindness and friendship to me. I would also like to thank both the old guard of

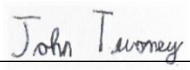
the office who welcomed me, as well as the incoming cohorts of PhD students and research assistants (Niamh, MaryClare, Fiona, Helena, Alison, Kelly and others) who have put up with my stressed-out ramblings. There have been a number of short-term visitors to the department who have also made the PhD office a better and more collaborative place.

Hanke, you have been there for everything, seen me at my most tired and stressed and you provided constant love and support. In the past few months, you have really kept me going through it all and I will not need to be reminded to do the same for you when your thesis is at its end stage. I must also thank both of your parents for opening their beautiful home in Germany to me yearly as a well-earned escape from this thesis.

Thanks to all my friends who have offered me support and escape these past few years, especially Sam, Cal, Jackie, Dan, Orla and Michael Manley. Thank you also to my family for their constant support. Particularly, my siblings; Donal, Liam and Sinead, as well as my uncle Mark and his family. A special thanks to my mother, Angela Fitzgibbon. You cared about my education more than anyone else.

Declaration of Originality

"This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property."

Digital signature of the candidate: 

Publications arising from this thesis

First author publications

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

First Author: preprints currently under review for publication

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*.

Twomey, J., O'Mahony, C., Linehan, C., & Murphy, G. (2025). The Development and Validation of the Deepfake Myth Acceptance Scale (DMAS). https://doi.org/10.31234/osf.io/6qn7y_v1

First Author publications: Not in the thesis

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 ACM Conference on Human Factors in Computing Systems* (pp. 1-8).

Second author and other publications during the PhD

Published

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

Robinson, S., Buckley, J., Ciolfi, L., Linehan, C., McInerney, C., Nuseibeh, B., ... & McCarthy, J. (2024). Infrastructural justice for responsible software engineering. *Journal of Responsible Technology*, *19*, 100087.

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.

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*, *CSCW2025. 9*(7), 1–23.

Abstract

Deepfake technology has emerged as a novel form of manipulation for images and videos and has increased the prevalence and realism of false video information online. While there have been some positive uses of deepfakes in translation and entertainment, the technology is often defined by two negative use cases: political misinformation and non-consensual sexual imagery. A further harm comes from the lack of faith in real videos as a consequence of the existence of comparatively accessible video manipulation. There has been little psychological research which aims to explore the harms of deepfake technology. I report five studies which use both qualitative methods (thematic and discursive analysis) and quantitative methods (psychometrics and experimental design) to study the harmful uses of deepfakes. The first three of these papers seek to understand how the various harms of deepfakes are constructed among scholars, the general public and victims of deepfake abuse through qualitative methods. These papers involved case studies that explore the impacts of deepfake misinformation during the Russian invasion of Ukraine and the harmful role of deepfake abuse imagery. The final two papers seek to make a positive contribution to the psychological study of deepfake perpetrators by developing a psychometric measure to study deepfake abuse myth acceptance, and an intervention which aims to reduce perpetration. The thesis contributes a number of novel case studies on deepfake harms as well as a first-of-its-kind tailored measure and intervention targeting deepfake abuse. Another contribution is a critical understanding of the role of online identity in technology facilitated gendered violence and the need for social and psychological solutions to augment existing technological solutions. I also outline a number of possible future directions for the psychological study of the consequences of deepfake abuse imagery. These future directions include the potential for future research targeted towards different perpetrators of deepfake abuse and the need for improved deepfake education in schools.

CHAPTER 1: Introduction

1.1 A technical and general introduction to deepfake technology

> "*Zachariah Harker's face replaced Hale's. It would have taken a better than expert eye to detect the faint qualities of unevenness which might betray the fact that this was a synthesis of rearranged sound-waves and light-waves. Technically, even Zachariah, watching his screen from wherever he was just now could not deny he had spoken the words, for every sound he heard and every motion of his own lips was genuine*"
>
> Henry Kuttner (1950)

Once the realm of pulp science-fiction, deepfake technology has made high quality manipulation of videos and audio a reality. In their most simple forms, deepfakes are a type of audiovisual media creation technology which utilises generative AI to produce false representations of individuals sayings or doing things which never happened (Whittaker et al., 2023). The purpose of my thesis is to study novel deepfake technology and how deepfakes can harm individuals and society. The thesis aims to study the harms of deepfake technology raised by a variety of stakeholders: academic scholars, the general population, victims of deepfake harms and perpetrators of these harms. Before we can outline how deepfakes harm, it is vital to understand what they are and how has this technology developed since their initial creation in 2017. Since their inception, a wide range of false videos have been referred to as deepfakes. The EU AI act defines deepfakes as AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful (Act, 2021). While the EU definition covers an extremely broad range of technology, in practice deepfake software usually involves the swapping of one face with another or the artificial movement of lips to change uttered speech (e.g., lip-sync videos). The first publication in this thesis (Chapter 2) discusses in detail the host of "definitions" which are used for deepfake technology. Instead of repeating that work here: I will

instead present a technical introduction to the main types of deepfakes; face-swaps and lip sync videos.

The original deepfake videos used so- called “face-swapping” software which swapped the face of an actor in a video with one of your choosing. The technology gets its name from a user of the social media site reddit (titled u/deepfake) and from the online community they founded (r/deepfakes). The original deepfake videos were specifically designed to target celebrity women, who were non-consensually edited into internet pornography. The original reddit user u/deepfake took an established practice of photoshopping celebrity women into nude imagery (r/cheapfakes) and developed their own tools to produce realistic videos in which the faces of two people were swapped. While the original deepfake community was shut down, deepfake technology had found its place in numerous apps and software repositories online (such as GitHub repositories). Two such examples of publicly accessible deepfake software include DeepFaceLab and DeepFaceLive, both of which significantly broadened the accessibility of the technology. The processes of creating a deepfake through face-swapping software such as DeepFaceLab generally requires three steps (Perov et al., 2020): extraction, training and conversion (see Figure 1.1).

1. Extraction - Initially the data is provided with a source (to provide the desired face) which consists of images and video data of a person (Perov et al., 2020). The software is also provided with a destination video (the video that will have the face swapped into it). These videos are extracted as a series of images and the software detects and aligns the faces in both the source and destination image sets. After the faces have been detected a neural network is used to carry out face segmentation, which isolates the visible part of the face excluding ears, hair, etc.
2. Training - At this point in the process there are numerous methods of training to software into creating an AI-generated face (Swathi and Sk, 2021). Generative adversarial networks are often used to make the videos more realistic (Nguyen et al., 2022) where the face generator competes against a discriminator neural network, which

is designed to tell real videos apart from fakes (Albahar & Almalki, 2019). The testing process is repeated until the generator produces videos of a high enough quality that the discriminator cannot tell if it is real.

3. Conversion - The final stage of the process is to interpolate the generated face of the source, into the destination video. The final process involves blending the face into the new video, accounting for different skin tones, facial features, and lighting (Albahar & Almalki, 2019). The video is also often sharpened to highlight details in the face.

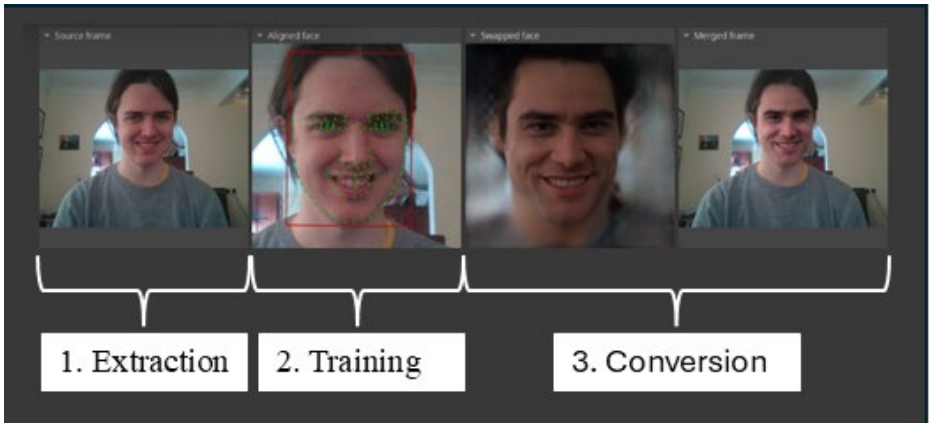


Figure 1.1: The three stages of deepfake production as outlined by (Perov et al., 2020), illustrated with an image of the author (John Twomey) face-swapping with the actor Jim Carey.

While deepfake videos that involve swapping women into sexual imagery are often referred to as “deepfake porn”, the terminology is critiqued in the same way as the term “revenge porn” was critiqued for emphasising the pornographic/entertainment nature of the content as opposed to the harm (Twomey, Foley, et al., 2025). While initially I referred to some the harms of deepfakes as pornographic, as the thesis has progressed (and as the terms used in the literature have evolved), I have moved away from that term and aimed to use more appropriate language. Based on established literature, I moved towards terminology such as deepfake abuse to describe the act itself (Flynn et al., 2025) and deepfake abuse imagery and Non-consensual Synthetic Intimate Imagery (NSII) to describe the content (Flynn et al., 2025; Umbach et al., 2024).

Outside of the original face-swapped deepfakes used in the creation of Non-consensual Synthetic Intimate Imagery, there have been significant concerns over the impact of deepfakes in political misinformation. Deepfakes of politicians do not usually involve videos swapping a face into an unrelated video. Instead deepfake misinformation often utilise the second most common type of deepfake: lip-syncing/facial puppetry. These techniques change the expressions and utterances of an existing video, by focusing on the lips of the target instead of the entire face and matching the lip movements to a new audio track (Suwajanakorn et al., 2017). Public concerns grew about the potentials of lip-sync deepfakes with the famous deepfake of Barack Obama which was made by film-maker Jordan Peele and released by the news organisation BuzzFeed (Pumphrey, 2024). The video purported to be a real video of the then US president, warning viewers to be on the lookout for new forms of dangerous fake information, before revealing Jordan Peele controlling the lip movements of the video. The tools for the generation of these types of lip-sync deepfake videos have become extremely accessible in the years since their initial development. It is currently possible for anyone with a modicum of computer skills, to generate videos online using websites specifically designed to make the experience as user-friendly as possible featuring the ability to directly manipulate videos from YouTube and generate false audio at the click of a button (Murphy et al., 2025).

Relatedly, lip-sync deepfakes often benefit from using AI-generated audio, where a large amount of audio data of a public figure is used to generate convincing audio of that person saying anything a deepfake creator wants (Aylett & Vazquez-Alvarez, 2020). While these tools historically only worked on public figures due to the need for large amounts of data, again there have been significant leaps forward in their accessibility and the development of browser based AI “cloning” websites which require less than a minute of existing audio to generate a unique copy of the voice saying anything which a deepfake creator desires (Murphy et al., 2025).

1.2 Theoretical approach: the social construction of deepfake technology

As various forms of AI and machine learning become more prevalent, understanding how different stakeholders understand deepfake technology becomes vital. Social Construction

of Technology (SCOT) theory involves understanding technology as artifacts which develop through relationships with a variety of social actors who have both problems with the technology and actively play a role in solutions which further develop the technology (Pinch & Bijker, 1984). The study of technology from a SCOT perspective requires researchers to consider the place of technology within a larger context of similar past and present technologies. It is vital to understand technological artifacts not as the sum of existing technologies but as part of a constantly negotiated process between a variety of "actors" concerned with the problems and benefits of a technology. Deepfakes serve as an excellent example to illustrate SCOT. One distinct focus of a SCOT approach is the actor-network theory, where which focuses on understanding the social relationships that technology is embedded into (Baron & Gomez, 2016; Pinch & Bijker, 1984).

Deepfake technology serves as an interesting example of the development of technology being shaped by social relationships in how it very clearly emerged from an online community (Han et al., 2025), and open-source technology repositories (Winter & Salter, 2020), rather than academic or commercial origins. Because of their social origin, a significant amount of theoretical and empirical work has been dedicated to understanding the structure of these groups. Qualitative work studying public response to deepfakes online includes work by Gamage and colleagues analysing reddit data from the initial phase of deepfakes on the social media platform reddit (Gamage et al., 2022). The dataset of Gamage's paper also gives a useful historical and sociological insight into the initial phase of deepfake technology on reddit before deepfake abuse imagery was banned. The work also highlights the economy behind deepfakes, which has been further studied in work by Timmerman which qualitatively analysed the beliefs of perpetrators on deepfake abuse imagery forums (Timmerman et al., 2023). My contribution to that field of interest was as a co-author on a paper focusing on critiquing the development of deepfake technology from a responsible software engineering approach (Robinson et al., 2024). Robinson's approach focused on the social characteristics of deepfake abuse imagery infrastructure, including how deepfake abuse imagery is part of an installed base of deepfakes

(e.g. abuse imagery was linked in software repositories), learned as part of membership (where deepfake communities specifically train users in deepfake abuse imagery generation) and part of conventions of practice (often deepfake forums dedicated towards developing the technology and creating NSII).

However, the existing research studying the actors behind harmful deepfakes mainly focus on the creators/users of deepfake videos and technology. There is not a clear way in which the victims of deepfake imagery fit into actor-network theory. Traditionally the Social Construction of Technology mainly illustrates the importance of users and producers of technology (Oudshoorn & Pinch, 2005). Similarly, the dominant ideology of HCI as a discipline has historically been concerned with the "use" of technology and so focuses explicitly on the active use of a technology (Baumer & Brubaker, 2017). While the studies of users are of benefit to those designing technology and technological solutions, it does not serve as a holistic approach to technologies which harm others. In the case of deepfakes, I am concerned with those who engage passively with technology (the exposure to deepfake content online created by others) or find themselves actively targeted by a technology. In Chapter 5 of the thesis, I shift focus onto various types of non-users of the technology, including who Eric Baumer calls "usees" (Baumer, 2015). Usees are non-consenting, unaware targets of a technology. While they do not take an active role in the creation or use of a technology, they have a significant role in how technology is socially constructed as they are the people who deepfake technology is used on.

When discussing the actor-network which underlies the evolving social construction of deepfake technology, the perspectives of "usees" give a vital insight into the harms of deepfakes. The negative uses of deepfakes in political misinformation and deepfake abuse imagery are both tied into its potential to create usees through the non-consensual creation of deepfakes of unaware targets. Because usees play a role in the actor-network behind deepfake technology, prioritising the experiences of deepfake usees over creators/users may play an important role in shaping the future development of deepfake technology and research. After

developing an understanding of the current targets of deepfakes in Chapter 5, I aimed to actively impact the actor-network of deepfake use/creation by forcing the users/creators of deepfake technology to reckon with the experience of usees in Chapter 7.

1.3 The moral wrong of deepfakes

My thesis aims to challenge the two dominant philosophical concerns with the harms of deepfake technology: ethical concerns and epistemic concerns. The most significant ethical concerns with the technology revolve around its uses in political misinformation and the creation of non-consensual sexual imagery of women. The epistemic concerns rely on how the social construct of a deepfake video undermines our ability to trust videos wholesale.

It is impossible to discuss the impact of deepfake technology, without having to determine what constitutes as a "positive" and "negative" use. In the debate over whether deepfakes are morally wrong, I have drawn from de Ruiter's work on the morality of deepfakes which describes deepfakes as inherently morally suspect but not necessarily wrong. What renders them morally wrong is when they involve the representation of someone's digital identity in a situation they would not agree with (de Ruiter, 2021). Non-consensual manipulation of identity is vital to the moral harms of deepfakes. Throughout the thesis I emphasise the role of manipulated identity as being fundamental to deepfakes: manipulated imagery often means the manipulation of someone's identity. The public persona we take on is not an output of an interior self but an interaction with the world which can have significant psychological implications. Deepfakes are fundamentally a tool to manipulate a person's image.

The epistemic harms of deepfakes overlaps with the harms of deepfakes in political disinformation and manipulated pornography. However, the epistemic harms of deepfake technology are not harms of deepfake videos themselves but rather a harm to trust caused by the concept of deepfakes. Video recordings have often served as what Rini calls "an epistemic backstop" for testimonial evidence (Rini, 2020). Deepfakes inherently challenge the reliability of video evidence and thus may lead to a rise of individuals who are overly sceptical of some or

all video evidence. This scepticism ties into what Chesney and Citron call the “liar’s dividend”, where the existence of deepfake benefits liars who seek to discredit real evidence (Chesney & Citron, 2019). Chesney and Citron’s paper moves the discussion beyond the impact of individually harmful deepfakes, to highlight the harms that the existence of deepfakes causes to wider trust in media encountered online. Videos inherently carry information about the material world and because they are now easy to manipulate, they carry less reliable information, undermining our information environment (Fallis, 2021).

There are some researchers who have disagreed that deepfakes have distinct and novel epistemic harms by drawing on broader histories of image manipulation. Specifically, dissenting researchers acknowledge the damage of deepfakes to trust but point out that it has always been possible to manipulate and misrepresent video content (Habgood-Coote, 2023). The epistemic harms of deepfakes are critical to Chapter 3 of the thesis, which discusses the epistemic harms of deepfakes online during global conflicts. The epistemic questions of deepfakes are also related to its use in deepfake abuse imagery. Some scholars have posited that the existence of deepfake abuse imagery may offer a benefit to victims of IBSA (Viola & Voto, 2023) due to their epistemic harms. Reflecting on Chesney and Citron’s (2019) Liar’s dividend may offer victims of traditional Image-Based Sexual Abuse (IBSA) the opportunity to deny that real imagery exists by claiming that real footage is manipulated. However, the other side of that epistemic concern, is that AI-generated abuse imagery may be treated as real footage (Harris, 2021). Chapter 5 of the thesis shows that the victims of deepfake abuse imagery are concerned that AI-generated imagery may be believed as real.

At the outset of my PhD research, I planned on including case studies on both the positive and negative applications of deepfake technology. However, as my understanding in this area grew and as the field itself rapidly evolved since I first enrolled in my PhD in 2021, I ultimately chose not to focus on positive uses of deepfakes in the thesis. The negatives of deepfake technology significantly outweigh the positives and desperately require the intervention of psychological/human-computer interaction (HCI) research. As previously stated,

though not inherently wrongful, deepfakes have been described as “morally suspect” as even supposedly positive uses (such as the resurrection of dead actors to complete movies) may often still involve the non-consensual manipulation of another’s identity (de Ruiter, 2021). The negative consequences of the use of deepfakes in film and television was a growing social concern across the PhD, with the Screen Actors Guild–American Federation of Television and Radio Artists in the United States going on strike, partly over the use of AI recreations and actors (Thomas, 2024).

In Chapter 2 of the thesis, I criticise the idea of deepfakes as having equal amounts of positive and negative uses (deepfakes as a “double-edged swords”). I also take umbrage with the unevidenced assumption that because there may be a positive use of deepfakes, we should distribute work towards positives. I believe that it is vital to study the potential harms of a novel technology before engaging with the potential hedonistic benefits that the technology can offer. Though intentionally absent from the thesis, I have also carried out work on these supposedly positive uses of deepfake media. Included in the appendix is a first author paper I wrote on the use of manipulated audio in political censorship (Appendix I). The paper, which was published as part of the alt.chi track of CHI 25, is a satirical paper on educational or restorative deepfakes (Twomey, Foley, et al., 2025). It highlights the inherent falsity of AI restorations of censored footage and the dangers of the use in AI in historical educational materials. As a second author I have also been involved with work focusing on the use of deepfakes in entertainment, both in the recasting of film stars (Murphy et al., 2023) in Appendix II and artistic digital resurrections of dead celebrities in a music video (Ching, Twomey, Linehan, et al., 2025) in Appendix III. What emerged in these papers is that there are also reasons to be concerned with the beneficial uses of deepfakes. Both of those works give insight into the uncertainty around positive uses of deepfakes and concerns about the falsity of the content and the damage of automating the roles of artists and creators. In terms of the positive uses in education and restoration, there is a lack of convincing arguments as to why a piece of real media could not accomplish the same solution without the downsides that come from the use of AI-generated materials.

Outside of the epistemic and moral issues within the technology itself, there are broader concerns with the ethics of AI as a technology which should be mentioned here. These concerns influenced the decision not to focus on the "positive" uses within the overall thesis. Firstly, there are a number of ethical issues behind the design of AI technology itself relating to the lack of consent in gathering data on which the models are trained (C. P. Lee et al., 2025). There are also concerns that the use of open source deepfake technology actively legitimises the abusive content which they are trained to create (Robinson et al., 2024; Winter & Salter, 2020). In Chapter 5 I briefly question the ethics of the use of deepfake technologies which are designed for deepfake NSII creators. Other broader critiques of generative AI involve the environmental damage which occurs from the collection/storage of data and draining of resources such as water and energy (Bender et al., 2021; Jegham et al., 2025).

1.4 A psychological approach to the study of deepfake media

Initially the work reported in this thesis drew from a Human Computer Interaction (HCI) background, focusing on how various stakeholders are impacted by novel deepfake technology. As the work developed, it made more sense to lean into an approach informed by forensic psychology, which has its roots in the design of interventions and measures which aim to reduce aberrant behaviours. While the broad range of theoretical influences is a strength of the overall thesis, it was a challenge to try and establish a concrete identity for the thesis as a work of Applied Psychology while reading such a wide range of literature. The scoping review in Chapter 2 highlights how the overwhelming amount of work in the field is in computer science and law and that psychological study of the harms of deepfakes has fallen well behind. Throughout the PhD, I switched focus to specifically understanding deepfakes within the emerging concept of technology facilitated gendered violence (TFGV). TFGV is an umbrella term to help understand a variety of novel offences enabled by new technological tools and often directed towards women and gendered minorities. These include traditional acts of harassment and abuse translated to an online setting such as active campaigns of harassment and threats directed towards feminist activists online. There are a number of novel offences related

to the creation and sharing of intimate imagery online such as a variety of forms of IBSA. As deepfake videos often play a role in the generation of abuse imagery, TFGV become a useful lens to study the consequences of deepfake technology towards women (Umbach et al., 2024) The concept originates from the criminological study of technology (Henry & Powell, 2018). While originating in criminology, the study of novel forms of gendered violence online may be an area in which psychological research may contribute. Specifically, Forensic Psychology, Psychology of Law and Behavioural Psychology approaches may contribute to the study of perpetrators characteristics and the study of perpetration-supportive cognition.

There are only a handful of psychological papers that take such an approach, employing the psychological study of TFGV. Of the existing literature, the most relevant examples are Fido and colleagues' work on the psychological profile behind deepfake abuse imagery perpetrators (Fido et al., 2022). They have successfully identified a number of factors which play a role in the creation of abuse imagery with deepfakes, but their work does not (yet) aim to actively reduce these behaviours. My work seeks to further contribute to this area of study by developing measures and interventions which may augment the future study of deepfake abuse imagery online. The development of psychological tools and interventions aim to encourage psychological, sociological and educational solutions as opposed to techno-solutionism for deepfakes.

The thesis privileges psychological solutions to the harms of deepfakes over the currently advocated technological solutions. In her book "Artificial Unintelligence: How Computers Misunderstand the World", Meredith Broussard coined the phrase techno-chauvinism to describe the current dominant philosophy around harms online: "technology is always the solution" (Broussard, 2018). In the context of the development of novel technology, techno-chauvinism involves blind optimism about technology and lack of caution about the use of novel technologies. Habgood-Coote (2023) applied the concept of techno-chauvinism to deepfake misinformation, highlighting three constituent philosophies behind techno-chauvinism: techno-optimism, techno-fixation and techno-solutionism. Techno-fixation involves

underplaying the importance of the designers and users of technological systems which has already been discussed regarding how the social construction of technology focuses on actor networks over non-consenting and targeted stakeholders.

Here instead, it is worth commenting on how the techno-optimism as a critique has been challenged by the rapid development of the technology. In the context of deepfakes, techno-optimism is the idea that technology is inherently competent at what it produces. For deepfakes, techno-optimism has consisted of an over-focus on the harms of what I call a "perfect deepfake" that is a professionally made deepfake video containing misinformation which is completely indistinguishable from a real video for the majority of the population. Existing research evidence has shown that even high quality deepfakes are not uniquely dangerous than other forms of fake images and texts (Ching, Twomey, Aylett, et al., 2025). However, it should be noted that even poor quality deepfakes may be enough to encourage broad scepticism of real media. Outside of deepfake misinformation, deepfake abuse imagery is really not as concerned with being real as it is with using high and low quality deepfakes in the abuse and harassment of women and violation of their sexual privacy by developing often-violent sexual materials without consent.

Techno-solutionism remains as a vital area of critique of existing deepfake research and practice and remains a larger motivator for the work of the thesis. Much existing responses to deepfake media have been characterised as techno-solutionist: that is to say solutions tend to "repackage social problems as technological problems" (Habgood-Coote, 2023). Habgood-Coote specifically critiques the focus on the harms of deepfake misinformation which focus on the qualities and application technology itself and justifying potential technological solutions. These technological solutions (such as debunking or detecting deepfakes) miss out on the underlying social problems which include broader issues with trust in the "post-truth" information environment such as lower trust in institutions and governments globally. They are also never as effective or widespread as the harms of the technology which they aim to stop. Though Habgood-Coote's paper only focuses on misinformation, it is very easy to draw

parallels with the current “solutions” for deepfake abuse imagery. One can also make the claim that the response to deepfake abuse imagery also tends to focus on the technology itself, focusing mainly on developing automated systems and reporting for deepfake abuse imagery. A focus on the social problems of deepfake abuse would focus instead on the growth of technology-facilitated gendered violence and increasing violations of sexual privacy online.

1.5 Thesis overview

The overall goals of the thesis are as follows (Figure 1.2):

1. To understand what deepfakes are, what makes deepfake technology different from what has come before, and to better establish the potential threats and benefits of the technology
2. To provide real world evidence for the current harms of deepfake technology across two domains (misinformation and abuse imagery)
3. To contribute a validated measure and intervention to aid the psychological study of the most ubiquitous harm of deepfake technology: deepfake abuse imagery

The overall aim of the thesis is to broaden our understanding of deepfake technology and to form an evidence base and tools for researchers who seek to challenge the harms of the technology. The initial aim of the PhD was to focus specifically on using design fiction to focus on the various hypothetical harms of deepfake technology that would emerge in the review. It quickly became apparent during the initial screening of the scoping review that the literature was already focused overly on the hypothetical at the expense of studying existing real harms. The discourse that surrounds deepfake technology has frequently been compared to a moral panic (Broinowski, 2022). The Russian invasion of Ukraine, and the subsequent use of deepfakes in that conflict, made it apparent that deepfakes were already seriously impacting the world and that it would be better for the thesis to focus on the real-world current harms. Thus, I investigated public reactions to deepfakes in the Russo-Ukraine war, where the harm of the technology was not only deepfakes themselves but how the discourse around novel AI-

generated technology informed widespread scepticism. I then moved my focus in the rest of the thesis to focus specifically on the use of deepfakes to generate non-consensual synthetic intimate imagery. The third paper focuses both on how the most common victims (celebrity women) construct their abuse and recourses. In that paper I discovered a number of false beliefs used to blame victims of deepfake technology. My final two papers focus on measuring these false beliefs and designing an intervention which reduces them and the perpetration of deepfake abuse.

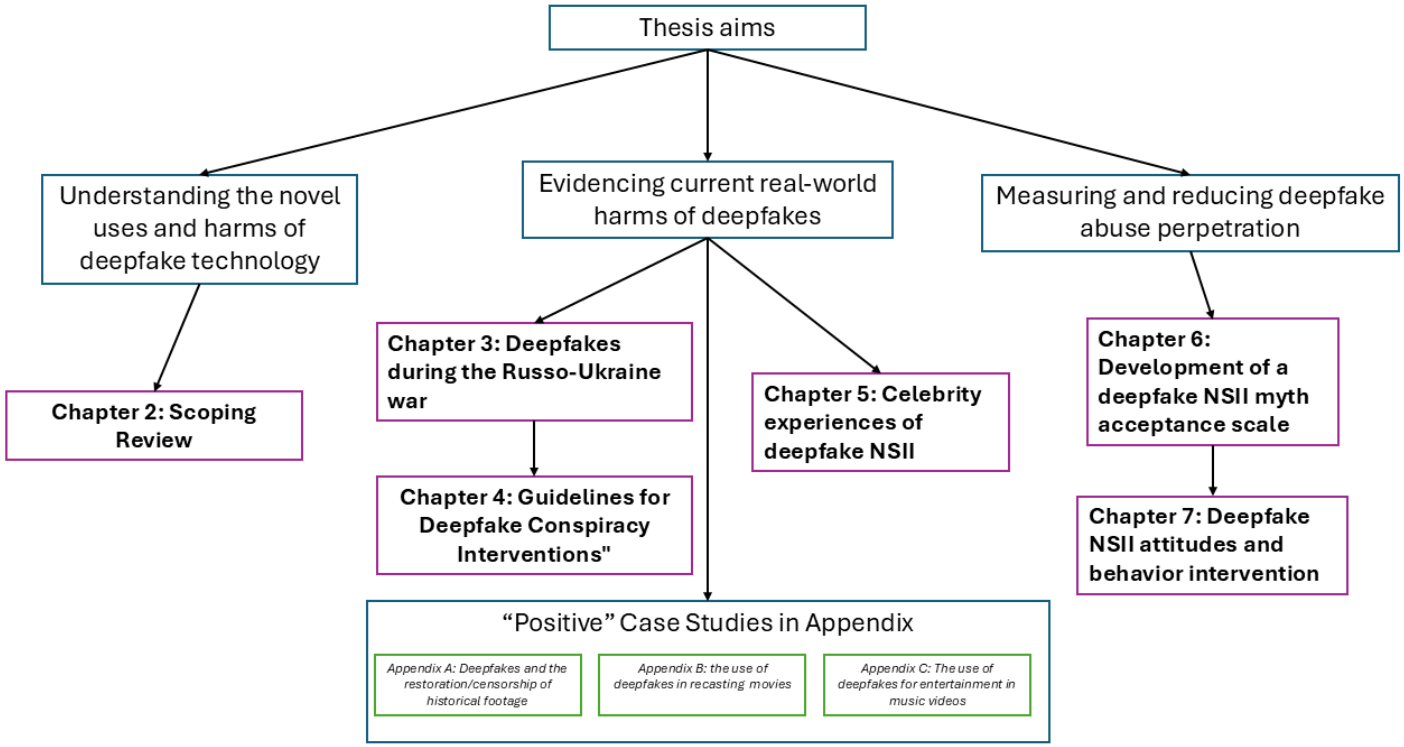


Figure 1.2: Structure of the papers in the thesis by aim

In terms of research methodologies, the work within the thesis is part of a larger literature of psychological research conducted using the internet and targeting behaviour and attitudes related to the use of online technologies. The psychological study of the internet and related technologies generally draws from three different approaches (Gosling & Mason, 2015): phenomenological, novel methods and translational methods. Firstly, phenomenological methods involve the psychological study of novel phenomena and topics which have emerged from internet technologies or are mediated by the internet. These include various forms of problematic behaviours mediated by the internet such as cyberbullying and online political polarisation. Every research paper in the thesis is concerned with the study of unique online phenomena relating to the creation and spreading of deepfakes online including deepfake misinformation (Chapter 3 and 4) and deepfake abuse imagery (Chapters 5, 6 and 7). Secondly, novel methods involve the use of new internet-based research methodologies as a means to study the internet including the analyses of language and social behaviours online. Two of the papers in the thesis employ qualitative analysis of secondary online data, natural data from the general public gathered from a social media website for Chapter 3 and data from online vlogs, podcasts and public statements for Chapter 5. The benefits of online qualitative analyses as novel methodologies are that they allow for the study of real-world information environments and allow for qualitative research where in-person research is impossible (due to the lack of reported victims of deepfakes in Ireland at the time of data collection). Finally, the third approach to internet research is the use of translational methods. Translational methods involve the translation of existing psychological research methods to an online environment. For the purpose of this thesis, translational online research methods include the use of online survey software/online recruitment to develop and implement a novel psychological measure and intervention for deepfake abuse (in Chapters 6 and 7). The benefits and downsides to the use of online populations and surveys is discussed further in both papers, but it allowed the research to directly study perpetrating populations in a way that is usually hampered by their anonymity.

1.6 Thesis outline and research questions

The present thesis consists of five research papers which explore the current and potential harms of deepfakes from a Human-Computer interaction and Applied Psychology lens. As a foreword to each paper, the motivation of the work is given along with a brief description of the authors contributions and publishing/review history. The aim of the scoping review (Chapter 2) was to answer some of the fundamental issues with deepfake scholarship around what deepfakes are. Namely, the misunderstanding of what a deepfake is, the assumed novelty of the technology versus existing forms of image manipulation and what current and potentials uses of deepfakes are currently constructed by academic scholars. I felt these concerns had to be dealt with before any empirical work could be carried out. This paper is published in IEEE Transactions on Technology and Society.

The research questions are as follows

1) How do researchers define deepfakes?

2) In what ways do scholars describe deepfakes as being different to other fake media?

3) What current and future uses and implications of deepfake technology are currently observed/ predicted?

Chapter 3 seeks to provide empirical evidence of the harms of deepfakes through a case study. During the scoping review, it had become apparent that there was a lot of focus on hypothetical harms and not the real-life use of the technology. The creation and broadcasting of a deepfake version of the Ukrainian president Zelenskyy during the Russian invasion of Ukraine served as the impetus of the work. The work was published in PloS One. The research questions are as follows:

1. How have people responded on Twitter to deepfake content during the Russian invasion of Ukraine?

2. Does Twitter discourse around deepfakes provide real-world qualitative evidence of the epistemic harms of deepfake technology which have been previously theorised and experimentally studied?

Chapter 4 consists of a follow up work to chapter 3 which builds on the findings of the previous chapter regarding deepfake conspiracies and the design of interventions for deepfake misinformation. I also introduce the concept of deepfake conspiracies. Deepfake conspiracies emerged as a finding in the previous chapter, where we identified a number of conspiratorial claims in our dataset which relied on decrying real videos as deepfakes. I critically reflect one existing deepfake literacy interventions and their tendency to increase scepticism in real media. It was published in ACM Interactions. There were three aims of the work:

1. An introduction to deepfake conspiracies and how increasing awareness of deepfakes may backfire
2. How does this line up with existing work on current conspiracy interventions and misinformation/fake news
3. To provide a critique of current deepfake misinformation interventions and to give possible recommendations for future deepfake misinformation/conspiracy interventions.

The paper reported in Chapter 5 focuses on studying another example of the current harms of deepfakes but changes its focus to that of celebrity women and non-consensual sexual imagery. As my understanding of the harms of the technology grew, I was faced with the stark realisation that deepfake abuse imagery constitutes the dominant harm of deepfakes. The paper is currently under review at ACM Transactions on Computer-Human Interaction and available as a preprint online. The research questions aimed to both understand the experiences of victim-survivors with the abuse itself and the various obstacles to justice:

1. How do celebrity victims of deepfake abuse construct being usees of deepfake technology (being unaware and non-consenting targets of technology)

2. How do they negotiate infrastructural and social obstacles against seeking recourse for deepfake abuse?

My next paper (Chapter 6) seeks to take an active role in developing psychometric tools to study deepfake abuse perpetration. It follows up on one of the findings of my analysis of celebrity interviews: deepfake myths. Deepfake myths are internet discourse which celebrities constructed as undermining their experience as a victim, minimising the harms of deepfakes and justifying similar or related abuse/harassment. I believed that the endorsement of these myths may be essential in understanding both the motivation of deepfake perpetrators and social stigmas which may reduce help-seeking behaviours. I developed a Deepfake Myths Acceptance Scale with the aim to test the validity of the scale and if it did correlate to deepfake abuse perpetration intentions. The research aims of the study were:

1. To test whether the factor structure of the Deepfake Myths Acceptance Scale is valid.
2. To assess whether the measure of deepfake myths correlates with problematic deepfake behaviours as well as other gendered false beliefs (misogyny and myths of sexual aggression).

The final research paper of the thesis (Chapter 7) builds directly on Chapters 5 and 6 by developing an intervention aimed to reduce the endorsement of deepfake myths and the future perpetration of deepfake behaviours. The research question is as follows:

1. Can a values-based intervention reduce deepfake abuse imagery myth endorsement and perpetration behaviours?

The last chapter, the discussion, reflects on the journey of the PhD as a whole and highlights a number of significant contributions of the work. There are also a number of areas of future interest for research in the area which are outlined

CHAPTER 2: What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology

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*.

This paper was published in IEEE Transactions on Technology and Society on the 18th of November 2024

Study Motivation

Before I could study specific uses cases of deepfakes it was vital that I developed an understanding of 1) what deepfakes were 2) how are they new and 3) how can they harm. The definition of deepfakes is a tricky subject as the term originates not from a commercial use case (such as Adobe Photoshop) but the name of a reddit user and subreddit. There has been a disciplinary crisis in the study of deepfakes. While there have been a number of attempts to synthesise a definition, these have not solved the problem (Altuncu et al., 2022). Relatedly, understanding what deepfakes are required also understanding how they are related to other forms of audiovisual manipulation. In many ways deepfakes are just another iteration of popular media editing technologies (such as Adobe Photoshop). We wanted to see if academic literature included fearmongering about “shallow-fakes”, forms of image/video editing which have existed for as long as videos and images themselves have.

Lack of existing psychological work in this area required the thesis to draw from research and theoretical approaches outside of psychology. In particular the decision was made

to take from legal databases because much of the pivotal work in this area has been done by legal scholars anticipating the social and ethical challenges of the new technology.

Author Contributions

This study was conceived by JT, DC and the supervisory team. The study protocol and initial screening were conducted by JT and DC under the overview of CL, GM, MA and MQ. The analysis was performed by JT under the guidance of CL and GM. JT wrote up the paper with guidance from the supervisory team. All authors approved the final version of the manuscript.

2.1 Abstract

Deepfakes are a form of synthetic media that uses deep-learning technology to create fake images, video, and audio. The emergence of this technology has inspired much commentary and speculation from academics across a range of disciplines, who have contributed expert opinions regarding the implications of deepfake proliferation on fields such as law, politics, and entertainment. A systematic scoping review was carried out to identify, assemble, and critically analyse those academic narratives. The aim is to build on and critique previous attempts at defining the technology and categorizing the harms and benefits of deepfake technology. A range of databases were searched for relevant articles from 2017 to 2023, resulting in a large multidisciplinary dataset of 102 papers, 181,659 words long, which were analysed qualitatively through thematic analysis. Implications for future research include questioning the lack of research evidence for the supposed positives of deepfakes, recognizing the role that identity plays in deepfake technology, challenging the perceived accessibility/ believability of deepfakes, and proposing a more nuanced approach to the dichotomous "positive and negatives" of deepfakes. Furthermore, we show how definitional issues around what a deepfake is versus other forms of fake media feeds confusion around the novelty and impacts of deepfakes.

2.2 Introduction

Deepfakes are a form of synthetic media that uses deep-learning technology to create fake images, video, and audio, which are then usually inserted into real footage. Early academic and media discourse around deepfakes has been characterized by conceptual vagueness. For example, the term “deepfake” has been used to refer to a range of different technologies and types of content (Albahar & Almalki, 2019). Early deepfake literature is also characterized by a tendency from authors to predict a wide range of harms upon future proliferation of the technology. Due to the conceptual vagueness of early literature, it is often difficult to understand which of these predicted harms are new or unique to deepfake technology and content. Through reviewing the academic discourse about deepfakes across a wide range of disciplines, this paper aims to provide the reader with a stronger understanding of the many ways that experts see deepfakes impacting upon individuals, organizations, and society.

2.2.1 Critical Qualitative Work on Deepfake Narratives

Previous qualitative research into the harms of deepfakes has focused on understanding news media narratives and how they focus on highly fearful hypotheticals as opposed to practical existing cases. For example, an analysis carried out by Gosse and Burkell (2020) of news media articles found that news media overly focuses on the hypothetical harms of deepfakes as the “ultimate threat to a democratic society”. They also found that news media narratives also tend to speculate on how the average person may potentially be misled by future deepfakes, instead of discussing the existing harms of real deepfakes. The Gosse and Burkell (2020) study highlights the benefit of using qualitative approaches to critically analyse the emerging narratives around this technology. Yet little qualitative research has applied a similar critical lens specifically to academic narratives across disciplines. Most previous review work seeking to understand what deepfakes are and how they cause harm. have often drawn from quantitative analysis or an attempt to synthesize a definition for deepfakes (Altuncu et al., 2022; Whittaker et al., 2023).

2.2.2 Our Scoping Review Approach

A scoping review was chosen as the review methodology as it is appropriate for work which aims to provide a critical qualitative synthesis of academic articles across disciplines. While previous reviews have attempted to synthesize both the impacts and definitions of deepfakes, our research instead critically challenges a broad range of existing academic narratives across multiple disciplines using qualitative analysis. Our research also intends to challenge the definitional issues of deepfake literature. Our review examines what researchers feel is novel about deepfake technology and provides implications as to what current deepfake researchers can learn from the technologies that have preceded deepfakes. We also seek to understand the "positive" and "negative" uses of deepfakes through a more critical multi-disciplinary lens.

2.3 Background

2.3.1 The Origins of Deepfake Technology and Definitional Confusion

The term deepfakes is a portmanteau of 'deep learning' (AI technology) and 'fake' which originates from a member of the Internet forum Reddit who went under the name "deepfakes", and the community that he founded, 'r/deepfakes' (Vasist & Krishnan, 2022a). This individual took a previous Internet practice of photoshopping female celebrities into pornographic photographs ("r/celebfakes") and used modern artificial intelligence technologies to create fake pornographic videos. A typical face-swap deepfake video from the original online community, involves three processes; the extraction of facial data from a source and destination video; training the software to generate a face swap using encoder networks and refining this using generative adversarial networks; and finally, the insertion of the original face from the source video into a destination video (Perov et al., 2020). While the term deepfake originally was used to describe generative AI enabled fake pornography videos, the term has since been used more broadly. It has been described as "digital manipulation of sound, images, or video to impersonate someone or make it appear that a person did something—and to do so in a manner

that is increasingly realistic, to the point that the unaided observer cannot detect the fake" (Chesney & Citron, 2019a). Indeed, while the general understanding of deepfakes includes audio-visual media modified using deep learning, there have been a number of other forms of media referred to by academics as deepfakes such as AI-generated images, text and digital models (Aliman & Kester, 2022). Vasist and Krishnan (2022), in a review of 68 deepfake-related journal articles and their definitions found that apart from the role of sophisticated generative AI techniques in their creation, there is little agreement as to what constitutes a deepfake and urged future review work to develop a precise conceptualization of deepfakes.

2.3.2 Further Review Work on Deepfake Definitions

Subsequent efforts to synthesize a definition in review work and theoretical work have just added to the confusion. In a meta-review by Altuncu and colleagues (2022) of 12 deepfake papers, a much broader definition of the technology was proposed: deepfakes are highly realistic/ natural media of any kind generated by a sophisticated automated method (usually AI). This would include media such as generative AI images of geography and medical scans made by programs such as OpenAI's Dall-E 2 (M. C. Williams et al., 2023). Whittaker and colleagues (2023) synthesized a narrower definition for deepfakes from 80 research articles published between January 2017 and June 2021, defining deepfakes as "synthetic media generated using artificial intelligence and deep learning technology which produce realistic yet fake representations of people undertaking actions or saying words in the form of video, image, or audio content". However, review work which focuses on synthesizing a definition does not critically reflect on the nature and impact of differing disciplinary perspectives on deepfakes. It is also important to question what the existing frameworks offer in the context of differing disciplinary perspectives. Because of these gaps in existing review work it is necessary to carry out our scoping review, which seeks to critically analyse how people are defining deepfakes across academic discipline and criticize the current definitional inconsistencies and confusion.

2.3.3 Defining Deepfakes in Relation to Other Technologies

The definitional issues around deepfake technology are inextricably linked to how it is understood relative to other technologies. In attempting to understand how previous research defines deepfakes, it is also necessary to understand how they are being compared to the technologies which preceded them. In Vasist and Krishnan's (2022) review of deepfake definitions, they also put forward that deepfakes may not constitute a completely novel phenomenon but rather constitute latest in a long history of media editing technologies. If so, it becomes vital to understand how existing responses to deepfake misinformation (false information) and disinformation (false information spread maliciously) can be broadened to include those of similar current and future technologies (we use the term misinformation throughout the paper as it is the broader term, except when referring to deepfake information spread maliciously).

2.3.4 How do Deepfakes Compare to Other Forms of Edited Media?

Deepfakes are often described as "photoshop on steroids" (Dixon Jr, 2019), and current attitudes around deepfakes are linked to the fears which have surrounded earlier image editing technologies. As well as relating to the definitional crisis of deepfakes, the novelty of the technology is important when considering current narratives around the impacts of deepfakes. Some critics have pointed out that most novel technologies have been associated with similar fearmongering at their onset (Broinowski, 2022) and early discourse tends to focus on fearful possibilities as opposed to any evidenced harms (Taylor, 2021). Many of the other harmful uses of deepfakes can also be seen as continuations of traditional forms of video and image editing. For example, the original deepfake community replicated the already-existing use of photoshop towards female celebrities and extended them to video, while reducing the technological/ artistic ability required. The question of what makes deepfakes different from previous media editing technologies such as photoshop has not been interrogated in detail in previous review work. Our scoping review seeks to tackle the lack of existing literature and review work considering the novelty of deepfake technology. It is essential for research and policy on deepfakes to consider

whether deepfakes are a novel phenomenon or a symptom of larger problems with novel media editing technologies and society, especially with advances in generative AI. Otherwise, one risks designing solutions for a technology and not for the broader harms of media editing technology wholesale. If the harms of deepfakes are not novel then we need to focus on placing deepfakes within existing work and taxonomies of harms such as privacy, election interference and cybersecurity.

2.3.5 Narratives Around the Current and Future Uses of Deepfakes

The original deepfake community was banned on Reddit as a result of public outrage brought about by news coverage of the practice (Wilkerson, 2021). However, forums dedicated to the making and sharing of deepfakes still exist today (Timmerman et al., 2023). As a result, both news media and academic organizations have expressed concerns about the potential harms of deepfake technology. In their seminal paper, Chesney and Citron (2019) put forward their view of the key potentials for harm, and benefits, from deepfake technology (summarized in Table 2.1). The nonconsensual pornographic related harms of deepfakes are the most immediate and widespread, with pornographic content representing approximately 96% of deepfake videos currently online (Kerner & Risse, 2021). Outside of pornographic content, deepfake technologies have demonstrated a host of uses, both benign and malicious (Liv & Greenbaum, 2020). Fuelled by the fear that deepfakes make more believable fake news than previous technologies, journalistic discourses around deepfakes often involve conjecture around the future use of deepfakes by “resource-rich bad actors” (Wahl-Jorgensen & Carlson, 2021). Notably, some commentators have characterized the typical academic, institutional, and journalistic responses to deepfake technology as “techno-paranoid narratives” (Broinowski, 2022) and as a “moral panic” (Cover, 2022). This stems from the asymmetrical coverage of deepfakes in news media which focuses on the potential harms as opposed to the current uses of deepfakes (Karnouskos, 2020). Indeed, academic discourse around deepfakes tends to focus on the negative impacts of technology (note, for example, the disparity between positive and negative examples in Table 2.1).

Table 2.1 The key beneficial/ harmful uses of deepfake technology according to Chesney and Citron (2019)

Uses of Deepfakes	Explanation
Positive Uses of Deepfakes	
1., Education	Educational deepfake videos for the classroom
2. Art	Creative uses often involving the resurrecting of actors.
3. Autonomy	Digital self-expression – particularly for people suffering with paralysis or other restrictive health conditions
Negative Uses of Deepfakes	
1. Individual and organizational harms	
Exploitation	Theft, abuse, and blackmail facilitated by deepfakes
Sabotage	Sabotaging the reputation of others using defamatory deepfakes
2. Societal harms	
Distortion of democratic discourse	Deepfake misinformation may harm political discourse
Manipulation of elections	Use of deepfake videos to attack election candidates
Eroding trust in institutions	Prevalence of deepfake videos will damage trust in institutions
Exacerbating social divisions	Deepfake misinformation might encourage and stoke social issues, dividing and polarizing communities.
Undermining public safety	Deepfakes could encourage public panic
Undermining diplomacy	Deepfakes may inflame relations between countries
Jeopardizing national security	Deepfakes could damage the reputations of militaries
Undermining journalism	News organizations may be challenged to verify deepfake misinformation/ disinformation.
The liar's dividend	Deepfakes could make it easier for liars to deny truth

2.3.6 Critical Review Work on the Uses of Deepfake Technology

A systematic review using network analysis carried out by Gamage and colleagues (Gamage et al., 2021) on the societal implications of deepfake technology highlighted the diversity of possible harms and threats to society in the literature. They found that security-related harms were dominant of these, with psychological, legal, and political harms also

emerging. Similarly, research into the media coverage of the technology has also highlighted the dystopian and utopian narratives around the technology replicating previous hysterias over the emergence of technologies such as radio and television (Broinowski, 2022). The media focus on deepfakes has also tended to centre on the implications for knowledge and truth, despite pornographic deepfakes making up the majority of deepfake related content. By relegating deepfake pornography to the history of deepfakes, and by not prioritizing it as the most pertinent of deepfake harms, news coverage of deepfakes devalues the current harms of deepfake pornography (Gosse & Burkell, 2020). Many technical review papers for deepfakes have produced summaries of the positive and negative uses of deepfakes (Albahar & Almalki, 2019; Dagar & Vishwakarma, 2022; Mahmud & Sharmin, 2021; Rana et al., 2022; Westerlund, 2019), generally focusing on the effects of deepfakes on entertainment, fake news, and the value of video evidence. However, there have been few reviews which have critically examined the contents of those arguments using a multi-disciplinary and qualitative approach.

As well as highlighting the definitional issues of deepfakes in their previously mentioned review, Vasist and Krishnan (2022a) also found what they call “divergent perspectives” across different academic disciplines. These were differences in opinion as whether or not deepfakes are harmful and the accuracy of overly fearful journalistic concerns. They found that some disciplines were split between discussing both the positive and negative impacts of deepfakes on their discipline (e.g., advertising) and they found some positive opinions around the technology in research fields which were at risk of the harms of deepfakes. Vasist and Krishnan also called for and subsequently carried out a review of existing qualitative deepfake literature (Vasist & Krishnan, 2022b) and developed a conceptual framework of deepfake engagement using social shaping of technology theory. Their framework provides a comprehensive understanding of the socio-technical elements which underpin the fabrication of the videos as well as their dissemination/ sharing. However, their research is also predicated on deepfakes being classified into either positive or malicious deepfakes. There seems to be little work done in questioning this dichotomous classification. The current scoping review intends to

build on existing review work which state the pros/ cons of deepfakes across multiple disciplines and to engage more critically with the impacts of deepfakes and to question what exactly is so beneficial/ harmful about them.

2.3.7 Aims and Research Questions

The emergence of deepfake technology has prompted responses from a wide range of academic disciplines including computer science/ information systems, humanities, social sciences (such as psychology and sociology) politics, and law (Vasist & Krishnan, 2022a). There is little consistency between how these papers define deepfakes, in what ways they describe the technology as novel or different, and how they envisage the potential impacts of deepfakes. While the early response to deepfakes typically tended towards sensationalism, as with the introduction of each generation of new media (Bazarkina & Pashentsev, 2020), we are at a point where it is important to clarify whether there is any good reason why we should consider deepfakes as uniquely harmful or worrying. We have chosen to carry out a scoping review over a systematic review as we seek to study a large wealth of papers across disciplines with differing or no study design (Arksey & O'Malley, 2005). Our scoping review seeks to collate and review the specific concerns and predictions across disciplines to understand whether there is something qualitatively different about deepfakes and their harms from previous media editing technology. The goal of this review is to answer three questions:

1. How do researchers define deepfakes?
2. In what ways do commentators describe deepfakes as being different to other fake media?
3. What current and future uses and implications of this technology are currently observed/ predicted?

2.4 Methods

We chose to use a scoping review for this research as it is useful in reviewing academic literature more broadly than other methodologies. Specifically, we focus not on the quality of specific empirical papers but instead to summarize and disseminate broader research findings, to more broadly examine the themes of academic literature and to identify where existing research, reviews and theory are falling short (Arksey & O'Malley, 2005). The search was carried out on the 11th of January 2023 across seven databases, limited to English language journal articles published between the 1st of January 2017 and the 11th of January 2023. The search terms used were (deepfake* or deep-fake*) within the categories of title, abstract or keywords. These search terms were kept intentionally broad to include the largest possible amount of deepfake articles. Similarly, the databases were chosen to gain a wide, multi-disciplinary dataset across the relevant research domains. The databases chosen were Web of Science. SCOPUS, Hein Online Legal Databases and three EBSCO academic databases (Academic Search Complete, Business Source Complete, APA Psychinfo). The protocol of this scoping review was preregistered on the Open Science Framework (OSF) website (https://osf.io/xc7mk) and the findings are reported using the PRISMA-SCR checklist (Figure 2.1).

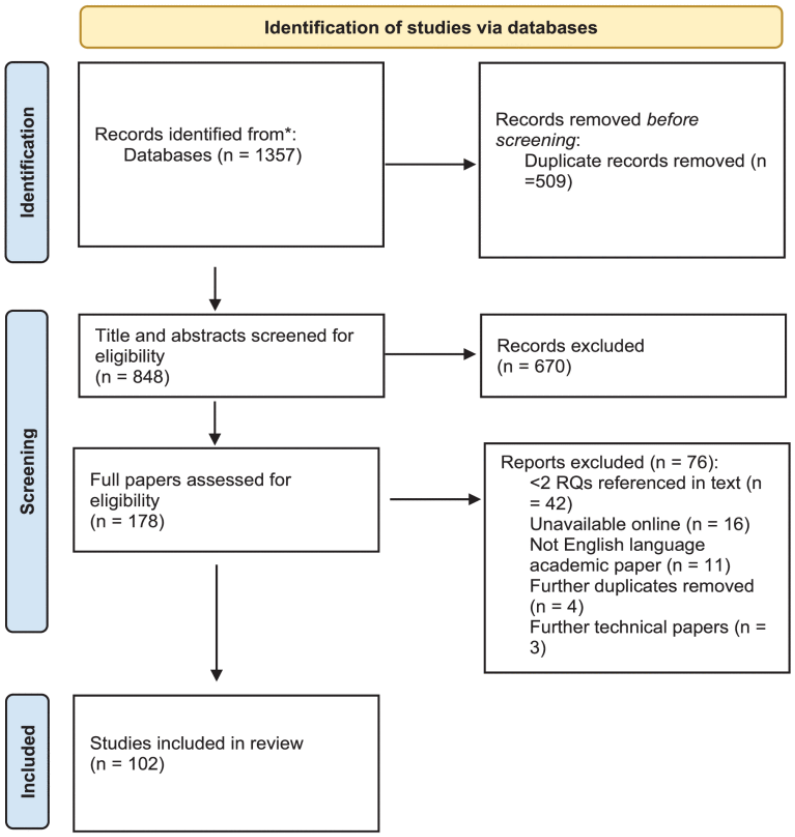


Figure 2.1: Prisma flow diagram for search results

2.4.1 Screening Process

To be included in the scoping reviews, papers had to come from a relevant academic discipline such as Information Systems, Computer Science, Law, Sociology, Philosophy, Media, or Psychology. We excluded approximately 261 papers that were solely concerned with deepfake detection/ mitigation or production, as this has already been carried out comprehensively by previous deepfake review papers (e.g. Whittaker et al., 2023). We did not search for grey literature on this topic due to our review's specific focus on academic narratives. Papers published before 2017 were also excluded as part of the search as they predate the emergence of deepfakes. At this stage, most of the duplicates were identified based on title and removed. The Rayyan open-source software for systematic reviews was used to store the data and two reviewers used the software to screen the papers. Papers were first excluded where the abstracts and/ or title did not mention at least one of the three extracted topics (definitions of deepfakes, comparisons with other technologies, and potential effects). Papers were then

excluded if they did not contain two or more of the three variables to be extracted in their full text.

2.4.2 Data Extraction

There were three qualitative variables extracted for the scoping review: definitions of deepfakes, novelty of deepfakes/ comparisons with other technologies, and potential effects of deepfakes. To be included in our dataset as a definition, papers had to mention deepfakes as a form of media/ content or a technology which produces media/ content which reflects both the original 2017 deepfake content and definitions for the technology which have been established by previous reviews (Altuncu et al., 2022; Whittaker et al., 2023). For text to be included in the analysis of novelty, the paper had to mention deepfakes as novel, position the technology in comparison with other technologies/ forms of media, or talk about other technologies in the text in the context of deepfakes. Finally, impacts had to be concerned with what deepfakes have, can, should or will do. The final dataset consists of 102 papers and is 181,659 words long (a full list of papers included can be found on the project's OSF page https://osf.io/bxnds).

2.4.3 Data Analysis

For our analysis we broadly follow the steps outlined by Braun and Clarke (Braun & Clarke, 2012) for conducting and reporting thematic analysis. We coded each extracted textual unit or paragraph, as opposed to coding the text line by line due to the size of the dataset, and used these codes to generate our themes. This method was derived from how other methodologies (such as applied thematic analysis; Guest et al., 2011) describe coding larger textual blocks. While time-intensive, this method was invaluable as a way of qualitatively synthesizing a broad range of articles. As for similarly intensive qualitative review methodologies (Stokes et al., 2023), this meant that we were not able to include more recent work, though we did keep abreast of relevant review work published during the analysis. We then related the resulting themes back to the data corpus as well as the existing theoretical frameworks, examining differences and common factors. While entire paragraphs were coded,

when reporting the data in the analysis we went back over the codes to gather illustrative examples of the codes for the reader, similarly we also gathered word counts from the different themes as a way to use corpus data to further illustrate for the reader how often certain concepts emerged.

2.5 How are deepfakes defined in the literature?

2.5.1 Deepfake Definitions Across the Dataset

We found 86 separate definitions of deepfakes in our dataset. Of those definitions, some were cited repeatedly (Supplementary Materials). Specifically, two definitions were cited over five times in the dataset: 1) Westerlund (Westerlund, 2019): “A combination of “deep learning” and “fake”, deepfakes are hyper-realistic videos digitally manipulated to depict people saying and doing things that never actually happened” 2) Chesney and Citron (2019): “It [Deep-fake technology] leverages machine-learning algorithms to insert faces and voices into video and audio recordings of actual people and enables the creation of realistic impersonations out of digital whole cloth. The end result is realistic-looking video or audio making it appear that someone said or did something” One essential characteristic that both these definitions have in common is in how deepfakes can be used to represent someone doing or saying something which never actually happened. The capacity of deepfakes to mimic and appropriate identity was an integral part of how the impacts of the technology related to the definition in our analysis. Similarly, there was an assumption of untruthfulness around deepfake technology, and the technology was seen as manipulating media for the purposes of falsification and fabrication. Deepfake technology was often defined by its ability to blend the fake into the real, through interpolation, as in face-swap videos.

2.5.2 Common Factors of Definitions

Our analysis identified a set of common factors across deepfake definitions (Figure 2.2). As expected, we found that deepfakes were generally defined by their use of deep-learning technology, a subset of machine learning and artificial intelligence (e.g., “a technique for human

image synthesis that employs artificial intelligence and deep learning technology to produce false yet deceptively realistic images and videos"), though definitions shared little detail on the technical process underlying the technology (Conte, 2019). Some papers limited the modality of deepfakes to videos, in line with the original deepfakes published in 2017: "At the most basic level, a deepfake is created by taking an input source of videos or other images of an individual and outputting a manipulated video of that individual's face" (Langa, 2021). However, deepfakes were more broadly viewed as consisting of video, image or audio and often multi-modal, involving audio and video. The broader definition was justified as being necessary in light of developments in deepfake technology to capture "the full range of hyper-realistic digital falsification of images, video, and audio" (Chesney & Citron, 2019a). There were references to deepfake text, referring to AI-generated text such as that created by large language models (e.g., Anna, 2021), but this was not common. Considerations of the modality of deepfakes ended up drawing from comparisons of the technology to previous media editing technologies, connecting these two research questions, Deepfakes were often presented as one of a host of synthetic media technologies (e.g., "Deepfakes are a new form of synthetic media that broke upon the world in 2017"; Kerner & Risse, 2021).

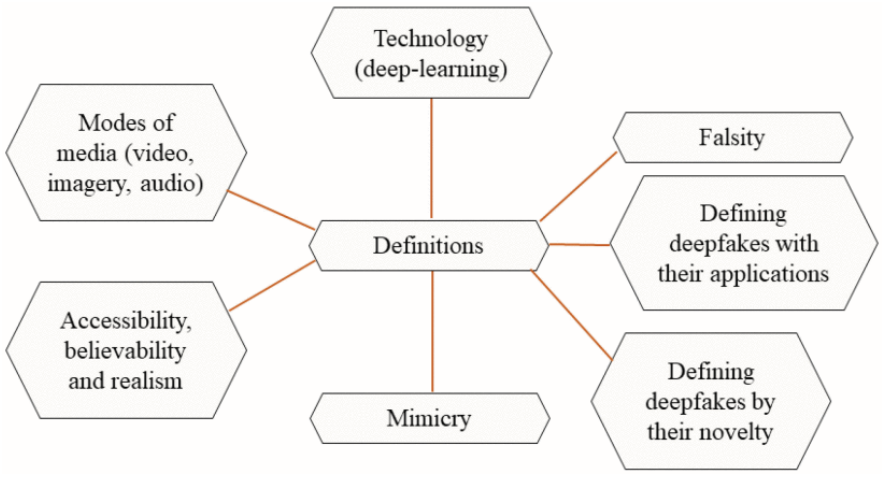


Figure 2.2: Themes drawn from deepfake definitions.

2.5.3 Points of Confusion Around Definitions

The need for greater clarity and precision in deepfake definitions became apparent when analysing the impacts of the technology. The dataset included a mix of real-world examples of the impacts of deepfakes and projections of potential future impacts. Confusingly, some examples given in the dataset regarding the real-world impacts of deepfakes referred to traditionally edited media instead of, or along with, incidences of deepfakes. “In 2017, researchers successfully created deepfake stills of former President Barack Obama saying things that he never actually said but were voiced by actor, Jordan Peele. (deepfake). The same technology can create false videos inciting plans to carry out political assassinations, seemingly private conversations featuring elected leaders, or Speaker Nancy Pelosi slurring her speech (not a deepfake, just slowed down video)” (Nour & Gelfand, 2022). Using examples of the harms of photoshop and rudimentary video editing technologies occasionally made sense, as these examples were used to illustrate the current harms that deepfakes evolve upon, however in some cases the traditional examples were used interchangeably with deepfake examples. Worryingly, some papers used a combination of real examples of deepfakes and traditional CGI/ photo editing without clarifying the differences between the technologies. Switching between examples of deepfakes and other false media has the effect of exacerbating the definitional issues of deepfakes and highlights the need to clarify what exactly defines deepfakes in comparison with other technologies.

2.6 The novelty of deepfakes

2.6.1 Accessibility and Believability: The Two “Mores” of Deepfakes

Two concepts were consistently observed in our data set, where authors discussed the novel nature of deepfake technology: accessibility and believability (Figure 2.3). Many viewed deepfakes as being alike to previous technologies but “more” in a certain way, particularly around the technology’s accessibility (e.g., “In the last four years, deepfakes have developed from an academic project to a rapidly evolving and accessible form of technology”; Krishna,

2020) and believability (e.g., “Concerns that the audiovisual nature of deepfakes makes them inherently more believable than previous fake news forms leads to worries over the impending weaponization of deepfakes by resource-rich ‘bad actors”’ (Wahl-Jorgensen & Carlson, 2021). In discussing the accessibility of deepfake technology some authors emphasized how comparatively easier it is to produce deepfake videos than equivalent methods of fake-video production. These narratives often revolved around the cost-free availability of open source deepfake technology to the tech savvy Internet users and the availability of consumer deepfake applications. “While Hollywood has for many years been able to generate synthetic content, it has been very cost, time and resource intensive. However, the technology associated with deepfakes requires a lower threshold of skills and time, and the resources used to create such materials are readily available and accessible to individuals globally” (Cross, 2022). The focus on the average person having access to deepfake technology highlights fears around the democratization of harmful AI technologies. While the general view of deepfakes is that they are comparatively easier to make than other forms of edited videos, it must also be noted that realistic deepfake videos still require specialized technical skills.

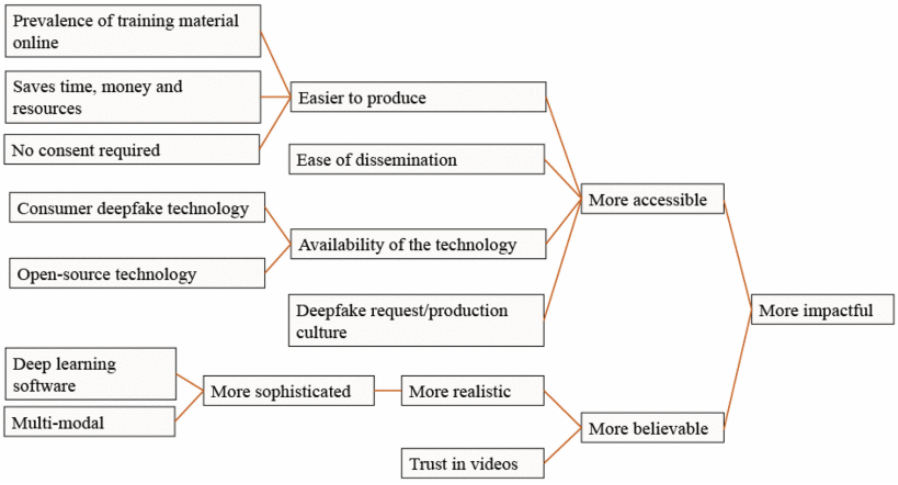


Figure 2.3: Themes and subthemes relating to the accessibility and believability of deepfakes

2.6.2 Broader Impacts of Accessibility and Believability

The realism and believability of the technology exacerbate both the social harms of the technology (through misinformation) and the potential for reputational damage (as most often

reflected in the harms of deepfake pornography). These characteristics were seen as integral to the technology's unique capacity for harm. The harms were frequently described either as novel or, less often, as an evolution of the similar harms of previous technologies. Deepfakes harms are novel in the way they create realistic faked faces with less effort than previous technologies: "most everyday manipulation or remixing fell short at generating moving images and sound that could convince a relatively discerning viewer that it was recorded footage" (Cover, 2022). Deepfakes have the capacity to manipulate both audio and visual data and as a result was seen as "better" than previous techniques such as CGI and Photoshop which are visual editing technologies. However, similar technologies such as photoshop have been commonly used for defamatory content, pornography, and misinformation for two decades prior, so deepfakes mostly serve as "the latest iteration of a longstanding practice" (Molina, 2021). The positive uses of deepfake were similarly discussed as being facilitated by the accessibility of the technology. While talking about the entertainment usage of digital recreations and recasting, papers were quick to point out how time-consuming and poor-quality previous manual attempts have been. "The deepfake technology available to consumers superficially resembles Hollywood special effects, the sort where dead actors are revived anew onscreen. But these professional techniques are extremely expensive and time consuming. What makes deepfakes remarkable is that they provide an approximation of the same effect, much more cheaply and quickly" (Rini, 2020). There is also a temporal element to the availability and realism of the technology, as the years since early deepfake research have seen an increasing amount of cheap or free deepfake applications. Few papers considered that being more accessible and believable than previous video editing technologies does not mean that deepfakes themselves are very realistic or very accessible. Notably no papers presented any evidence that deepfake images (not videos) are more realistic than traditionally manipulated images.

2.6.3 Comparisons With Other Technologies

There were three different ways in which papers attempted to relate deepfakes to other media technologies; the deepfake – shallowfake dichotomy, the cheapfake-deepfake spectrum,

and as part of a taxonomy of synthetic media (Table 2.2). The deepfake-shallowfake dichotomy (Gregory, 2020) is a means of classifying media editing based on whether deep-learning technology is involved. “Deepfakes should be distinguished from shallowfakes, which are also manipulated media, but manipulated through human intervention rather than artificial intelligence” (Sylvester, 2020). The deepfake-cheapfake spectrum (Paris & Donovan, 2019) organizes media technology on a spectrum based on the sophistication of the technology used to create the fake media. “At the other end of the deepfake spectrum, less sophisticated actors can create “cheap fakes,” or lower-quality deepfakes, using consumer-grade software or simple video-editing techniques” (Geddes, 2020). The terms “cheapfake” and “shallowfake” are notable, as many forms of traditional editing can be quite complex and expensive and is neither cheap nor shallow. It is also interesting to note how these two categorizations interact with the ideas of deepfakes being more believable and accessible. The view of previous technologies as “shallowfakes” lines up with the increased sophistication of deepfake technology (and hence realism and thus believability). In contrast describing previous technology as cheap is inconsistent with the view of deepfake technology being a more accessible and cheaper way to manipulate videos. These definitions/ comparisons serve an important purpose; delineating deepfakes from the spectrum of prior media editing technologies and thus establishing its novelty. However, these two terms often muddy the waters of deepfake definitions, especially as they refer to forms of image manipulation that have been defined for decades prior. Finally, deepfakes were often characterized in relation to other so-called “synthetic media” (Barnes & Barraclough, 2020) (as opposed to recorded media). Deepfakes are a means of producing a type of audio-visual synthetic media using deep learning technologies.

Table 2.2: Approaches to organizing deepfakes in comparison to other media

Typology	Source/sources	Explanation
Deepfake – shallowfake dichotomy	(Gregory, 2019)	Deepfakes are made with sophisticated AI, shallowfakes are not. Shallowfakes often consist of videos which have been relabelled and re-uploaded with a different context. If shallowfakes are manipulated, it usually involves simple edits, or a new audio track.
Cheapfake-deepfake spectrum	(Paris & Donovan, 2019)	The authors propose a spectrum charting audio-visual manipulation techniques based on their technical sophistication, barriers to entry, and techniques. In their most complex and expensive form, deepfakes include completely virtual and realistic performances. "Cheaper" techniques include the typical lip-synching and face-swapping deepfakes. At the other end of the spectrum, "cheapfakes" at their least complex consist of recontextualized videos. At their most complex they involve rotoscoped face-swapping and sped-up/ slowed down videos.
Deepfakes as part of a taxonomy of synthetic media	(Barnes & Barraclough, 2021; Millière, 2022)	Synthetic media can be partially or totally synthetic. Partially synthetic media involves the modification of archival media with new media created using a generative component. Totally synthetic audio-visual media is not produced by modifying pre-existing archival media but instead involves generating new audio or visual media. Deepfakes originally referred to face-swapping videos but now refer to a wide range of deep-learning based audio-visual manipulations which can be partially or totally synthetic. Limiting deepfakes to these taxonomical distinctions misses the effect deepfake technology may have across all synthetic media

2.7 The existing and future uses of deepfakes

2.7.1 Potential Uses of Deepfakes

We identified a number of potential harms and uses raised by experts across the disciplines in our dataset. Through our analysis, we classified these into three chief types of harms:

1. Personal harms of defamatory pornographic deepfakes, particularly harassment and objectification.

2. Societal harms of defamatory political deepfakes, enacted through social manipulation.
3. The epistemic harms of deepfakes to video (harms to truth and evidence).

The dominance of these three types of related harms is apparent in the corpus (Table 2.3), showing a focus on terms referring to the pornographic deepfakes (e.g., pornography, sexual, women), political deepfakes (e.g., political, election, disinformation) and the epistemic harms of deepfakes (e.g., trust, evidence, epistemic). In contrast, a small minority of papers suggested that there may also be political and societal benefits. For example, improved scepticism in light of misinformation deepfakes and more trust in established and reputable politicians/ media (Langa, 2021). Some papers expressed that deepfakes would empower political activists through satire (van der Sloot & Wagensveld, 2022) and noted the potential use of deepfakes in advocacy videos by charities (Kerner & Risse, 2021). However, these papers were in the minority. While some papers in the dataset discussed the positive uses of deepfakes, this was not common, as reflected in the corpus data. This illustrates either (or both) the dominance of dystopian narratives around deepfake technology, or the lack of many real-world examples of the potentially positive uses unique to the technology. Narratives around the current and potential use of deepfakes were generally characterized by warnings that the technology has caused or will cause more harm than good. “One might well leave this discussion with a lingering sense that the potential for more damage is enormous and already rather concrete, whereas the benefits are much more uncertain” (Kerner & Risse, 2021). Other papers proposed that reasonable responses to deepfake technology involved balancing the positive and negatives. A minority of papers expressed that the technology was not as bad as people believed “The reality is that as the technology has progressed, a different picture of how deepfakes will be used has emerged. They may not be such a problem at the highest levels of security and defence. Instead, the problems that deepfakes pose lie at the lowest levels” (Ascott, 2020). There were also papers which expressed the opinion that political and news media concerns around deepfakes consisted of a moral panic and commented on the extreme harms of deepfakes as

being a mostly hypothetical and unevidenced. “Past experience suggests that “the panic around” deepfakes will turn out to be a “false alarm,” as one commentator recently concluded. As far as the courts are concerned, I predict that this panic will prove largely unfounded. The nation’s courts are robust institutions that have shown themselves capable of handling each new variant of the age-old problem of fakery” (Pfefferkorn, 2019a). The victims of deepfake harms in the text were split between individuals and organizations/ institutions. Public figures emerged as a significant target of deepfakes. In line with the general focus on political and pornographic harms of the technology, as well as gendered narratives around the technology, these public figures tended to be male politicians and female celebrities (though there were some exceptions to both).

Table 2.3 Wordcounts of deepfake-related words in the impacts section of our dataset.

Deepfakes	1987	Epistemic	107
Deepfake	1345	Impact	106
Video	877	Victims	104
Technology	750	Voice	103
Videos	576	Data	103
Media	487	Users	100
Fake	483	Recordings	98
Public	353	Uses	98
Social	346	Intelligence	98
Content	342	Security	97
Political	307	Problem	97
Deep	303	Nonconsensual	97
People	288	Footage	96
News	265	Spread	96
Potential	252	Citron	96
Information	251	Individual	95
Fakes	242	Internet	95
Evidence	235	Possible	94
Online	216	Effects	93
Pornography	212	Legal	92
Digital	208	Threats	91
Real	195	Speech	90
Harm	193	concerns	90
Individuals	190	Chesney	90
Images	183	Fact	90
Threat	163	Journalists	90
Person	158	Abuse	89
Pornographic	157	Virtual	89
Face	156	Reality	87
Trust	154	Future	87
Porn	147	Misinformation	87
False	147	Manipulated	85
Audio	144	Artificial	85
Sexual	140	Revenge	85
Women	130	Sex	84
Image	129	Elections	83
Technologies	127	Undermine	83
Disinformation	126	Truth	82
Ai	125	Malicious	82
Actors	119	Manipulation	81
Society	116	Realistic	81
Election	115	Different	80
Time	115	Opportunities	79
Harms	114	Negative	79

2.7.2 Personal Harms of Defamatory Deepfakes

Defamatory deepfakes were associated with a series of personal harms. They undermine the consent of the parties in the videos and are then used to harass and defame the targeted individuals. This was clearly exemplified by the use of deepfakes to generate pornographic

videos of female celebrities. In many ways, this harm was seen as an evolution of previous forms of image-based sexual abuse, now being able to target anybody (Gosse & Burkell, 2020). It was also seen as part of a larger pattern of technology being used to invade female celebrities' privacy and to silence activists (Maddocks, 2020). The use of pornographic deepfakes to intimidate, harass, objectify, and humiliate targets was identified commonly as one of the biggest threats of deepfakes. These defamatory harms of deepfake technology were associated with significant adverse personal responses. These include 1) the loss of control over how one's identity is visually represented as well as 2) the disempowering potential of the technology as part of the social and individual consequences of deepfakes. "Much like the painful accuracy of non-consensually disclosed pornography, the misrepresentation of deepfake pornography impacts one's ability to control their sexual identity" (Kugler & Pace, 2021). In many papers that were themselves unrelated to pornography, it was still referred to as part of the origin of deepfake technology. "While deepfakes originated in the pornography industry, there are growing concerns about how they could interfere with politics and elections and threaten individual privacy" (Wilkerson, 2021). The reputational damage of deepfakes can potentially cause severe financial consequences and career damage if the fake content reaches employers. Defamatory deepfakes can also harm an individual's self-image, forcing them to doubt and question themselves in the face of convincing visual evidence.

2.7.3 Social Manipulation Using Defamatory Deepfakes

The other prevalent harm involved the use of deepfakes in social manipulation, particularly the use of defamatory political deepfakes to spread lies and misinformation. Many papers argued that deepfakes give people an increasing number of tools for imitating public figures and tampering with existing videos. These concerns often focused on the harms to governments, online platforms, journalism, and news media and business organizations. "Deepfakes can be used to spread fake news, influence elections, introduce highly realistic fake evidence in courts and make fake porno movies. Each of these applications potentially has a big impact on society, social relationships, democracy and the rule of law" (van der Sloot &

Wagensveld, 2022). Many papers also considered the cognitive effects of socially manipulative deepfakes, discussing the use of deepfakes in misinformation and how they may encourage confusion, reinforce previous distrust and conspiratorial thinking, perpetuate biases, influence individuals' ideologies, and induce false beliefs/ memories (Langa, 2021; Liv & Greenbaum, 2020; Murphy & Flynn, 2022). Other concerns focused on hypothetical scenarios of companies' CEOs being defamed, leading to their market value plunging (Mustak et al., 2023). There was also a concern around existing financial crimes such as blackmail, identity theft and fraud which could be augmented using deepfake technology. AI-generated fraud phone calls emerged as a distinct harm of deepfakes in this field due to a small number of real-life examples where fake video calls were purportedly used to extort millions from companies (de Rancourt-Raymond & Smaili, 2022). Most papers, particularly law and politics papers, focused on institutional harms and particularly the destabilizing effect of deepfakes on politics, international relations and democracy. Deepfakes may spread disinformation through its capacity for novel forgery and identity theft, augmenting text-based disinformation with fake profile pictures and deepfake videos. Or, with the advancement of real-time deepfake technology, there is a possibility of deepfake video calls being used to spread disinformation and commit fraud (Anna, 2021).

2.7.4 Socially Manipulative Deepfakes in Conflict

There were fears that socially manipulative deepfakes would be utilized in future conflicts as a weapon for sabotage or as propaganda. Some papers also expressed a fear that deepfakes would be used to garner supports for terrorists or to be used for terrorism in the form of cyberattacks. The most extreme securitization fear around the technology involved the hypothetical use of deepfakes to start a war through deepfake misinformation. "In our present age of misinformation, society will soon have to deal with deepfakes that can threaten national security. Consider a deepfake of President Trump announcing impending nuclear missile attack on North Korea "(Harris, 2018). Often these worries were characterized by fears of "bad actors", usually political agents or in many cases so called "foreign actors". One paper alluded to the military benefits of deepfake technology in military engagements as a means of creating

propaganda and fake news (Langa, 2021): “In the military and national security context, the same Al technology used to create deepfakes could be repurposed to assist intelligence agencies in surveilling hostile areas by detecting human threats with facial recognition software, thereby countering guerilla warfare and insurgency. The U.S. military could also use Al technology to achieve a strategic advantage over enemies with lesser capabilities”. Similarly, some papers viewed deepfakes as giving people a new means of political expression through deepfake political satire. This highlights the perspectival element of deepfake harms and benefits where the positives of deepfakes are often seen as negatives from different points of view.

2.7.5 Epistemic Harms of Deepfake Videos

A third, broader harm in the dataset involved the epistemic harms of deepfakes. Deepfakes were seen as a threat to the general public’s trust in video evidence. Some papers expressed that video records are anomalous in human history as previous forms of novel media have all been shortly followed by the ability to be manipulated and forged. These epistemic concerns related to the broader societal and philosophical impact of deepfakes, particularly how deepfakes threaten truth by undermining faith in videos. “While videos and audio remain powerful pieces of evidence, their authenticity will be hard to gauge in the era of deepfakes” (McPeak, 2020). In turn the damaged trust in videos undermines trust in documentary records, evidence, institutions, and governments (Geddes, 2020). Similarly, deepfakes were seen to challenge the freedom of activists and protestors, especially through undermining low-quality videos and thus challenging the validity of civilian journalism. “As deepfakes become easier to generate, and more common, media outlets may be well-advised not to take for granted the accuracy of video footage from unknown sources. In this way, some video footage will be unsuitable as evidence” (Harris, 2018).

2.7.6 Epistemic Harms of Deepfake Narratives

The epistemic harms of deepfakes have similar implications to the political harms of deepfakes, as they both draw from the undermining of trust in politicians and institutions/

organizations. However, it should be noted that much of the epistemic harms of deepfakes drew from the harms of the narratives around deepfakes as opposed to the actual use of deepfakes. Legal papers tended to focus on this harm, expressing worry about the damage of this lack of trust to courtrooms, by devaluing video evidence and forcing legal teams to have to use expensive computer forensics to decide whether a technology is a deepfake or not (Pfefferkorn, 2019b). The so-called 'liar's dividend' emerged from legal discourse as a unique harm of deepfakes (Chesney & Citron, 2019a). This is the effect of epistemic doubt around videos as providing liars a new form of deniability. With the epistemic damage of deepfakes, future politicians may have to regularly deny deepfake misinformation. Many papers highlighted the role of deepfakes in facilitating lies and liars. However, there were fewer empirical papers which explored the extent to which deepfakes may exacerbate or cause epistemic damage. While the harm of deepfakes to truth and trust fuel academic narratives around the negative aspects of deepfakes, there is little real-world evidence given for these harms.

2.7.7 Artistic, Educational and Self-Expressive Deepfakes

The positive uses of deepfakes, expressed across disciplines, focused on creative and artistic uses or educational uses. Deepfakes allow the creation of media that depicts things that did not really occur, or were at least not filmed, and while this has a significant impact on truth and misinformation, it will also increase what people are able to creatively express. The use of deepfakes allows users more creative expression in how they can present humorous content, as well as resurrecting dead celebrities, gender swapping and exploring alternate realities. For example, a deepfake audio of U.S. president John F. Kennedy of the speech he was planning to make before his assassination or a deepfake of Nixon presenting a speech he would have made had the Apollo 11 astronauts died (Mustak et al., 2023; Reynolds, 2020). Deepfakes were seen as having the potential to influence entertainment and art, either through the use of deepfake technology in online content and memes or to improve fictional films and TV. In papers which focused on media/ entertainment, Deepfakes were seen as encouraging new practices in Hollywood (Holliday, 2021), where films and media could be changed significantly by using

deepfakes to recast actors and to digitally resurrect deceased actors (Kerner & Risse, 2021). “Created by a number of online video artists, these convincing ‘mash-ups’ playfully rewrite film history by retrofitting canonical cinema with new star performers” (Holliday, 2021). There were still harms noted in these cases, particularly concerning the ethics of appropriating the identity of dead actors to feature them in movies without their consent. This was particularly relevant for deepfaked celebrity advertisement, where there was also the potential for false endorsements and the disrespect of using dead celebrities to market products. Further criticisms highlighted how the positive entertainment uses of deepfake tended to target less marginalized individuals versus the harms. “The many creative involvements with Cage’s star image and his emergence as a regular target for Deepfakes technology has further surfaced how the cyberstars of such ‘take two’ cinephilic activities are typically cisgender white males” (Holliday, 2021).

2.7.8 Economic Benefits of Educational Deepfakes

The positive economic effects of deepfakes in tourism/ marketing literature drew from the belief that educational historic deepfakes were one of the main positive uses of deepfakes. The widespread incorporation of these were felt to potentially increase visitor engagement and thus tourism revenue for nations, towns, and businesses (Shin & Lee, 2022). Deepfakes were generally seen as having educational benefits including the use of deepfakes in museums or schools to make educational videos more immersive. Some papers problematized this however, stating “Some perceive the idea that Napoleon would give a history lesson in secondary schools a nice prospect, others find it problematic because children would be confronted at an early age with blurred lines between fact and fiction” (van der Sloot & Wagensveld, 2022). No paper questioned the extent to which the educational role of deepfakes could be similarly accomplished with impersonators in costumes and there was no empirical evidence given for the education benefits of deepfakes. Positive educational narratives also focused on the use of deepfakes to create easy lip-synched multilingual videos (Diakopoulos & Johnson, 2021). This has the potential to make video more accessible, reducing the cost of having to remake videos in different languages and providing more support for lip-reading than traditionally dubbed media.

One positive element of the ability of deepfakes to erase the identity of the source was its use in anonymization, particularly for documentary movies. “Deepfakes can be used to empower the marginalized to allow them to share their stories using authentic emotion without fear of identification. Welcome to Chechnya, a documentary which explores the persecution of LGBTQ people in Chechnya, uses deepfakes to anonymize the subjects of the documentary by transplanting the faces of New York-based activists onto the faces of the Chechen documentary subjects” (Whittaker et al., 2021). In this way, deepfakes could protect anonymous whistle-blowers from harassment and reprisal. However, it has not been established if this form of anonymization improves over existing practices of anonymization besides subjectively making content more “expressive”.

2.7.9 Deepfakes as a “Double-Edged Sword”

The positive uses of deepfakes were seen by some papers (mostly legal papers) as balancing out the existing harms of the technology. Narratives around balancing the impacts of the technology described deepfakes as a two-sided coin (Molina, 2021) or a double-edged sword (Nema, 2021) as a way of equivocating between the positives and negatives of the technology. This falsely implies parity between the positive and negative uses of the technology, which does not reflect the incidence rates of both (Ajder et al., 2019). While discussion of non-consensual pornographic related harms was prevalent in the dataset, the academic discussion does not reflect the dominance of pornographic deepfakes online. While many papers suggested correctly that pornographic deepfakes are the current dominant use of the technology, some papers played down their harms in comparison to the more hypothetical and apocalyptical political impacts. “But what happens when deepfakes stop being the fare of redditors with free time and a fixation on Game of Thrones actresses and become a weapon for state actors interested in destabilization of governments like the United States?” (Yadlin-Segal & Oppenheim, 2021).

2.8 Discussion

Our research builds on and extends the work of previous systematic and literature reviews across our three research questions. The scoping review methodology (Arksey & O'Malley, 2005) allows us to collate a broad range of research and theory to give us insight into the existing narratives around deepfake technology. Regarding existing work on deepfake definitions, our research finds similar themes to the synthesized definition established by Whittakker and colleagues (Whittaker et al., 2023). However, in place of Whittaker and colleagues' theme of "saying" (which they defined as the generation of "spoken words which are seemingly verbally communicated from the person depicted in the deepfake"), our critical analysis found the importance of identity and mimicry in how deepfakes are defined (instead of the generation of spoken words). This is more consistent with the popular established definitions we identified which describe the deepfakes as mimicking people "saying" or doing things which they never actually said or did (Chesney & Citron, 2019a; Vasist & Krishnan, 2022b) and even implicitly within the definition they provide themselves. This work also advances the work of existing categorizations of deepfakes by identifying problems with the use of existing terms like "shallowfake" and "cheapfake". These terms seemed to add unnecessary confusion as to what constitutes a deepfake. In contrast, we emphasize Millière's (Millière, 2022) philosophical paper which categorizes deepfake technology alongside other synthetic media (defining deepfakes as being audiovisual media modified through AI technology and not purely AI-generated images or text). This categorization is not only fundamental for researchers outside of philosophy to understand how deepfakes are different to other media, but it is also a useful guideline as to what constitutes a deepfakes. Further theoretical contributions of the work include identifying numerous unsubstantiated claims in the literature regarding the novelty, accessibility, believability, and harms/ benefits of deepfakes. We also discuss implications for how researchers approach deepfakes, outlining potentials for approaches which consider the role of identity and consent. The review also encourages the potential of a more holistic approach to deepfakes. Our review problematizes the commonly assumed dichotomy of positive and

negative deepfakes which is implicit in most review work on the topic (Albahar & Almalki, 2019; Dagar & Vishwakarma, 2022; Mahmud & Sharmin, 2021; Rana et al., 2022; Westerlund, 2019). Instead, our scoping review takes a critical approach to existing theoretical and review taxonomies on the harms and benefits of deepfakes and highlight how the harms and benefits are often two sides of the same coin, and the positive uses of deepfakes may have unintended harms.

2.8.1 Deepfakes are Inconsistently Defined

The deepfake literature is mired in the definitional issues which plague news media narratives around deepfake technology. The definitions for deepfake technology were inconsistent throughout the literature, with 16 papers not containing a definition for deepfakes at all. Traditional methods of generating synthetic media such as CGI and video editing were reframed as either “cheapfakes” (Paris & Donovan, 2019) or “shallowfakes” (Gregory, 2020). This review highlights how this confusion leads non-deepfakes being used to describe the potential harms of the technology. While this may have been important at the very advent of deepfake technology, there are now a multitude of examples of real cases of deepfake harms to cite. Our work encourages future work which seeks to reference other types of fake videos/ images to use terms that are already associated with traditional video and image editing such CGI, photoshopped, photogrammetry, and fake videos (in cases where it is clear that a video is fake, but the type of fakery is unclear).

2.8.2 The Importance of Mimicked Identity to Deepfakes

Our analysis highlights how the importance of mimicry in deepfake definition is understated in previous systematic and scoping reviews on deepfake definitions. Across disciplines, the primary harms of the technology draw from the technologies ability to defame through false visual representations of individuals. While most of the legal literature concerned with deepfakes has touched upon this damage, it is rarely recognized that most deepfakes inherently challenge representations of identity. In particular, pornographic deepfakes were seen

as causing defamation and irreparable harm to individuals self-image and a loss of control over their identity: “the misrepresentation of deepfake pornography impacts one’s ability to control their sexual identity” (Kugler & Pace, 2021). The importance of identities’ role in deepfake impacts has been understated by past research and it may be important to consider many harms related to deepfakes as drawing from the capacity of the technology to control or appropriate the identity of its target and to erase or disguise the identity of its source. Our review identifies the need for novel approaches to deepfakes which consider the role of manipulated identity as fundamental to the personal harms of deepfakes. Our review highlights a research gap in the study of identity and deepfakes which has broader implications for work focusing on challenging deepfake perpetration. Not only do deepfake communities trade in identities, but the specific identities of deepfake targets reflect their values. The defamatory harms in our analysis were viewed as currently and potentially targeting women as well as racial, gender, and sexual minorities and a means of silencing targeted activists. Existing work has highlighted the need to consider how deepfakes further social injustice and the role of intersectional approaches in responding to the damage of deepfakes (Karnouskos, 2020). Future research from the social and political sciences into the personal and privacy harms of deepfakes needs to privilege the concept of identity in their approach to deepfakes, focusing on how the technology serves as a vehicle to appropriate and disguise visual representations of marginalized groups. Similarly, policymakers and legal scholars who seek to restrict the harms of technology must recognize deepfakes as part of a larger pattern of technologies which facilitate the theft and appropriation of identity.

2.8.3 Deepfakes Are a New Way of Causing Old Problems

Our analysis questions what exactly academic researchers have been classifying as novel about deepfakes. We found that deepfake impacts are an evolution of the harms of previous technologies, and reflective of the online culture in which deepfakes emerged. The pornographic harms of deepfakes are an evolution of a previous practice using photoshop, the use of deepfakes in mis/ disinformation evokes the use of photo editing in fake news. This once

again ties into the accessibility of the technology, as while photoshop was not the first means of image manipulation, it made image-editing more accessible to the average person (Fineman, 2012). It is interesting to consider the prevalence of these fears in the political/ legal literature around deepfakes in contrast to lack of academic literature on traditional photo-editing technologies and their harms in misinformation (Weikmann & Lecheler, 2022). If the foundational harms of deepfake are exemplified by existing examples of non-deepfake content (e.g., slowed down videos being used to defame someone as drunk), researchers must consider whether it is worthy in studying deepfakes or simply the impacts or harms themselves. Unless deepfakes can be empirically proven to be significantly more convincing or uniquely harmful, studying the impact of all fake media on defamatory content will have broader impact, particularly in light of future advancements in generative AI which may have an impact in its own right on society as well as potentially increasing the accessibility of deepfake technology (Shoaib et al., 2023).

2.8.4 The Novel Epistemic Harms of Deepfakes

We call for further research, (particularly from social/ political sciences) questioning how the prevalence and accessibility of these technologies will affect online spaces, whether it be through the increase in AI-generated “spam” content online, or the increase in personalized defamatory deepfake content. In particular, the epistemic harms of deepfakes may be most affected by the prevalence and accessibility of deepfake videos. An increase in the accessibility of deepfake video production would lead to an increase in the prevalence of deepfake videos online and thus lead to increased damage to the value of real media in our information environment. In this way, we propose that deepfakes should be considered a novel method for “epistemic pollution”, false information which can worsen the quality of our information environment (Ryan, 2018). If the number of fake videos increases online due to deepfakes and other kinds of false videos, our current online information environments, online forums, social media etc. could become significantly less usable as users seek to navigate increasing amounts of manipulated information. Of particular concern to legal scholars is the way deepfake footage

damages the epistemic value of video evidence. Particularly how deepfakes allow individuals to falsify or incorrectly challenge CCTV evidence (Apolo & Michael, 2024). Mitigating the amount and damage of generative AI misinformation online is an essential goal for policymakers, media companies and researchers in political science and across disciplines.

2.8.5 Deepfake Accessibility and Believability: Cause for Concern, Strategies for Solution and Need for Empirical Evidence

Our scoping review highlights how the accessibility and believability of deepfakes is central to both the impacts and novelty of the technology. Deepfake technology, while more complex than other forms of media editing, is often hidden behind user friendly tools (Karnouskos, 2020). Existing work has studied how the accessibility of deepfakes is a byproduct of the open-source movement. The lack of oversight of open-source technology fuels negative uses of the technology (Widder et al., 2022). We propose that the extent to which deepfakes are both believable and accessible are constantly in flux, depending on people's awareness of the technology and the prevalence, not only of deepfake technology but the expert/ hobbyist communities dedicated to their production. Our work calls for increasing empirical and theoretical work which seeks to understand the role of accessibility and believability in the potential uses of deepfakes. The multi-disciplinary nature of this review highlights how the design of solutions to deepfakes and future research into the technology across disciplines need to consider the function of accessibility and believability. For example, legal and political efforts to mitigate the harms of deepfakes should consider controlling the prevalence and accessibility of particularly harmful versions of these technologies and challenging the content available in consumer deepfake applications. Similarly, the perceived believability is often used to argue for how dangerous political deepfakes can be, but there is little evidence of this. Research on the believability of deepfakes should interrogate whether hyper-realistic deepfakes effect people's attitudes more than lower quality deepfakes and if, without the obvious novelty of the technology, they provide anything useful for advertisers or other "positive" users. Similarly, research on the educational uses of deepfakes should question if "believable"

educational historical deepfakes are more exciting to students than the use of costumes and actors. Research should question whether educational content must be hyper-realistic and believable. Other supposedly positive uses of deepfakes such as multi-lingual media do pose some advantages over using traditional dubbing, notably that it may help people who rely on lip-reading, but these positives need to be tested empirically by future research.

2.8.6 Academic Narratives Around the Positive and Negative Uses of Deepfakes

Our findings build upon early legal scholarship done by Chesney and Citron (2019) which categorized positive and negative deepfake applications. While their taxonomy is still foundational to understanding the impact of deepfakes, our analysis takes from multiple disciplines over the first five years of the technology and highlights the complexities and problems behind categorizing deepfakes as purely positive and negative and bound within either personal/ organizational or societal domains. As shown in our data, there were more references to real world incidences of deepfake harms than of their benefits. Often, the discussion of the supposedly positive uses of deepfakes in disciplines such as entertainment (e.g., resurrecting dead actors or recasting movies) were problematized by legal criticisms that it violates intellectual property, takes away roles from actors and runs into the same issues with consent that were more dominant in discussion around the pornographic use of the technology. Existing philosophical work on deepfake technology shows how lack of consent often underlies the harm of deepfake videos (de Ruiter, 2021). Our findings encourage a more critical and consent-based approach to deepfakes, which is particularly relevant to researchers studying the supposedly positive uses of deepfakes. These researchers need to consider how even positive uses of deepfakes may undermine the consent of individuals involved and may damage people's trust in video media. They should also question if the novelty and strangeness of the content draws attention away from the positive value of deepfake content.

2.8.7 Reconsidering the Categorization of Deepfakes Uses

A significant contribution of the review to academic theory around deepfakes is the need to apply a more holistic lens to deepfakes (i.e., to take an approach to deepfakes which more accurately conveys the complexity behind positive and negative uses). Much of the literature, particularly law and politics literature viewed defamatory pornographic deepfakes as being one of the main harms of deepfakes and focused on the technology as having personal (individual) harms. The distinction between these harms is fundamental to much of the established literature on deepfakes including Chesney and Citron's work (2019). Yet, what also emerged in the analysis is that the defamatory harms of deepfakes are in themselves political (and thus societal). They often target specific groups such as politicians, celebrities and activists and are used as part of campaigns to silence and harass their mostly female targets. It should also be noted that while in general the political harms generally exist on a social level and the pornographic harms exist on an individual level, both types of harms have the capacity to exist both socially and individually. The social harms of deepfakes will have effects on individuals and the reputational effects of deepfakes will have broad effects on society in general. Building on this, one should consider that every harm of deepfake has both a personal, inter-personal and a societal element and most positive uses of deepfake have a corresponding (and often significantly more impactful) harm. While taking care to avoid overly fearful conjecture (Vasist & Krishnan, 2022a), researchers should consider the uses of deepfakes as being mostly negative (or morally suspect as proposed by Rini (Rini, 2020). Our work has implications for disciplines participating in the "beneficial" uses of the technology, where they should consider any associated or possible harms across individual and societal domains as well as within and outside of their own discipline.

2.8.8 Further Implications for a Holistic Approach to Deepfakes

Our work calls for deepfake literature to focus less on potential future harms and to take a more holistic view on deepfakes. We provide two contributions of a holistic approach to future research to interrogate the gaps in the literature exposed by this review. First is the necessity of

more empirical research on the present impacts of deepfakes. This should specifically focus on the lived experiences of the targets of deepfakes. This would provide insight into how deepfakes harm and whether these harms are unique or an evolution of existing harms. Secondly, the fixation on the future of deepfake technology provokes an alarmism that needs to be studied in more detail. We have found similar evidence of alarmism in academic literature to alarmism in the media narratives around deepfakes analysed by existing research (Wahl-Jorgensen & Carlson, 2021). The harms of deepfakes to trust in video evidence often rely on the ability of deepfake awareness (Chesney & Citron, 2019a) to undermine faith in real video evidence (Chesney & Citron, 2019b). Existing review work has previously critiqued the perceived fears and stigma around the use of deepfakes in academic literature (Vasist & Krishnan, 2022a). Perhaps the fearful speculations around deepfake technology identified in our review of academic literature is a further indicator of the harms of broader social narratives around deepfakes. Academics across disciplines need to consider how awareness existence of this potentially harmful technology evokes overt scepticism and fearfulness of the future, detracting from the ability to focus on the present values and harms of the technology.

2.9 Limitations

Deepfakes are an emerging form of technology and any attempt to survey current attitudes and narratives around the technology must be disclaimed with the fact that this field is constantly evolving. As such, our dataset provides a broad approximation of narratives from within the first four years of academic research across stakeholder domains. We highlight the necessity of understanding divergent viewpoints across disciplines on deepfakes. Though future work making empirical claims as to differences between disciplines is necessary, it will only be valid when more research is published as outside from law and IT, as there is a comparative dearth of deepfake research outside these disciplines. While our findings benefitted from the breadth of our review, it impacted how we carried out our qualitative analysis of our dataset. The dataset contained over 180000 words, and while the resulting document provided a wealth of multidisciplinary discourse around deepfake technology, its size hindered how reflexive the

qualitative analysis could be. Qualitative analysis is influenced by the quality and quantity of the data it studies (Morse, 2000). As a result, our thematic analysis was more focused on identifying broader ideas and narratives within the corpus as opposed to drawing out smaller themes within each use case. Furthermore, while a second reviewer was involved in deciding on which papers were included/ excluded and read through the full text of each, producing inter-rater reliability would not have been practical given the size of data required to study and in light of broader theoretical concerns as to the legitimacy of inter-rater reliability as a construct (Braun & Clarke, 2021b). Similarly, we could not include every single paper on deepfakes in our analysis - while our inclusion criteria were broad, it also had to be somewhat economical to avoid including papers which only tangentially mentioned one of the themes leading to the dataset becoming too large to analyse. This practicality also led us to focus distinctly on journal articles as many of the conference proceeding focused specifically on deepfake generation and production which were outside this reviews scope. However, we recognize the importance of conference proceedings to academic work especially in AI related research.

2.10 Conclusion

Our scoping review challenges current academic narratives around the definitions, novelty and impacts of deepfakes using a large multi-disciplinary dataset. There are three contributions of the review to academic theory around deepfakes and to current and future deepfake research. 1. Deepfakes are inconsistently defined but there is some consensus as to what constitutes a deepfake. 2. Deepfakes are an evolution of previous technologies, it is better to understand how deepfakes uniquely (often by being more accessible/ believable) augment and challenge existing practices and solutions. 3. There is a conceptual issue with the assumed dichotomous positives and negatives impacts of deepfakes across distinct domains. We would like to elaborate further on each of these points. Firstly, our research has highlighted the evolving definitional issues around deepfakes. We highlight how they have caused confusion in discussions of the novelty and impacts of deepfakes. Yet also, on a more positive note, we have shown that academics are arriving at some commonality between the use of deepfakes with

more recent papers more likely to include a definition or cite notable previously established definitions. The way deepfakes have been defined has evolved during the early years of the technology to reflect academia's updated understanding of the phenomenon and the requirements of different disciplines. Our work expands and further refines previous review work on the definitions of deepfakes. Particularly we update existing theory on deepfake definitions (Whittaker et al., 2023) by highlighting that mimicked identity is integral to deepfakes. This review questions the purpose of research that focuses specifically on novel technologies as having unique hypothetical harms. Research in this field should explore the harms in themselves and understanding what novel technologies change and do not change about existing solutions. Furthermore, the lenses of believability and accessibility must be considered when exploring the impact of deepfake technology. Deepfakes can be seen as one of many evolutions to synthetic media in the last decade which have furthered both the accessibility and believability of fake information online. We call for more empirical work on the existing harms of deepfakes which engages and tests the factors of accessibility and believability of the technology. We further problematize the current attitudes towards deepfakes as having a clearly established positive and negative across distinct personal, organizational, and societal levels. There is an over-focus on dystopian and utopian potentials of deepfakes in comparison with the ethical issues with the technology itself. Clearly, broader critical issues around consent, copyright, and identity underly even the positive uses of deepfake technologies. These are vital factors to consider in the study of deepfakes. We also highlight the overlap between the personal, organizational, and societal levels of the technology. We critique existing models of deepfakes uses which divide uses into personal and societal harms/ benefits.

CHAPTER 3: Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine

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.

This paper was published in Plos One on the 25th of October 2023.

Study Motivation

This is the first of two papers which shifted focus from the potential harms of deepfakes to what is happening right now. The impetus of this paper was reading of the news in March 2022 about a deepfake video of the Ukrainian president Volodymyr Zelenskyy surrendering. The deepfake had been broadcast on a real Ukrainian news website that had been hacked and aimed to convince Ukrainian soldiers to surrender. The deepfake of Zelenskyy made it extremely clear that while the literature was focusing on potential harms of deepfakes; deepfake videos were actively being used in wartime dis/misinformation.

There were a number of qualitative case studies on deepfakes before this paper, but they almost exclusively focused on entertainment deepfakes (Bode, 2021). The aim of this work was to use social media posts as a case study for the effects of deepfake misinformation.

Author Contributions

This study was conceived by JT, DC and the supervisory team. The analysis was performed by JT under the guidance of CL. JT wrote up the paper with guidance from the supervisory team. All authors approved the final version of the manuscript

3.1 Abstract

Deepfakes are a form of multi-modal media generated using deep-learning technology. Many academics have expressed fears that deepfakes present a severe threat to the veracity of news and political communication, and an epistemic crisis for video evidence. These commentaries have often been hypothetical, with few real-world cases of deepfake's political and epistemological harm. The Russo-Ukrainian war presents the first real-life example of deepfakes being used in warfare, with a number of incidents involving deepfakes of Russian and Ukrainian government officials being used for misinformation and entertainment. This study uses a thematic analysis on tweets relating to deepfakes and the Russo-Ukrainian war to explore how people react to deepfake content online, and to uncover evidence of previously theorised harms of deepfakes on trust. We extracted 4869 relevant tweets using the Twitter API over the first seven months of 2022. We found that much of the misinformation in our dataset came from labelling real media as deepfakes. Novel findings about deepfake scepticism emerged, including a connection between deepfakes and conspiratorial beliefs that world leaders were dead and/or replaced by deepfakes. This research has numerous implications for future research, societal media platforms, news media and governments. The lack of deepfake literacy in our dataset led to significant misunderstandings of what constitutes a deepfake, showing the need to encourage literacy in these new forms of media. However, our evidence demonstrates that efforts to raise awareness around deepfakes may undermine trust in legitimate videos. Consequentially, news media and governmental agencies need to weigh the benefits of educational deepfakes and pre-bunking against the risks of undermining truth. Similarly, news companies and media should be careful in how they label suspected deepfakes in case they cause suspicion for real media.

3.2 Introduction

Synthetic media are a type of audio-visual media which has been partly or fully generated/modified by technology (Millière, 2022). Deepfakes are a new form of synthetic media which interpolates artificially generated media into an existing video, often with intent to imitate or mimic an individual. Researchers and commentators have argued that deepfakes have the potential to undermine truth, to spread misinformation and to undermine our trust in information, media and democracy (Chesney & Citron, 2019a). The increasing prevalence of fake videos could undermine what we know to be true (Fallis, 2021). Specifically, academic researchers believe that deepfakes could create a situation where fake videos are believed to be real and conversely, where real videos are denounced as fake. Fears of deepfakes being used to spread disinformation have been realised during the Russo-Ukrainian war. We have seen fake videos of both Russian President Vladimir Putin and Ukrainian President Volodymyr Zelenskyy, as well as many satirical and entertainment uses of deepfakes around the crisis. Our paper seeks to provide empirical evidence for the hypothesised and forecasted harms of deepfakes to truth and knowledge. The aim of this paper is to understand the nature of deepfake discourse on social media in the context of the Russo-Ukrainian war.

In this paper, we analyse Twitter discourses around deepfakes in relation to the Russo-Ukrainian war by carrying out a thematic analysis on relevant tweets during the first months of the 2022 invasion. Our study is the first empirical analysis carried out on the use of deepfakes in wartime misinformation and propaganda. As deepfake technology becomes increasingly accessible, it is important to understand how such threats emerge over social media. Understanding the current threats of deepfakes will have implications in how social media companies and academic researchers deal with harmful deepfakes online. Understanding how the threats of deepfakes emerge online is a significant step in learning how to mitigate their harms. The current paper also has numerous implications for academic research on deepfakes. Our research explores the epistemic harms of deepfakes in practice, as opposed to the theoretical discussions of the concept in academia (Kerner & Risse, 2021). We also provide a

non-exhaustive timeline of the use of deepfakes and other synthetic media in the Russo-Ukrainian war. The provided timeline is important as both a record of the uses and the impact of deepfakes during the Russian invasion of Ukraine and as a means to gauge the type and quality of synthetic media content created during the conflict.

3.2.1 Background: Deepfakes and misinformation

Deepfakes are an emerging form of synthetic media with potentially harmful implications for truth and democracy (Diakopoulos & Johnson, 2021). While there is not a consistent definition of the technology, deepfakes are generally understood as audio-visual media which have been manipulated using artificial intelligence technologies (Celli, 2020). Deepfakes most often consist of videos where a fake "face" constructed by artificial intelligence has been merged with an authentic video. The result of this is a realistic, mimetic recording of someone doing and saying things which they never did or said (Sylvester, 2020). Deepfakes were brought to media attention in early 2018 when a reddit community (r/deepfakes), used artificial intelligence to swap celebrities faces into pornographic videos (Winter & Salter, 2020). The potential harms of deepfakes include the erosion of trust in institutions, democracy and journalism (Wilkerson, 2021). Some journalists have even hypothesised that deepfake misinformation could lead to a world leader declaring nuclear war over deepfake videos (Langa, 2021).

Concerns about deepfakes have centred primarily on their capacity to deceive. Misinformation and disinformation refer to false information utilised to mislead people. The difference between the two concepts is that disinformation is done purposely, while misinformation is not (Søe, 2021). Deepfakes have the capacity to be used for both misinformation and disinformation by depicting events that never happened (Paris & Donovan, 2019). While deepfakes are usually created with deceitful aims, they are often shared online by people who believe them to be unedited videos. For example, in a study carried out by Ahmed (S. Ahmed, 2021b) in the USA and Singapore, it was found that over 30% of their sample had inadvertently shared a deepfake. Quantitative research around deepfakes has broadly focused on

individuals' ability to detect deepfakes (Korshunov & Marcel, 2021), and factors which lead to accidental sharing of deepfakes (S. Ahmed, 2021a). In contrast, qualitative research has focused on the supposedly positive elements of deepfakes on social media, focusing on entertainment videos on YouTube (Bode, 2021; Y. Lee et al., 2021). There has also been some quantitative research into the effects of deepfakes on trust. Vaccari and Chadwick (Vaccari & Chadwick, 2020) found that deceptive deepfake videos were significantly more likely to elicit uncertainty compared with the same video containing an educational disclaimer. However, they did not compare this to a control group who had not been exposed to a deepfake, which poses a broader theoretical question as to whether or not educational deepfakes may be harmful to truth in themselves and if the very existence of deepfake technology may pose a threat to truth.

The broader theoretical concern as to how the very existence of deepfakes may cultivate misinformation is exemplified by the liar's dividend. As described by Chesney and Citron (Chesney & Citron, 2019a), the liar's dividend refers to a situation in which people may avoid accountability for real audio and video evidence as a result of public scepticism. As the public becomes more and more aware of novel video manipulation technologies, real video evidence will be increasingly denounced as fake. For example, in a hypothetical situation, where a politician is faced with compromising video evidence, they could deny a factual incriminating video of themselves as a deepfake. The prevalence of the liar's dividend also increases proportionately with the public's awareness of deepfakes (Chesney & Citron, 2019a). Encouraging scepticism towards visual media may unintentionally have the adverse effect of making it more likely that people will flatly deny real video evidence with which that they are presented (Chesney & Citron, 2019b). While pre-emptive misinformation interventions, known as pre-bunking, can be used to combat other forms of misinformation, pre-bunking deepfakes may only increase their epistemic harms (Roozenbeek et al., 2020). Quantitative research into the efficacy of deepfake misinformation interventions has found that while interventions improved people's ability to identify deepfakes it also reduced participants' ability to correctly identify real media by 9% (Tahir et al., 2021). While the capacity of deepfakes to undermine

people's ability to distinguish truth has thus been shown in experimental contexts, there is a lack of research qualitatively exploring how this epistemic distrust may manifest online.

The liar's dividend demonstrates deepfakes' potential to encourage deniability of real video evidence and harm the work of journalists, governments, and the legal system (Whittaker et al., 2020). The doubt that deepfakes cast on the truthfulness of video evidence has been described as the biggest epistemic threat of the 21st century (Carlson, 2021). Interestingly, despite these drastic projections made by commentators when deepfakes first emerged, there have not been many high-profile or large-scale uses of deepfakes for misinformation purposes. This has led some academics to consider the epistemic harms of deepfake overblown and to suggest that deepfakes may conversely have a positive effect on trust due to an increased scepticism towards dubious online sources and a move to more verifiable content (Etienne, 2021). In practice, most existing uses of the technology revolve around harassment and bullying, particularly image-based abuse of women (Winter & Salter, 2020). Any uses of deepfakes in the political sphere have largely been benign, such as a Belgian political message spoken by a deepfake of Donald Trump, which was used to highlight climate change (Ullrich, 2022). The use of deepfakes on social media during the Russian invasion of Ukraine indicates that this may be changing.

3.2.2 Deepfakes, disputed deepfakes and other synthetic media during the Russian invasion of Ukraine

The Russo-Ukrainian war began in 2014 with the Russian invasion and annexation of Crimea (Kuzio, 2018). In the years after this, Ukraine saw increased acts of aggression initially by Russian-proxy forces and Russian backed separatist militias in the Donbas region of Ukraine (Marples, 2021). Later responses to Ukraine's attempts at a counterattack were foiled by covert troops and weapons from the Russian armed forces (Katchanovski, 2016). In the later months of 2021, Russian forces were amassed at the two countries' borders under the pretence of a training exercise. At this time, Russian governmental figures made incorrect and disingenuous claims of widespread Russophobia and Nazism in Ukraine (Kuzio, 2023). This was further fuelled by a

controversial address made by the Russian president Vladimir Putin which incorporated similar "conspiracy theories and lies" to attempt to undermine the history and existence of Ukraine as a Sovereign state in its own right (Brunk & Hakimi, 2022). On the 24th of February 2022, the armed forces of Russia launched a full-scale military invasion of the country of Ukraine (Ebim, 2022). These actions, along with the brutality of the Russian invaders, drew widespread condemnation from many governments and organisations both within Europe and worldwide (Asadzade & Izadi, 2022).

Public outrage and interactions with the Russo-Ukrainian war have been mediated by social media platforms and technology. Cell-phone journalistic practices mean that first-hand accounts of the atrocities of the conflict are readily available online (Liu et al., 2022). Internet activism has meant that individual users of the internet have been able to co-ordinate distributed denial-of-service (DDoS) attacks on Russian websites (Husák et al., 2022). While modern technologies have facilitated the spreading of first-hand accounts of the brutality of the war, they have also been used to spread misinformation and propaganda (Henkhaus, 2022). The Russo-Ukrainian war has seen for the first time, the use of deepfake technology in wartime propaganda and misinformation. Figure 3.1 provides an overview of deepfakes (as well as similar synthetic media that have been misidentified as deepfakes) used in the Russo-Ukrainian war during the first four months of the invasion. In this section we will discuss the key events relating to deepfakes, contested deepfakes and other synthetic media in the war in more detail.

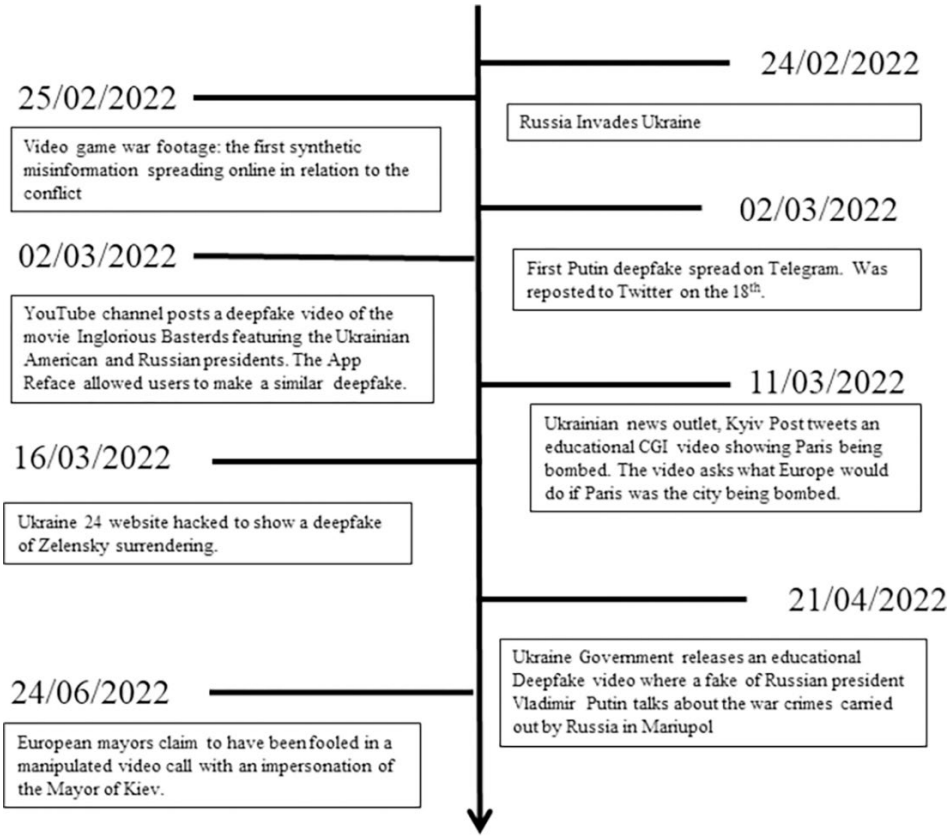


Figure 3.1: Timeline of the use of deepfakes/synthetic media in the war.

3.2.3 Synthetic video footage of a Ukrainian fighter jet.

Misinformation in the early days of the war was not characterised by deepfakes. Instead, most misinformation consisted of the misuse of videos from the earlier conflicts in Ukraine. Notably, the second day of the war saw the spread of video game footage as evidence of the urban myth fighter pilot “The Ghost of Kyiv”. This video footage was taken from the video game Digital Combat Simulator. It was spread around social media during the first few days of the invasion purporting to show a Ukrainian fighter jet (Coleman, 2022). The Ukraine Department of Defence even tweeted the viral footage. This serves as the first known example of synthetic media concerning the Russian invasion of Ukraine.

3.2.4 Deepfake of Putin announcing peace with Ukraine.

In early March a deepfake of Russian president Vladimir Putin emerged, showing the Russian president announcing peace with Ukraine. The deepfake was first published online in

the first week of March on the reddit r/sfwdeepfakes and r/ukraine communities (Reddit (u/reizuki), 2022). It was posted with an acknowledgment that it was fake and the user who submitted this video claimed to have found it on the social media site Telegram and added their own subtitles. The deepfake was then published on Twitter on the 18th of March. This is the version that was reported on by news agencies. The version posted on Twitter did not contain subtitles, suggesting it may have come from Telegram. This deepfake was unique out of the major examples of the technology in the conflict as it has been suggested that the audio was also generated using AI (Reuters, 2022).

3.2.5 Deepfake of Zelenskyy surrendering to Russia.

In the months leading up to the invasion, the Ukrainian government warned of the potential for Russia to employ a deepfake of Zelenskyy surrendering to the Russian Government (Pearson & Zinets, 2022). On the 14th of March 2022, a Ukrainian news website was hacked to display a deepfaked message of Zelenskyy surrendering. Concurrently, the News Ticker of Ukraine24's television channel was changed to read a similar message. Despite media coverage of this event accusing the government of Russia of orchestrating the hack, they have not claimed responsibility to date, nor has any significant evidence been found indicating their involvement. This event is significant as it is one of the first cases of a deepfaked politician being intentionally used to spread disinformation. The quality of the deepfake was poor and quickly debunked by Zelenskyy himself. However, this incident highlights the potential harms of deepfakes. The usual indicator of truthfulness and trustworthiness of online information, the source of the video, was undermined by the hack. If the video had been more realistic and more widely believed, it may have had a more harmful impact.

3.2.6 Educational deepfakes from the Ukrainian government and news media outlets.

On the 21st of April, the government of Ukraine released an educational deepfake video in partnership with Reface, a Ukrainian AI company. Titled "Putin telling truth, huh? Well, we

tried to imagine what he'd say if he did", the footage showed a deepfake Putin walking around the Ukrainian city of Mariupol and describing the war crimes carried out by the Russian forces (Ukraine / Україна [@Ukraine], 2022). A similar non-deepfaked video was posted the month before consisting of a CGI video of Paris being bombed, asking how European citizens would feel if their cities were being attacked. These two examples are relevant as they show the educational and political communication uses of synthetic media and deepfake technology.

The Reface app also allowed users to swap their faces into Zelenskyy's speeches, or onto images of Ukraine soldiers (Reface, 2022). The app encouraged users to faceswap themselves and Putin into a scene from the Quentin Tarantino movie "Inglorious Basterds" in which a Nazi is violently beaten to death by an American soldier. Versions of these videos including Zelenskyy as the American soldier also went viral online (Reddit (u/HauntingFly), 2022). This is one example of the numerous online videos recasting the two presidents into Hollywood movies and other satirical or humorous videos. Most of these videos involved using AI to insert the Russian and Ukrainian presidents into comedic movie scenes or dancing/lip-synching to music. Oftentimes, political figures in the war would be recast as famous movie villains or heroes. These videos were clearly satirical. While intended to influence, they were obvious parodies and probably not intended to deceive.

3.2.7 European mayors receive possible deepfake video call from the mayor of Kyiv.

On the 24th of June, three European mayors came forward as victims of a deepfake video call. The mayors of Berlin, Madrid and Vienna had believed they were talking to the mayor of Kyiv, Vitali Klitschko (Oltermann, 2022). After a certain point, odd and seemingly nonsensical questions alerted the politicians that this was not the real Vitali Klitschko. This potential deepfake is both unverified and contested, with some suggesting that the hack does not contain deepfaked footage and the video consisted of reused video clips (Laufer, n.d.). Two Russian comedians, Lexus and Vovan have claimed responsibility for the video, but have not stated how the video was manipulated (Hoppenstedt, 2022). This case is quite similar to an

incident in 2020, in which many mayors decried the use of a supposedly deepfaked video call with Leonid Volkov. It was in fact a prank call also orchestrated by Lexus and Vovan. The supposed deepfake was no more sophisticated than make-up and cleverly chosen, obfuscating camera angles (Vincent, 2021). The pair of comedians, who are notoriously pro-Putin, have attracted attention with similar actions during the Russo-Ukrainian war, imitating Zelenskyy during an audience with noted author J.K.Rowling (Peiser, n.d.).

While it is possible to deepfake a face in a video-call with software such as DeepfaceLive (Gerstner & Farid, 2022), the software to accomplish this is in its infancy and while capabilities have advanced significantly in the last two years, it still often requires significantly more work to get to a lower level of quality than traditional deepfakes. It is more likely that the video call used either manually (non-deepfaked) stitched together clips from a previous interview or lip-syncing technology (which is considered by some as a type of deepfake). Evidence for this is the fake mayor's insistence on using Russian despite their fluency in German and how the video frames seem to come from an earlier interview from the mayor. Nonetheless this is an example of the capabilities of fake videos in warfare and the current vigilance towards potential deepfakes. It also highlights how news media and governmental agencies can label traditionally manipulated videos as deepfakes, contributing to a fear of the capabilities and prevalence of the technology.

3.2.8 The current study

The Russo-Ukrainian war has highlighted the impact of novel misinformation technologies on online media spaces. We have seen the realisation of many fears around deepfake technologies, in which deepfaked presidents have been used to affect public opinion regarding an active war. However, it is unclear how this new form of wartime misinformation was perceived online at the time, and what harms were caused. It is vital to assess this, so governments, social media organisations, and other stakeholders are able to understand the harms deepfakes may cause to social media discourse in present and future conflicts. We used qualitative analysis to explore Twitter users' comments on the topic of deepfakes during the

Russo-Ukrainian war. This data provides a unique opportunity to assess the potential harms of deepfakes in a real-world situation, as well as assessing evidence for some of the fears around deepfakes, such as the liar's dividend and the harms to epistemic truth.

Our research questions are as follows:

1. How have people responded on Twitter to deepfake content during the Russian invasion of Ukraine?
2. Does Twitter discourse around deepfakes provide real-world qualitative evidence of the epistemic harms of deepfake technology which have been previously theorised and experimentally studied?

3.3 Methods

3.3.1 Data collection

The decision to study Twitter was made because the majority of the most prominent cases of deepfakes (and similar synthetic media) identified in our timeline involved the social media platform. We used the Twitter developer API to capture tweets relating to deepfake content from the 1st of January 2022 to the 1st of August. The API was accessed using a python program based on one from the Twitter GitHub page (GitHub, n.d.). The search term used was "(deepfake Zelenskyy OR deepfake Ukraine OR deepfake Putin)–is:retweet". We excluded retweets in our search as they would have been duplicates of an original tweet. We extracted the textual content, the date, and the tweet ID of each tweet into a CSV file. Our final dataset included tweets in English, German, and French. This was done to consider a broader range of western European perspectives rather than a purely English-language dataset, particularly considering the salient German language discourse around the Klitschko video call. Furthermore, these were languages that two members of the research team had experience with, which avoided the pitfalls of relying on purely machine-based translation. Russian and

Ukrainian language discourse were excluded out of respect to those involved in the war, to reduce the chance of researcher harm resulting from exposure to sensitive personal and combat footage being included in the dataset, and because we are primarily concerned with the discourse around the war and not the experiences of the war itself. This decision was also informed by challenges with searching the Twitter API for multiple search terms in different languages, the increased likelihood of mistranslation due to the team's lack of fluency in either Ukrainian or Russian, and issues with sample bias due to the comparative unpopularity of Twitter in Eastern Europe. Similarly, we avoid making claims as to whether tweets were posted by Ukrainian or Russian Twitter users, considering the existence of Russian troll farms which assume false identities and automatically generated accounts to spread misinformation (Lange-Ionatamishvili et al., 2015). Any claims made in online research regarding the participants or victims of this war faces issues around the prevalence of these appropriated identities. Research on the actual experiences of Ukrainian nationals with deepfake disinformation during the invasion requires careful ethical consideration and knowledge of identity which cannot be securely gained by this type of broad qualitative research of an online dataset.

As with all qualitative research, is also important to consider the biases and positionality of the research team. It is impossible to make claims to the demographics of the dataset, except that both it and the research team broadly consisted of privileged voices from western Europe and north America. Our research questions do not focus on the specific experiences of people involved in the war with deepfakes, as the specific demographics and anonymity of Twitter reduce any claims to the national identities of any of the participants. While we have justified our decisions to work on this demographic, it does highlight possible alternative avenues for the analysis of deepfake content on social media. Any qualitative research on the experiences of the Ukrainian people with deepfake disinformation (such as the Zelenskyy deepfakes) would require a different mode of analysis, and a different method of data collection which would avoid the risk of studying false identities online.

In total, 4869 tweets were extracted between the 1st of January 2022 and the date of data extraction, the 28th of August 2022. The extracted dataset was saturated with articles from news media outlets. 1984 tweets were excluded as they consisted of links to news articles with news headlines and quotes and no other relevant reactions and commentary. Similarly, a further 212 only consisted of links to images and videos with no significant textual content. 590 of the posts were excluded for being duplicates, usually consisting of spammed posts. While we aimed to reduce the number of fake posts in our datasets, we limited ourselves to excluding tweets which were obviously spam as further measures of bot-detection would run risks of excluding real data. 691 tweets were excluded because they were in languages not included in our inclusion criteria such as Spanish. Overall, 1392 tweets were included for analysis. 84% of these were English, 12% were German and 4% were French. This is generally unsurprising considering the use of English language search terms. The high prevalence of German tweets in the dataset is most likely a result of the fake Vitali Klitschko video call which affected two German-speaking cities. During the analysis it was found that 161 of these did not contain enough lexical density to be used in our analysis because they for example only consisted of self-promotion or spreading news without commentary. In total, our data collection provided 1231 tweets to be analysed. Due to ethical considerations, any links to Twitter pages or attachments were replaced with summaries after the analysis.

3.3.2 Prevalence of deepfake discourse during the start of the war

Deepfake related content steadily grew during the week leading up to the war and the first few weeks of the war (Figure 3.2). This wave of deepfake discourse peaked on the second of February, with 441 tweets being extracted on this day. This largely is the result of spam content and news coverage of the US and Ukraine's warnings of deepfake content. The largest number of tweets occurred between the 16th to the 18th of March, correlating broadly with the release of the fake Zelenskyy surrender video and news coverage of the supposed Putin deepfake. A smaller peak occurred during the 4th to 26th of April coinciding with the supposed deepfake video call between Klitschko and the European mayors.

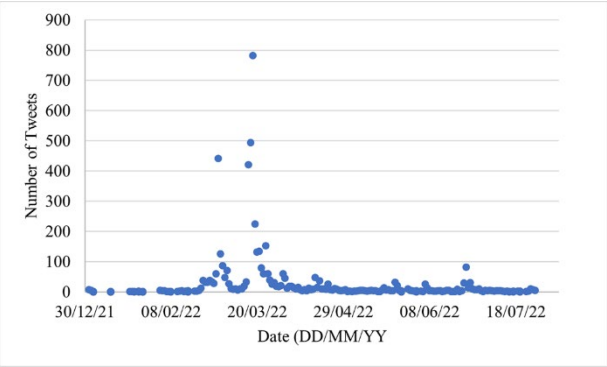


Figure 3.2: Prevalence of tweets over time in the dataset.

3.3.3 Ethics

Ethical issues with researching Twitter have been well-established in prior literature. Our ethical processes were designed to mitigate the harms associated with online research, particularly the inability to gain informed consent. Specifically, we followed relevant academic ethical guidelines (W. Ahmed et al., 2017; Fiesler & Proferes, 2018; M. L. Williams et al., 2017) as well as the relevant Twitter developer and privacy policies. We used anonymisation and non-traceability as a means of mitigating this risk. Despite existing as a public online platform that is open to inquiry by researchers, Twitter users still deserve care and anonymity when being studied online, especially in ethnographical research where informed consent cannot be easily gathered. Similarly, many Twitter users are not aware that their content is not only open to use by researchers and businesses, but actively encouraged to be used in this context, and often used in a way which is traceable back to their account (Fiesler & Proferes, 2018). Because of this, it is important to make sure users are not publicly identifiable from research papers. To ensure users' anonymity and non-traceability, any examples of tweets in this paper have been replaced with fully fictionalised tweets which represent broadly the original data. Ethical approval to carry out this research was granted by the Social Research Ethics Committee of University College Cork.

3.3.4 Data analysis

We used and inductive, thematic analysis (Braun & Clarke, 2012) to understand and illustrate patterns in our data. We judged thematic analysis as the most appropriate way to

address the data set, since the approach is flexible, allowing for many different levels of interpretation. Moreover, there is no existing well-defined theory of how people understand and respond to deepfakes that could guide a more deductive or theory-driven approach. The thematic analysis broadly followed the six steps outlined by Braun and Clarke (Braun & Clarke, 2006). Initially, while applying the inclusion and exclusion criteria, we familiarised ourselves with the data, drafting up a series of rough codes in a text document. We then used NVIVO software to descriptively open code the data. Once open coding was complete, we carried out a number of iterative rounds where we sought to meaningfully define observed patterns in the codes. Specifically, we first arranged the codes into small groupings and sub-themes within NVIVO and then worked on developing larger overarching themes. The majority of this work was done by the primary researcher (White-Irish, Male), with regular consultation with the broader research team throughout the process to iterate on codes and themes. We chose to have one primary researcher do this work as the research often required making judgements to the truth or falsity of deepfake claims and the primary researcher was most aware of both the news media within the dataset (for exclusion of news headlines) and the specific cases of deepfakes used in the war. This was an inductive and iterative process. For example, while initially we approached formulating the codes as a typology of responses to deepfake misinformation, but we found that the range of novel themes and usages of deepfake content outside of misinformation necessitated a broader and more comprehensive set of themes.

After generating and refining a set of three themes, we carried out further analysis, borrowing methodologies from discourse analysis (Potter & Wetherell, 2004), which allowed us to critically appraise the social constructions behind the textual data (Braun & Clarke, 2021a). We considered how the tweets reflected and drew meaning from the real-world incidences of deepfakes and we incorporated discursive methodologies such as the use of framing and interpretative repertoires. Analysing power relations, positioning, and comparing the discourse in individuals' tweets such as in critical thematic analysis (Lawless & Chen, 2019) was impossible due to the anonymity of our dataset, the character limit of tweets and the lack of

conversational context. While this research did not involve a content analysis (Mayring, 2004), we include the number of references to give the reader an idea of the prevalence of themes in our interpretation of the dataset.

During the writing up process, to ensure the validity of the final quotations, the analysis was written up with the actual quotes before those quotes were then replaced with paraphrased versions. We chose to anonymise the quotes during at the end of the write-up process to ensure that the fictionalisation not affect the analysis. The full research team refined these quotes to ensure they adequately represented the original quotations which included utilising similar typos, using synonyms and rewording the order of sentences. This was done to accurately reflect the original tweets and how they draw from the "new language" of online discourse (Varnhagen et al., 2010). The methodology was reported in this paper according to the Consolidated criteria for reporting qualitative research (COREQ) checklist (Tong et al., 2007), though this had to be adapted due to the unique participant and methodologies of online qualitative studies in contrast to research carried out using in-person interviews.

3.4 Results

Our analysis produced three main themes of online content related to deepfakes and the Ukraine crisis; 1) deepfakes and misinformation, 2) deepfake fuelled scepticism, and 3) non-misinformation related deepfake discourses. The first and third of these are relevant to our first research question: understanding how people reacted to deepfakes online. The second of these, addresses our second question showing epistemic distrust as a result of deepfakes. Each of these will be discussed in the following section using fictionalised versions of the quotations from the dataset.

3.4.1 Deepfakes and misinformation (interactions with news articles, deepfakes in warfare, solicitation of deepfakes)

A significant portion of our dataset consisted of commentary and general reactions to deepfake-related news articles (n = 246). There was a disproportionate amount tweets which

consisted entirely of news articles in the dataset (which were excluded) in comparison to the deepfake discourse that was included for analysis. Users often emotionally reacted to the news about deepfakes with negativity (n = 29), worry (n = 27), shock (n = 6) and confusion (n = 6), mostly targeted towards news of the deepfake Zelenskyy surrender. This highlights how reactions to news articles seemed to overshadow independent discourse relating to deepfakes. Many tweets in the dataset reacted to the news of the deepfake as something that had been expected (n = 36), highlighting the many attempts of pre-bunking carried out by the Ukraine government." Well this confirms the warning that Russia would use a deepfake of Zelenskyy surrendering". The reaction to deepfakes in the war was generally negative. "Deepfake tech has arrived, and it's absolute madness" Many tweets sought to explain deepfakes and deepfake news for other users, spreading information about their prevalence and defining the technology. Some spoke to the anecdotal experiences they had had with deepfakes during the war (n = 3). A significant portion revolved around the technical appraisal of the technology, often reflecting on their poor quality (n = 109). Specifically, many users criticised the quality of the Zelenskyy deepfake, which users felt did not live up to warnings of deepfakes or their own expectations of the technology. "You've seen the deepfake of Zelenskyy, now we've seen a deepfake Putin declare surrender. This is clearly better quality than the Zelenskyy cheapfake but it's still poorly made. Unfortunately, the technology will get better with every advance of the technology". Some Twitter users reacted positively to deepfakes used during the war, particularly those used for entertainment value (n = 5), especially demeaning videos such as those putting Vladimir Putin's face onto Gollum from the Lord of the Rings movies. Other people positively reacted to the deepfakes used for misinformation (n = 15), especially towards those made of Vladamir Putin. In some cases, they framed them as revenge for the Zelenskyy deepfake. "In an outstanding retaliatory move by pro-Ukraine VFX artists, a deepfake of the Russian president has been released as a response to the deepfake of the Ukrainian president. Gives insight into the future and how edited videos can be used as a response to deepfakes."

One element of deepfakes that tweets seemed to engage in strongly was the potential of deepfakes in warfare. Deepfakes were seen as a new form of weapon (n = 28) or a new kind of propaganda (n = 10). “You’d think deepfakes are harmless, if you’ve only seen silly videos of deepfaked Keanu Reeves. Unfortunately deepfakes can be a new and vicious type of propaganda. We’ve seen it now with deepfakes of the Russian and Ukrainian leaders” In January and early February, people feared deepfakes would be used as a false flag to start the war (n = 22). “Russia have plans to use a deepfake video to justify an invasion of Ukraine. Before making a statement I need some time to process this”. As a response to this, many tweets also focused on detecting and preparing for deepfakes (n = 17) and criticising government bodies for perceived inaction (n = 20). In particular, German language tweets tended to express doubts about their own politicians’ abilities to detect deepfakes. The Klitschko hoax phone call was perceived as a worrying or embarrassing indicator of government cyber-security. “Some political actors are very behind the times. The stupidity of using technology which is not compliant with modern internet safety protocols. There’s no excuse when there are means of authenticating video calls”. The doubts over cyber-security led to many users encouraging stricter media controls (n = 14). These often framed the rise of deepfakes and misinformation as a justification of the decision made by the Ukrainian government to take over private news agencies. “Ukraine is completely justified in taking over news companies, it is very important to stop Russia’s misinfo. Putin used a deepfake of Zelenskyy surrendering and has used their propaganda news to rig a US election.” Some also demanded significant improvements in governments cyber-security or the banning of deepfake technologies altogether.

A disturbing trend in the dataset involved Twitter users, often jokingly, requesting that people create deepfakes of world leaders in incriminating or humorous defamatory videos. These often included placing world leaders into videos as dictators, movies such as Downfall (Der Untergang) or Charlie Chaplin’s the Dictator. “Anyone made a deepfake yet of Putin as Hitler in that Bunker scene from downfall?” However, a significant portion of the requests for

deepfakes involved the president of Russia (n = 190). People specifically requested deepfaked pornography, which was seen to humiliate or demean Putin, occasionally involving bestiality (n = 3) and BDSM (Bondage and Discipline, Dominance and Submission, Sadism and Masochism). “Would it be possible for some dedicated VFX [Visual Effects] guy to make a video of Putin, heh “developing a connection with a barnyard animal”?” A significant portion also desired to place him into pornography with other men (n = 30). While some justified this as a consequence of Putin’s own anti-LGBT (Lesbian, Gay, Bisexual, Transgender) policies, many created an uncomfortable association between sexuality and shame. “We need to stop this invasion. What if we show Russian politicians a high quality deepfake of Putin in an orgy with a ton of men. We ask them to stop the war or we will plaster the video all over the internet. No offense to gay people, but this could stop a war”. These requests were often accompanied by requests to disseminate and distribute the defamatory video (n = 92), such as hacking Russian news stations or televisions (n = 28).

Some users however did request more positive deepfakes of world leaders which would tell the truth (n = 23), for example to have Vladimir Putin admit to the war crimes carried out by Russian soldiers. “It’s time for hacker groups to join forces with VFX tech. Russia isn’t telling it’s people what’s really happening, so we’ll make a deepfake Putin tell them the truth. No lies, just showing them that civilians are dying and it is a real war”. While distinctly more noble this was still often accompanied by asking if anyone could hack Russian media to present the deepfake. “It’s time to play dirty as well, we should do a deepfake of the President of Russia. We can hack it into Russian news sites and show the real truth of what’s happened during this war”. This shows a worrying tolerance towards the production and distribution of deepfake disinformation when it fits one’s own political beliefs.

3.4.2 Deepfake fuelled scepticism

Many tweets in the dataset expressed a healthy scepticism towards deepfakes. People often warned about the dangers of the technology (n = 61) and mentioned how they were preparing for deepfakes (n = 3), and how to identify the technology (n = 7). “Prepare for

Russian deepfake disinformation. They are going to use deepfakes to spread propaganda and make excuses for their war crimes". Many users engaged in fact checking the media they were consuming and identifying signs that videos were (n = 11) or were not (n = 25) deepfakes. Good media practice was encouraged, tweets promoted healthy scepticism of online information (n = 39), these tweets occasionally highlighted the important work done by government and news agencies in detecting and rebutting fakes. "Sources say, the Russians are planning a deepfake of Zelenskyy soon where the president will falsely surrender. Be prepared".

Unfortunately, the majority of this type of Deepfake discourse during the war consisted of unhealthy scepticism fuelled by deepfakes. Fears of deepfakes often undermined users trust in the footage they were receiving from the conflict (n = 85) to the point where they lost trust in any footage coming from the conflict. "We can't always trust our eyes anymore. Not that this video is fake, but all possibilities need to be considered". Many Twitter users expressed the opinion that no information from the conflict could be inherently trusted, with the implication that sinister agencies were presenting false narratives. "None of the videos coming out of this war can be trusted. In the next few years, the media will be using sophisticated deepfake software. If you think we have a problem now, wait until we start seeing completely faked videos, as opposed to their current lies". This quote also highlights the scepticism of journalism and new media (n = 30) as an interpretive repertoire in the dataset. "This is a western media deepfake. These journalists are under the globalists thumb. We know what will happen next. Praise the lord". At its most extreme the anti-media sentiment seemed to be used to justify conspiratorial thinking and a distrust in reliable sources of news. This highlights how deepfake discourse can be used in arguments which undermine the veracity and trustworthiness of news media.

Most emblematic of the epistemic impact of deepfakes were the deepfake accusations, particularly in situations where real media was accused of being deepfake. Tweets in the dataset tended to label real media as fake more (n = 60) often than they correctly fact-checked deepfakes (n = 11). Often this media was of real footage or of footage taken out of context.

These accusations were often directed towards an incident where a video of Vladimir Putin appeared to show the Russian President's hand pass through a microphone. While this was later shown to be a video artifact, a significant number of tweets accused it of being a deepfake. "Look at his hand passing through the mic. This is a possible deepfake and a poorly made one at that. The Russian president wasn't in attendance". These claims were often used to justify the theory that Putin was hiding or that he was suffering from serious health issues. "Vlad's clearly at deaths door. He has had an abdomen surgery and those addresses have been deepfaked". Often users would call CGI footage deepfake (n = 12) or confuse the terminology of CGI and Deepfakes (n = 3) when accusing other users of spreading fake news. The word deepfake, took on the role of an adjective in these cases often referring to concepts or people instead of media (n = 15)."He's got the world consuming his Deep Fake News spreading his Deepfake ideas and his Deepfake beliefs". When used as a buzzword and applied to people and groups, the word deepfake was generally used insultingly. "This is all nonsense from the Deepfake liberals, they've made Zelenskyy into an idol". The proclivity of users to call people and governments deepfake and the tendencies to use the word deepfake as a buzzword are a worrying indicator of the lack of awareness of what deepfakes are.

In the most extreme cases of scepticism in the dataset, tweets contained conspiracy theories claiming real world events and individuals were deepfakes. The falsely accused video of Vladimir Putin was one of many examples of deepfake conspiracies theories which focused on deepfaked world leaders (n = 155) or deepfakes being used to represent world leader who were in hiding (n = 30). More broadly, users suggested the war was not as it seems, referring to the entire conflict as a deepfake (n = 21). "We are being deceived by Ukraine, they are probably laughing with Putin over our dollars. The war is a deepfake". These conspiracies often justified this accusation by relating it to larger criticisms of governments or passing the entire war off as anti-Russia propaganda. "We need to figure out what is really true. Most importantly, the truth that Ukraine isn't real, we are being fed lies against Putin. All our incompetent leaders in the west, under certain payrolls, are hiding this from us. Look at the Wikipedia page for deepfakes".

This quote shows how conspiratorial thinking was often related back to broader interpretative repertoires of conspiracies, often referencing interpretative repertoires of conspiracy theories such as deep states and antisemitic conspiracies (n = 10). “THE SEMITIC DEEPSTATE HAS HAD FREE RUN TO SPREAD THEIR PROPAGANDA AGAINST PUTIN AND THEY MUST BE HELD ACCOUNTABLE”.

Not all conspiracies were endorsed so wholeheartedly, some users explained the conspiracies as more a humorous thought experiment and reflecting on the implausibility of it. “The nutty conspiracy theorist in me thinks maybe it isn’t real. Either he’s worn the same outfit every day since the first video or he’s been deepfaked”. Some people created sarcastic conspiracies as a way of criticising other users’ opinions and claims (n = 12).”Putin was obviously created by a lab in China and will be broadcast to us all using 5G and our vaccines”. While a lot of these conspiracies were presented earnestly, the outlandishness of these theories often made it hard to judge. Some of the more preposterous conspiracies came from users and in response to media that was seriously posting the conspiracies. The theory that Vladimir Putin had a health issue and was temporarily replaced with a deepfake was picked up as a rumour and ran by a few tabloid newspapers. Some tweets in the dataset then presented this tabloid reporting on online rumours as evidence for the conspiracy. “This news report suggests that Putin has been using deepfakes for his public appearances, which have been pre-recorded to hide his deteriorating health”.

A less extreme aspect of conspiracy theories in the dataset also involved attributing the creation of the well-known deepfakes to government bodies, for example claiming the two deepfake presidents were made by the opposite governments (n = 33). The two presidential deepfakes were often framed as responses to each other in state sponsored cyber-warfare “Very cool. After Putin’s laughable deepfake of Zelenskyy, the Ukrainians have got their payback with this epic deepfake of Putin. Very funny!”. Some conspiratorial accusations in this vein involved claiming that Ukraine and Russia were making the deepfakes of their own presidents so they

could claim victimhood (n = 13) “Does anyone else think that the Russians created the Putin deepfake so it matches the Zelenskyy deepfake?”.

3.4.3 Non-misinformation related deepfake discourses

A significant characteristic of non-misinformation related deepfake discourses was humour, both through jokes (n = 84) and through reactions to humorous deepfakes (n = 45). Jokes were used to criticise politicians and to mock conspiratorial deepfake beliefs. Some users made requests for non-serious deepfakes based off current news stories. “If we made a deepfake of Putin making fun of Jada, Will Smith would slap him so hard it could end the war. MI6 should get onto this”. Similarly, humorous deepfakes of the figureheads of the Ukraine war were often used to criticise politicians by placing them into unflattering movie scenes. Users reacted positively to these deepfakes, especially the deepfake of Vladimir Putin being swapped into a violent death scene of the movie Inglorious Bastards. “There is a great deepfake of Putin getting his head caved in with a bat on YouTube”. A small number of people found these deepfakes bizarre and in poor taste. Humorous deepfake content when related to the war was framed as embarrassing. “Someone please get me off this wild ride, there’s a deepfake of Putin getting beaten to death in Inglorious bastards while Biden watches”. In contrast to the requests for demeaning and pornographic content, these requests were not seen as a weapon in the Ukraine war, nor did users request them to be used in hacks. Instead, they were viewed as strange indicators of the current state of internet discourse. “Social media really has the world’s worst opinions on Ukraine. There’s videos of people dying soundtracked by folky songs about ww3, people stanning Zelenskyy like Beyonce, Deepfake Putin doing TikTok dances”. Social media companies were also criticised for leaving up violent Putin deepfakes online with some Twitter users believing that they would’ve been taken down if they weren’t of Putin. “It’s a well-made-deepfake. Biden watching Zelenskyy beating Putin to death with a baseball bat. From Inglorious Basterds. If it was a video of Biden or Zelenskyy getting their heads bashed in people would be reporting it to YouTube for removal”.

References to educational deepfakes also existed in this dataset, consisting of users spreading an educational deepfake with a summary (n = 27). Many of these were positive, especially toward a deepfake of Vladimir Putin soliciting donations for Ukraine. However, reactions to the CGI videos posted by the Ukraine government often expressed distrust towards other media produced by Ukraine (n = 15). “Look how well made this Ukrainian deepfake is. You’d start to lose faith in how real their supposed evidence of Russian involvement in the MH17 plane crash”. Despite the positive messages of these deepfakes it was felt that by showing people they had the capacities to create fake footage, the Ukraine government undermined their own credibility.

3.5 Discussion

Since computer-generated deepfakes emerged in 2017, researchers have speculated that deepfakes would be used in warfare for propaganda and misinformation (Westerlund, 2019). Here we report the first study of deepfakes used in an active war. In line with researchers’ predictions about the harms of deepfakes in war time, there were a great many deepfakes shared online during the period studied. The most notable were the false surrender videos of the Russian and Ukrainian presidents. Though of very poor quality, had these deepfakes been believed, they may have had wide-reaching consequences for the war. In particular, the deepfake of the Ukrainian president undermines the idea that relying on reputable sources will prevent the harms of deepfakes (Sharma & Kaur, 2022), as the video was spread through the hack of a reputable Ukrainian news website. Though there were examples of harmful deepfakes in the war, the majority of deepfakes were not serious attempts to stop or start a war; predictions have severely underestimated the satirical and humorous use of deepfakes. Another important event during the war, the “deepfake” Klitschko video call, indicates the tendency of news articles and governments to label unknown video manipulations as deepfake. In light of the epistemic harms of deepfake discourse, the importance of technical literacy in news reporting around deepfakes is important.

This research had two aims: to study people's responses online to deepfake content during the Russo-Ukrainian war and to explore if there was practical evidence of the effect of deepfakes on epistemic trust. Our thematic analysis highlighted three main areas of deepfake discourse online; deepfakes and misinformation, deepfake fuelled scepticism and non-misinformation related deepfake discourses. We found evidence for the epistemic harms of deepfakes in cases where people doubted the veracity of real videos and in cases of deepfake conspiracy theories. In the following section we will discuss what these results teach us about how deepfakes are perceived online, and the implications for journalistic practice and future academic research.

3.5.1 Reactions to deepfakes in the Russo-Ukrainian war

We found that perceptions of deepfakes and deepfake news in the dataset were generally negative. Tweets generally focused on the harms of the technology, expressing fear and shock at the potential harms of the technology similar to the fears expressed in news articles and academic research on deepfakes (Gamage et al., 2021; Gosse & Burkell, 2020). Positive discourse around the technology often celebrated the use or potential use of deepfakes in cases where the target was seen as justified. Users often asked for deepfakes to be made of world leaders and to be disseminated through hacks. This evokes past research into deepfake communities on Reddit, where redditors often requested deepfakes for specific purposes (Gamage et al., 2022). This is indicative of the crowd-sourced nature of deepfakes and how the social demands for certain deepfakes influences what is created and spread (Gamage et al., 2022). Another element of deepfakes which became apparent through our analysis was the absolute prevalence of deepfake discourse relating to pornography. Past research has estimated that 95% of deepfakes online are pornographic in nature (Ajder et al., 2019). Despite the Ukraine war showing the potentials of the technology in war and misinformation, there is clearly a widespread tendency to use deepfake technology to create non-consensual pornography, even when relating to the leaders of military conflicts. This harm does not depend on the capacity of deepfakes to deceive it instead was seen as a way to publicly attack a

politician's image. Our research showed people's desire to shame and humiliate world leaders, but this technology is and will be used, to bully and harass people in their everyday lives (Kugler & Pace, 2021). The impacts of harmful defamatory deepfakes and deepfake pornography remain an important and understudied harm of the technology. The prevalence of this harm highlights the need for researchers and policy makers to tackle defamatory deepfakes through research and legislation.

A significant element of deepfake discourse highlighted in our analysis was the humorous and educational discourses around deepfakes. While these have generally been viewed as positive potentials of deepfakes (Westerlund, 2019), our dataset showed more complex views on the non-misinformation deepfakes. Because they used images of politicians and often related to current events, the humorous deepfakes in the war still served as a way for users to make political commentary through satire and parody. Humorous deepfakes were generally celebrated by users, but many found the videos uncomfortable, especially in instances where the deepfakes were graphic or violent. This highlights the need for more nuanced research into creative uses of deepfakes.

3.5.2 Deepfakes and distrust

This study provides tentative indications as to how the epistemic damage of deepfakes exist in an online environment and the complex interactions it encourages between truth and falsity existing on social media. The liar's dividend is the result of an information environment where real information can be easily discredited as fake (Chesney & Citron, 2019b). Our research provides insight into the creation of an information environment which may potentially lead to the liar's dividend. Real video and images were decried as deepfake, people often mistook CGI for deepfakes and used deepfakes as a catch-all insult for information they did not like. This shows the practical implications of the previously-studied relationship between awareness of deepfake technology and how it undermines one's belief in real media (Tahir et al., 2021). There were also two incidences in our dataset where tweets misidentified real war footage as deepfakes but because of our methodological restrictions on sensitive combat footage

we cannot make any claims as to the prevalence of this. While the generally assumed threat of deepfakes is their believability, the labelling of real media as deepfakes shows the capability of deepfake discourse to undermine truth (Whittaker et al., 2020). Though this paper cannot account for pre-existing conspiratorial beliefs or the impact of malicious "bad actors" using deepfakes to encourage scepticism for their own advantage, our results still highlight that deepfakes are being discussed in epistemically harmful ways. We found deepfakes were used for accusations and to cultivate unhealthy scepticism of real media, though future quantitative research is required to highlight both the prevalence of deepfake accusations and its potential effects on trust.

The use of deepfake and CGI educational media by the Ukrainian government was often viewed quite negatively in the dataset. Many users felt that by showing their ability to create deepfakes, the government had undermined the credibility of other authentic video evidence. This has broad implications for educational deepfakes and their potential harms. The current study suggests that educational deepfakes may damage the credibility of the reputable agencies which produce them. By making educational deepfakes, one is only showing the public that their company or agency can produce highly convincing fake videos. Previous quantitative research has shown that educational disclaimers can to some extent mitigate the harms of pure deepfake misinformation (Vaccari & Chadwick, 2020), but our research indicates that educational deepfakes may still undermine truth in comparison to unedited videos. The effects of educational deepfakes on deepfake related scepticism and the liar's dividend require further study, but we suggest that governments and organisations that rely on the public's trust should avoid using deepfakes as part of their messaging.

A novel finding of the analysis was the observed interactions between deepfakes and conspiracy theories. The use of image editing and video editing techniques in conspiracies is not new. The moon-landing conspiracy was fuelled by the current state of Hollywood cinema and visual effects at the time, specifically those used in Stanley Kubrick's 2001: A Space Odyssey (Gagliardi, 2017; May, 2017). Deepfakes and the liar's dividend may provide another tool of

deniability to conspiracy theorists. Conspiracy theories harness suspicion of video evidence, which deepfakes have the potential to fuel (Lamb, 2020). Advancement in modern video-editing technologies is likely to inform a host of theories revolving around deepfake replacements of world leaders or world leaders using deepfakes to improve their speeches or to hide their own illnesses. For example, on the 27th of July 2022, a 40 second video of US president Joe Biden in which he did not blink led to many Twitter users to decry the video as a deepfake and endorse similar conspiracies to those that emerged during our analysis (BBC, 2022). This paper is the first to find evidence of online conspiracy theories which incorporate deepfakes. Conspiratorial deepfake beliefs are a potential problem of deepfakes going forward and one that necessitates further research.

3.6 Limitations

An important limitation of this study is that Twitter may not be representative of other social media platforms. The format of tweets, limited to 280 characters and the prevalence of news sites and bots means the data gathered from the service tends to be less textually dense than information from Reddit or Facebook. Most tweets consisted of short sentences, and longer posts tended to be simple summaries of news articles. The lack of lexical density significantly limited the ability to derive some of the intended meanings, at times it was impossible to gauge if users were being serious or not with their posts. This is an issue that researchers have encountered when studying language, particularly the study of sarcastic language, on Twitter (González-Ibánez et al., 2011). This was especially true regarding conspiratorial posts where often the ridiculousness of the claim was not indicative of its seriousness. Future research could explore more textually rich online spaces such as online forums and blog posts.

Because the extracted data was from a Western media service, the search terms were in English, and the tweets were in Western European languages, the dataset has the potential for political bias. Telegram and Russian social media were significantly involved in the spreading of anti-Zelenskyy media and deepfakes which may not be fully represented in our dataset (Allyn, 2022). As such it is impossible for us to make conclusions as to whether the deepfakes

and discourse of the crisis were more anti-Russia (as generally observed in our data set) or anti-Ukrainian. Another bias is that the dataset did not include any tweets that had been deleted before data collection so it may not be representative of the misinformation spreading at the time.

A minor limitation of the research is the inability to determine if any of the content was produced by politically motivated "bad actors". Considering the unverified and anonymous nature of many Twitter users, much of the discourse undermining truth could potentially have been motivated in this manner. However, politically motivated or not, this discourse is still indicative of the real-world social media discourses around deepfake.

One further limitation of this research is the use of the word deepfake in the search terms. This excludes from our dataset incidences where people discussed deepfakes without knowing that they were watching edited footage. Though it would have proved unfeasible to carry out a search for media that people did not recognise as being deepfaked, it is still important to note that people who accepted or believed the deepfake media to be real are missing from our dataset, especially in the discussion of people's ability to be misled by deepfakes.

3.7 Implications

This research has implications into how news media and governments interact with the general discourse around deepfakes. In our dataset, the number of news articles and tweets spreading stories about deepfakes significantly eclipsed actual user discourse. Given that the liar's dividend is likely to grow with education and efforts to curb deepfakes (Chesney & Citron, 2019a), the dominance of news media in our dataset may give credence to claims that deepfake coverage has outpaced the use of actual deepfakes in misinformation (Gosse & Burkell, 2020). We need to consider if the news focus on deepfakes is disproportionate to the threat we are currently facing and whether this response is creating more distrust and contributing to an epistemic crisis. Warnings about deepfakes need to be measured against their

probability, and due diligence may be necessary before claiming videos which are traditionally faked are actually deepfakes. In many cases, preconceptions and fears about deepfakes informed the anti-media sentiment throughout the dataset. The implication of this is that news media needs to weigh the benefits of pre-bunking and information inoculations against the risks of the liar's dividend and the undermining of real information. News coverage should also be careful in the labelling of suspected deepfakes, for example news coverage of the dubious supposed deepfake video call with Mayor Klitschko, which possibly only used traditional video fakery, encouraged users to doubt other interactions between politicians and accuse them of being deepfakes. In an information environment where people are endorsing deepfake conspiracy theories and decrying real media as deepfake, it becomes vitally important that news discourse around deepfake encourages literacy in identifying deepfakes and other forms of fakery. News coverage of deepfakes needs to focus on educating people on what deepfakes are, what their potential is, and both what their current capabilities are and how they will evolve in the coming years.

This research has implications as to how deepfakes can be used in warfare. The harms deepfakes have on people's trust and their concept of truth significantly undermine faith in governments. In our dataset, some users responded negatively to the educational deepfakes, and CGI made by the Ukraine government. While well intentioned, the use of educational synthetic media encouraged Twitter users to frame the real combat footage in Ukraine as deepfakes. The timeline we have established for the Ukraine war shows that deepfakes are by the time of writing having real world impacts on the war. The Zelenskyy deepfake serves as a worrying indicator of the potential harms of the technology. The general playbook against misinformation and deepfakes often encourages people to check the source of any claims or videos. However, the deepfake of Zelenskyy was distributed in tandem with a hack of both the TV station and news site of a reputable media source. Deepfake literacy and a healthy scepticism of outlandish claims should be encouraged as a way to avoid the harms of deepfakes used in hacked media.

This is consistent with research on the protective nature of deepfake literacy and awareness against the harms of deepfake misinformation (Hwang et al., 2021).

In the dataset, almost no tweets involved individuals' belief in real deepfakes, instead most of the misinformation came from labelling real media as deepfake. The word deepfake tended to devolve into meaning "extra fake" when used as an adjective against individuals, CGI and other non-deepfake media. Deepfakes, already suffering from the lack of consistent definitions, seem to have become a buzzword and an insult. A lot of the deepfake scepticism in our dataset seemed to only have a tentative understanding of what deepfakes were. Non-media objects such as individual people or even the entire war were accused of being deepfakes. The lack of general knowledge of what deepfake means and the tendency to use it towards other fakes has implications for academic researchers. In research that measures proclivity to share deepfakes, the measurements of deepfake sharing are biased by individuals' perceptions of what deepfakes are. It is possible that if someone has only been presented with a rough definition of deepfake that they may be using the term to describe CGI or traditional fakery. To avoid this, academic research on deepfake sharing behaviours should encourage people describe any deepfakes they have fallen for qualitatively to ensure that they are indeed deepfakes.

3.8 Conclusions

The use of deepfake technologies in the Russo-Ukrainian war is a significant moment in the history of deepfake technologies. For the first time we have seen deepfake propaganda and misinformation that has attempted to influence a war. Despite the use of deepfakes in fake surrender videos, the chief focus of deepfake discourse is still the personal defamatory harms of deepfakes. Individuals tended to overlook or even encourage the harms of defamatory deepfakes when they were directed towards political rivals. The current research provides the first practical qualitative evidence of the of the epistemic harms of deepfakes on social media. In the dataset, real videos were accused of being deepfakes and deepfakes fuelled conspiratorial beliefs and unhealthy scepticism. The use of deepfakes in education and entertainment must also be reconsidered in light of the epistemic harms of deepfake technology.

Chapter 4: Do No Harm: Designing Interventions for Deepfake-Fueled Conspiracy Theories

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.

This paper was published in ACM Interactions on the 16th of October 2025.

Study Motivation

The motivation of this work was to briefly follow up on the findings of Chapter 3 by examining in more detail the concept of deepfake conspiracies. Deepfake conspiracies are one way in which deepfakes can harm trust in real media. Video evidence can now be potentially called into question. In Chapter 3, deepfake conspiracies focused on rumours on the health of the Russian President. I felt that this was a novel finding which required further elaboration.

Another motivation of the work was to further understand the usage of deepfakes in misinformation and the harms to epistemic trust. I wanted to branch out and link deepfakes to broader literature on scepticism and the existing research base on conspiracies. The paper gave me the opportunity to work with my colleague, Cian O'Mahony, who had just recently finished a scoping review on interventions for conspiratorial beliefs. The findings which emerged in that review also involved the counter-intuitive idea that education can make scepticism towards all media worse. I wanted to take a look at existing efforts at deepfake literacy and detection and see if they were also increasing scepticism. The recommendations on intervention design motivated my decision to end the PhD with a paper designing a deepfake intervention, though the aim of that intervention was to target the harms of deepfake abuse imagery.

Author Contributions

This study was conceived and written by JT and CO'M. with guidance and feedback from GM and CL. All authors approved the final version of the manuscript

4.1 Abstract

Conspiracy theories are a set of explanations about the world where significant (and often negative) events are believed to be orchestrated by clandestine groups. Frequently emerging during periods of tragedy or ambiguity (e.g., the Covid-19 pandemic, the September 11 attacks) conspiracy theories are typically an attempt by individuals to find closure and meaning in these circumstances. This belief system operates as a framework through which individuals seek clear answers, providing them with a structured understanding of how the world functions. The content of conspiracy theories is influenced by advancements in media editing and production technologies. As far back as the first moon landing, conspiracies referenced Hollywood's visual effects advancements of the time (e.g., Stanley Kubrick's *2001: A Space Odyssey*).

4.2 Insights

- Novel media-editing tools, such as deepfake technology, are used to create misinformation and disinformation
- Designing interventions to challenge deepfake conspiracy theories is crucial.
- Well-intentioned deepfake interventions may counterintuitively cause more harm than good by undermining trust in media.

4.3 The use of deepfakes in conspiracy theories

More recently, conspiracy theories claimed that the Sandy Hook school shooting and the murder of George Floyd were carried out by paid actors and faked by using video production tools. There are also popular conspiracy theories that similarly argue professional body doubles and actors have replaced individuals such as Melania Trump and Avril Lavigne. The advent of deepfake videos gives agitators engaged in these types of conspiracies a new weapon to undermine people's ability to trust the veracity of video evidence.

Deepfake videos use deep learning to insert someone's image into a video, often saying things they never said. Deepfakes have the potential to be used in disinformation and misinformation campaigns—for example, spreading a deepfake video defaming a politician the night before an election—as well as more broadly to undermine trust in video evidence (e.g., voters disregarding real and credible evidence of a politician engaged in something nefarious and rejecting the video as a deepfake). Novel media-editing technologies, such as deepfake technology, may provide a valuable tool for conspiracy theorists who seek to discredit factual video evidence. The first notable incident of a deepfake conspiracy involved former Gabonese President Ali Bongo. Many people believed that he died after being hospitalized for a stroke in 2018. While official sources released videos that proved the president was alive and well, conspiracy theorists discredited the videos as being deepfaked or using body doubles. The conspiracies created political instability in the country and lead to a failed coup d'état.

Our past research on the use of deepfakes during the Russian invasion of Ukraine highlighted the harms of deepfake conspiracies (Twomey et al., 2023). Specifically, we found that some users questioned the validity of all war-related media after being exposed to deepfake and CGI educational content about the war. A wide range of deepfake-related conspiracies emerged during the conflict, with many claiming that world leaders were using deepfakes for various reasons. One such conspiracy theory, which was picked up by tabloids, said Russia's president, Vladimir Putin, was desperately ill and using deepfakes to hide his condition. The supposed evidence for this theory—a video of Putin's hand passing through a microphone at a meeting—was debunked as an artifact of Twitter compression.

Such theories will only become more popular as deepfake technology becomes more prevalent. In the time since we carried out that research, a video of former U.S. President Joe Biden has also been falsely labelled a deepfake. The supposed evidence for the accusation was that the Biden "deepfake" did not blink (17 seconds between blinking is not impossible and there was also a cut within the video). The conventional wisdom around identifying deepfakes is undermining real information, and this needs to be considered in deepfake research.

4.4 Conspiracy Interventions: Avoiding a Backfire

Our research team reviewed 25 studies that analysed the efficacy of conspiracy theory interventions and found the majority were not successful (O'Mahony et al., 2023). Some interventions, however, did produce medium to large effects in reducing belief in conspiracy theories. These interventions mainly focused on preventative measures, such as priming participants to engage conspiracies with an analytical mindset or debunking conspiracy theories pre-emptively, before participants were exposed to them.

One intervention was effective in reducing a wide range of counterfactual beliefs by enrolling participants in a pseudoscience university class. Students were encouraged to both make arguments for various counterfactual beliefs and to actively evaluate those beliefs with the skills they learned in class. This group was compared to a standard research methods class. At

the end of three months students in the pseudoscience class believed in significantly fewer conspiracy theories than those in the research methods class (Dyer & Hall, 2019).

These findings highlight the importance of teaching explicit skills to analyse misinformation that people can put into practice, in contrast with peripheral knowledge that is passively consumed. Furthermore, previous interventions addressing misinformation have highlighted the importance of active interventions, as they motivate participants to be more inclined in the future to apply what they have learned. Importantly, the review also highlighted the many ways in which conspiracy interventions may actually increase belief in conspiracies, as several of the interventions identified in the review resulted in a backfire effect. Evidence of this effect has also been found in a handful of previous misinformation interventions, though some researchers argue the results are inconclusive.

4.5 Deepfake Misinformation Interventions

Current attempts to design interventions for deepfake misinformation have focused primarily on improving the ability of individuals to identify deepfake videos. As the quality of deepfake videos improves and other novel AI-generated media becomes more prevalent, however, the type of knowledge needed to recognize them will also change. There is already a so-called arms race between the technology that identifies deepfakes and one that generates them. As illustrated by the Biden and Putin examples, a layperson's understanding of deepfake features (e.g., the absence of eye blinks) can be incorrect or outdated, and it can be used to discredit real media.

The erosion of trust in media is fundamental in predicting deepfake harms. Indeed, "the liar's dividend" theorizes that, as fake videos become more prevalent, individuals will use public scepticism to cast doubt on real media. The awareness of deepfakes alone may be enough for people to lose trust in real media and for the political process to become destabilized. Thus, counterintuitively, education around the harms of deepfakes may actually increase belief in misinformation. Similarly, when people who are highly concerned about deepfakes accidentally

share a video that includes them, they tend to become highly sceptical of all news media. The public's decreasing trust in the validity of real media may be as harmful as people believing deepfake media.

Research has shown that exposure to false labelling of real videos as deepfakes may undermine authentic political speech (Hameleers & Marquart, 2023). A similar phenomenon was observed when researchers tested fake news interventions. Such interventions attempted to reduce participants' susceptibility to misinformation, but they resulted in participants rejecting both factual information and misinformation. As a result, we argue it is important that interventions for deepfake conspiracies reduce people's susceptibility to fabricated media without making them overtly sceptical of real media.

Many deepfake interventions have focused on teaching people how to identify fake videos and not how the interventions may damage the veracity of real videos. Rashid Tahir and colleagues conducted research (Tahir et al., 2021), showing one real video and three deepfakes to two groups of participants. One group was given a specialized training that walked them through specific examples of deepfakes and highlighted the features used by detection algorithms. Tahir and his colleagues found that, while the training increased the treatment group's ability to detect deepfakes significantly, it also reduced their ability to correctly identify real media by nine percent.

Further evidence comes from research carried out by the MIT Media Lab, which found that, by the end of a deepfake detection test, participants were both 18 percent better at detecting deepfakes and 18 percent worse at identifying real videos (Groh et al., 2022). In a test where half the media is fake and the other half is real, an increase in both identifying and misidentifying deepfakes does not affect overall measures of accuracy. In real-life information environments, however, where deepfake videos are not as prevalent as real media, increasing the amount of people who do not trust real media is disproportionately dangerous. Importantly, both pieces of research measured the false positives (people accusing real media of being fake),

which is a good practice in misinformation research and needs to be more common when evaluating the efficacy of deepfake interventions.

4.6 What Should a Deepfake Conspiracy Intervention Look Like?

We have already shown how the current interventions for deepfakes and conspiracies have significant limitations. The harms of deepfakes increase with the awareness of technology. The use of strategies such as inoculations and priming to deal with conspiracy theories may, counterintuitively, increase scepticism of real videos. To avoid this, we recommend that interventions for deepfake conspiracies be designed with the following principles in mind.

Consider the disproportionate impact of false positives and/or negatives. Interventions should consider the social impact of false positives and negatives. If a deepfake intervention causes some individuals to be overly sceptical of real media, then it is a poor intervention (assuming there is more real information in the media ecosystem than fake).

Focus on decision making over identification. Interventions should avoid treating deepfake susceptibility dichotomously, overly focusing on getting people to make snap judgments about whether something is a deepfake. The goal of deepfake interventions should not be to turn people into machines that can easily identify media as deepfake or real. Rather, interventions should encourage healthy suspicion of media and focus on how people treat suspected mis/disinformation.

Consider the effects of deepfake prevalence. Future interventions need to question the extent to which they carry information about prevalence. It is important to consider whether it is wise to train people on deepfake interventions using an equal number of deepfake and non-deepfake videos. This does not reflect the current media environment.

Design for implementation. Interventions should be designed with practical, real-world implementation in mind. Previous research has highlighted that many effective conspiracy interventions could not be realistically applied to settings outside a laboratory environment.

Measure effects over time. Interventions should assess whether the effects they induce persist over time. It is important to establish whether both positive and negative effects will last longer than a few minutes—something that is lacking in current deepfake and conspiracy theory research.

Encourage motivation to apply learned skills. Interventions should aim to motivate participants to sustain the use of the skills and tools acquired during the intervention. As such, future experiments should measure this motivation as an indicator of the participants' likelihood to persist in applying these skills and tools to potential deepfake content.

4.7 Conclusion

Deepfakes pose a threat to trust in the veracity of video information. Deepfake conspiracies provide a unique use case of the AI technology in misinformation, and we believe they require unique solutions. Existing deepfake interventions have counterintuitively increased the likelihood of participants misidentifying real media as deepfake. It is imperative to not only teach how to identify deepfakes but also how *not* to misidentify real videos as fake. As deepfake technology develops and becomes more accessible and prevalent, and many well-intended interventions and educational initiatives are launched, researchers must remember our guiding principle: Do no harm.

Chapter 5: “What do you expect? You’re part of the internet”: Analyzing Celebrities’ Experiences as Usees of Deepfake Technology

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*.

This paper is currently under review at the journal: The International Journal of Human - Computer Studies.

Study Motivation

The most significant motivator of this work was a highly publicised incident in which a male streamer (a streamer is a person who publicly records or live-streams themselves playing video games or making other content) revealed he had been looking at deepfake videos of other female celebrity streamers. Like the Ukraine case, we viewed this as a strong indicator of the harms of deepfake abuse imagery and decided to conduct a case study on this area. We specifically look at public statements from celebrity women about deepfake abuse imagery.

Initially, the use of deepfakes in creating intimate imagery was not going to be the focus of my PhD. However, I started to realise the extent to which abuse defines deepfakes (literally, as the term deepfakes were defined by a reddit user creating intimate imagery). I decided that studying the harm of AI-generated abuse imagery should be the major contribution of the work. Pivoting to this topic meant dedicating the last three papers of the PhD to this harm. I debated this decision over a number of occasions with my supervisors where I convinced the team that the harms against celebrity women were the most vital to study.

I also tried to justify why I wanted to study celebrity women specifically. I initially justified celebrities as “canaries in the coal-mine” who are the first sign of a harm which is soon

to be prevalent for women in general. As I wrote this paper, I realised that it is also vital to study celebrities as their own case. Celebrity women are at unique risk for large scale organised misogynistic harassment and this harassment is inherently linked to false beliefs about this abuse being part of their job. The false belief of some people "deserving" deepfakes shaped the idea of deepfake myths which was a novel finding of the work.

Author Contributions

This study was conceived by JT under the guidance of GM and CL. The data collection and analysis were performed by JT under the guidance of CL, GM and SF. JT wrote up the paper with guidance from the supervisory team and feedback from all collaborators. All authors approved the final version of the manuscript

5.1 Abstract

Deepfake technology is often used to create non-consensual synthetic intimate imagery (NSII), mainly of celebrity women. Through Critical Discursive Psychological analysis, we ask; i) how celebrities construct being targeted by deepfakes and ii) how they navigate infrastructural and social obstacles when seeking recourse. In this paper, we adopt Baumer's concept of "Usees" (stakeholders who are non-consenting, unaware and directly targeted by technology), to understand public statements made by eight celebrity women and one non-binary individual targeted with NSII. Celebrities describe harms of being non-consensually targeted by deepfakes and they describe various infrastructural/social factors which hinder activism and recourse. This work has implications in recognizing the roles of infrastructures underlying deepfake abuse and the potential of human-computer interaction to improve existing recourses for NSII. We also contribute to understanding false victim-blaming rhetoric about deepfake abuse. Future work should involve interventions which challenge values and false beliefs which motivate NSII creation/dissemination.

5.2 Introduction

Deepfake technology is the latest example in a long history of technology facilitating violence towards women (Giugni, 2021). Deepfakes involve the use of deep learning technology to create false images/videos of their targets. The vast majority of deepfake videos available online are pornographic (Ajder et al., 2019), and the people appearing in those videos often have not provided any form of consent. Deepfake technology is underpinned by infrastructure that is implicitly tied to the abuse of celebrity women (Winter & Salter, 2020), including software repositories and forums that produce non-consensual deepfakes of celebrities. Even supposedly positive uses of the technology in entertainment (such as the consensual resurrecting and de-aging of actors) often use software such as DeepFaceLab, which directly advertises forums dedicated to the production/dissemination of illegal content (Timmerman et al., 2023).

Most existing HCI research on deepfakes focuses on experimentally investigating perceptual differences in people's ability to identify deepfakes (Nissen et al., 2023; Shahid et al., 2022; Wöhler et al., 2021). There has also been an increasing focus on understanding the social context behind deepfakes and implications on society. For example, a synthesis of qualitative studies on deepfakes by Vasist and Krishnan has criticized existing literature for studying deepfakes in isolation without considering how deepfakes are encouraged by platforms (Vasist & Krishnan, 2022b). Their work also highlighted how much existing qualitative research on deepfake abuse imagery studies the platforms and communities which create the imagery as opposed to the experiences of victims. The limited qualitative work on deepfake abuse imagery from a human-computer interaction (HCI) perspective similarly focuses on understanding general public attitudes towards deepfake abuse material online, often through studying platforms such as Reddit (Gamage et al., 2023) and asking the broader general public on their experiences with creating and viewing deepfake abuse imagery (Umbach et al., 2024). Outside of HCI, other qualitative approaches to this topic highlight the lived experiences of deepfake abuse victim-survivors (Rousay, 2023) but do not situate it in the broader context of novel

technologies and how they interact with online infrastructure and communities. Our work focuses specifically on qualitatively analysing the statements of victims of deepfake nonconsensual synthetic intimate imagery (NSII) as a way of understanding current recourses and critical issues with AI technologies.

In the current paper, we examine the infrastructure of existing recourses available and identify that responsibility is unfairly placed on the victims of deepfakes to remove abusive content. Through qualitative methods we highlight that the discourse employed by deepfake victim-survivors provides insights into their experience of manipulated intimate imagery. We analyse how victims identify infrastructural factors in the production, distribution, and consumption of deepfake abuse imagery (or non-consensual synthetic intimate imagery- NSII). We argue that the celebrity victims of deepfakes should be viewed as Usees of the technology as defined by Baumer (2015): stakeholders who are non-consenting, unaware and directly targeted by technology. Celebrities are intentionally targeted by the deepfake software, they are often initially unaware that there are deepfakes made of them, and the videos/generative models override their consent. It is important to understand how these high-profile figures construct their experience of being Usees of deepfake technology as they are most often the targets of deepfake abuse. Examining these public statements may also more broadly help the design of general solutions/interventions for deepfake abuse which will become increasingly relevant as more people are targeted.

This qualitative study will analyse public interviews with celebrity targets of deepfake intimate imagery. Our research questions are:

- How do celebrity victims of deepfake abuse construct being Usees of deepfake technology (the experience of being unaware and non-consenting targets of technology)?
- How do they negotiate infrastructural and social obstacles against seeking recourse for deepfake abuse?

We analysed public interviews with celebrity targets of deepfakes using a form of Critical Discourse Analysis (Fairclough, 2013) known as Critical Discursive Psychology which is concerned with both the 'micro' analyses of discourse construction of discursive psychology alongside 'macro' considerations of societal injustice associated with Critical Discourse Analysis (Locke & Budds, 2020). Our research questions are inherently discursive as they concern the positioning and construction of deepfake targets and their experiences with deepfake content, as well as how current solutions are framed (such as removing content from search engines/social media, and potential legal actions against the creators/consumers of deepfakes). By taking these interviews as a corpus of public statements of the harms of deepfake abuse, we highlight the real-world experiences of current targets. Their status as celebrity women and non-binary individuals means that they 1) are uniquely targeted and victimized by deepfakes as a pattern of abuse against public women and gender minorities, and 2) provide insights into harms of technology which may soon become widespread, as deepfake technology becomes increasingly accessible. The contribution of our work is to critically examine how technology infrastructure and usage patterns create deepfake abuse, how this is constructed by the Usees of deepfake technology in their interaction with this infrastructure, and how this can inform software engineering, policy and interventions.

5.2.1 Note on terminology used

As HCI researchers it is important to be aware of the evolving terms around gendered violence to avoid stigmatization of survivors. There is limited guidance on the correct terminology to use for deepfake videos using manipulated intimate imagery in existing literature, despite criticisms of the phrase "deepfake pornography" (Rousay, 2023), stemming from similar criticism over the phrase "revenge pornography" (McGlynn et al., 2017). Existing terminology tends to imply sexual gratification and ignores both harm and lack of consent. This paper refers to deepfake videos made from pornographic videos as an example of non-consensual synthetic intimate imagery (NSII) (Umbach et al., 2024), or as deepfake abuse,

drawing from the current terminology for image-based sexual abuse (Patrick, 2019). Similarly, there is limited guidance on the preferred terminology for those who have experienced deepfake abuse. When drawing from Baumer's work (Baumer, 2015), we use the terms Usee or target. Otherwise, we use victim, while understanding that as this harm evolves, so will the preferred terminology.

5.3 Related Work

Deepfake technology generally involves audiovisual media that has been manipulated using AI to falsely portray a person in a situation that has not occurred (Westerlund, 2019). While deepfakes are not the first method for manipulating video, they are significantly more accessible than manually editing video footage. From their inception, deepfake videos have been used in the targeted harassment of public women. Deepfakes originated in late 2017, evolving from a previous internet practice of photoshopping celebrities' likeness onto adult images (on the subreddit r/celebfakes and on sites such as 4chan) (Caldera, 2019; Gosse & Burkell, 2020; Suslavich, 2023). A Reddit user under the pseudonym u/deepfake, applied novel video manipulation technology to create similarly modified adult videos which he shared on a new subreddit (Caldera, 2019). While the resulting subreddit for these videos (also bearing the name deepfakes) introduced this technology to a wider audience, it should be noted that similar photoshopped NSII (which were not generated using AI) existed on Reddit and other parts of the internet for years beforehand. The community was characterized as both a space for learning, where its users gave insight and technological advice on the production of the videos (Gamage et al., 2022), and as a place for the dissemination of NSII, specifically modified pornographic videos into which a celebrity's face was interpolated (Timmerman et al., 2023). That forum can be seen as an example of an online environment characterized by "networked misogyny" which combines the anonymity of most internet forums with organized harassment and hostility against women (Marwick, 2017). An investigative news piece in January 2018 increased public awareness of deepfakes and led to the original deepfake subreddit and r/celebfakes being shut down (Godulla et al., 2021). Since then, deepfake creators have

migrated to websites specifically designed for disseminating and creating manipulated NSII. These sites operate similarly to the original subreddit: with sub-forums categorized into general discussion, technical resources and assistance, creation resource sharing, content sharing, content requests and meta discussion (Timmerman et al., 2023). These forums are part of the increasing number of deepfake marketplaces online, who produce deepfakes of famous and non-famous targets for a paid fee (Rousay, 2023; Timmerman et al., 2023).

5.3.1 Celebrity Women as Deepfake Usees

Research has shown that celebrity victims of NSII are viewed less sympathetically than non-celebrities, especially when the celebrity is a woman (Fido et al., 2022). This is consistent with viewing behaviours of deepfakes: research by Umbach and colleagues (Umbach et al., 2024) who found that 9.9% of men and 2.8% of women reported viewing sexual deepfakes of a celebrity (versus 4.2% and 1.6% respectively for deepfakes of ordinary people). Feminist literature suggests that celebrities exist in a fragile and tenuous space between the private and the public where their career involves constant scrutiny on their identity and actions (Patrick, 2019). Celebrities (particularly celebrity women) have always faced constant surveillance and significant breaches to their privacy and personal agency (Gill, 2019). They are assumed to have traded control over their image to the public in return for their continued existence as cultural figures (Patrick, 2022). Information which is not public is considered a secret to be traded and sold as gossip (Patrick, 2022); this practice has only evolved with the advent of social media, with notorious publishing of private images by black hat hackers (Odunze, 2018). Feminist scholars have portrayed this violation of privacy as part of a distinct form of misogynistic harassment (Patrick, 2019): “Her images are stolen and circulated with a dual purpose: to both punish her for her public success, and to remind all women that their bodies/sexuality are never theirs, and are always prone to such violence, no matter how “successful” they become (p. 72)”. Deepfakes build on this in a literal sense, where the body and image involved does not belong to the target at all but is instead the result of a fabrication.

While deepfakes are in theory able to target anyone, most victims remain public figures, and are almost exclusively women (Karnouskos, 2020; Timmerman et al., 2023). A report by the cybersecurity company Deeptrace found in 2019 that 95% of deepfakes online were manipulated pornographic content and that the prevalence of deepfake content had doubled since the previous year (Ajder et al., 2019). This is a growing and global phenomenon. In recent years there has been an overwhelming increase in the targeting of K-pop idols, who now constitute around 25 percent of deepfake NSII online (Wang & Kim, 2022). The targets have also broadened to various forms of online celebrity including social media influencers and streamers (users of live-streaming/vlogging platforms such as Twitch) (Hawkins et al., 2025). As this technology becomes increasingly accessible, its potential to harm also broadens outside of celebrity targets. In Autumn 2024, the messaging company Telegram was investigated by South Korean police for hosting a group dedicated to deepfake NSII with over 200,000 members (McCurry, 2024). This was part of a larger investigation into deepfake groups within universities, where students created NSII targeting fellow students and teachers (Kelly, 2024).

In this work, we configure the targets of deepfakes within Eric Baumer's work on Usees (Baumer, 2015) and post-userism (Baumer & Brubaker, 2017). Previously, HCI research typically positioned individuals as belonging to one of two groups, the users and non-users of a given technology. HCI as a discipline was seen as concerning itself with users of technology, distinguishing the field from related disciplines. Baumer's work emerged from an increasing focus by HCI researchers on "non-use" of technology and contributes the novel idea of "Usees" of technology; a form of interaction which is neither use nor non-use of, but rather use by, technology. Usees are non-consenting and unaware persons who have been targeted by a technology or its users, a relationship with technology which is becoming more common. One example of Usees given by Baumer is that of stalkerware applications, where Foursquare's API was being used to track and map women without their consent (Baumer, 2015). We suggest that the targets of deepfake videos can also be described as Usees, as the content is made without their consent or awareness, and the software is designed to target a specific group: public

women. Our research aims to explore how non-consent, unawareness and being targeted underlie the experience of deepfake abuse.

5.3.2 The Consequences of Being a Deepfake Usee

We situate this work within existing HCI literature which aims to explore the role of technology and social media in encouraging gendered abuse, such as work analyzing the sentiment and hatefulness of the #Gamergate movement, which involved the harassment and defamation of female video game journalists and developers on Twitter (Chen et al., 2022). HCI work has also identified the importance of social media in challenging gendered abuse through activism online. Research focusing on the #MeToo movement showed how #MeToo encouraged Twitter users to talk about their experience with sexual violence and that the network of disclosure reduces stigma and encouraged other twitter users to speak out (Gamage et al., 2022). On the other hand, Moitra and colleagues' work with stakeholders in India, describe the marginalization of intersectional identities in activism, which make engagement with #MeToo challenging particularly for marginalized castes and sexual minorities (Moitra et al., 2021). Moitra et al. specifically state that the "burden of proof solely rests on the survivor to prove that they are not lying or "did not ask"" for their abuse in Indian justice systems.

There is limited work exploring the experiences of victims of deepfake abuse and the various socio-cultural and infrastructural obstacles which they face. Noelle Martin, a deepfake victim-survivor turned activist/scholar, describes two distinct phases of deepfake abuse: the "abuse phase" and the "removal phase" of deepfakes (Marwick, 2017). The lay advice given to victims of image-based sexual abuse (IBSA) often puts responsibility onto the victims; suggesting that they should reduce their public profile online and keep images of themselves off the internet (Karnouskos, 2020; Martin, 2021). We refer to this kind of public advice as "social solutions" in this text as solutions which require victims of abuse to change their own behaviours online. The "strategy" of removing oneself from the internet to prevent deepfake abuse is a form of victim blaming which deprives women (and particularly already marginalized groups such as women of color) equal access to the internet (Marwick, 2017). Social solutions

are also uniquely impractical for celebrities, (particularly internet celebrities) because they further threaten the labour undertaken by women and marginalized groups in the digital creative economy, which is already precarious under platform capitalism (Duquette, 2020). This strategy also emboldens the use of deepfakes in targeted harassment, where certain misogynistic communities online can police public women's behaviour and force them to withdraw from the internet (such as Gamergate; Karnouskos, 2020; Patrick, 2022). Just as "social solutions" position the onus of the work on the victims, technological recourses for deepfakes often require users to self-police illicit content of themselves (Umbach et al., 2024). In particular, Reddit practices towards harmful content have been criticized because, while problematic communities can be banned, individual infractions overly rely on user-level interventions, such as community moderators and user reporting (Gamage et al., 2023). In tandem, existing policies from search engines such as Google Search (Google, 2023) and Microsoft Bing (Microsoft, 2023) put the onus on users (or a representative) to collect screenshots of abuse imagery of themselves in their reports. This process must then be repeated for every search engine provider and does not take any action against the potential for future infractions (which may be increasingly common as the technical skills necessary to create deepfakes have reduced; Kwok & Koh, 2021). While this process is an improvement over the lack of any reporting mechanisms, the emphasis on user-based solutions is part of a wider issue with responsibility on platforms, where certain websites/services which are responsible for deepfakes face limited consequences for hosting illicit content. Instead, and in line with the dominance of victim-blaming discourses in society, they pass responsibility onto the victim to remove this content. The design of these services is a space in which HCI and particularly trauma informed HCI practice is required.

There are also legal recourses and infrastructure for deepfake abuse imagery though, like most technological solutions, they are also limited by focusing victims reporting known abuse and abusers. Intimate imagery generated through deepfake technology, like other forms of online non-consensual intimate images is a technology-facilitated crime (Falduti & Tessaris, 2023). There are various legal approaches that can be taken such as copyright and IP law (Tyagi,

2023), privacy (E. Gerstner, 2020) and defamation (Pascale, 2023) and image rights (Duquette, 2020). As with all image-based sexual abuse, there are various obstacles to victims achieving justice (Wieberneit et al., 2024). Unfortunately, most legal solutions to deepfake abuse imagery are further hamstrung by two specific factors. Firstly, the anonymity of the creator, enabled by existing digital infrastructure impedes prosecution. Secondly, the transnational nature of internet crime (Hall et al., 2023) means that even if perpetrators were identified, they may be operating in an area of the world with different or non-existent deepfake regulations. Once the imagery exists it can be copied, shared and stored across various locations internationally (Thombre, 2021). Existing research on deepfake abuse forums shows that many deepfake perpetrators know this, and while they are concerned at the possibility of being prosecuted for producing illegal content, they believe that the content will always exist as long as they have access to the open-source technology behind the videos (Timmerman et al., 2023).

As part of the potential of HCI research in challenging gendered violence, one must centre the experiences of those who have experienced abuse and how their abuse is impacted by technology (Kim et al., 2022), such as the development of care infrastructures for victims of domestic abuse, to respond to the threats they face to their privacy and security (Tseng et al., 2022). There is limited qualitative work on deepfake abuse which centres the victims of NSII. An interpretative phenomenological analysis (IPA) carried out by Rousay (2023) on both online interviews with deepfake abuse victims and published work of lived experience experts/activists generated three common themes. Firstly, victims reported experiencing gendered narratives around deepfake victimization, feeling denigrated and objectified, and experiencing gendered stigma when reporting the abuse. Secondly, victims reported long-lasting consequences including post-traumatic distress, anxiety and nightmares. They also experienced harassment including doxing, rape threats and ‘slut-shaming’. Finally, victims noted an effect on their identity where their reputation had changed, and they felt like they had “become pornography”. Existing qualitative research outside of HCI focuses on highlighting the experience of victims of deepfake abuse. Building on this, our work seeks to use Critical Discursive Psychology to

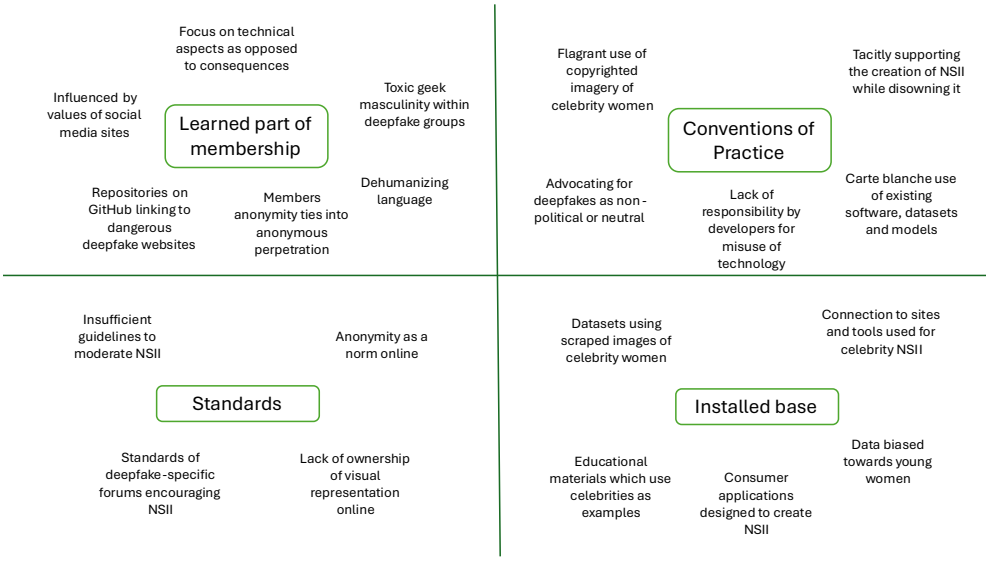
explore more critically how these experiences are constructed in relation to broader social repertoires and how they emerge as a consequence of the infrastructure of deepfake technology, recourses for NSII, and online culture as a whole. Specifically, our work seeks to use a Critical Discursive approach to 1) examine the discourse around how technology infrastructure and usage patterns fuel abuse, and 2) how victim-survivors understand, respond to, interact with, and critique that infrastructural environment when seeking recourse.

5.3.3 The Infrastructural Justice Approach to NSII

In the last twenty years, digital communication technologies have been embedded into varying existing and novel forms of gendered violence and abuse (Powell et al., 2021). This includes the abuse of public women using non-consensual intimate imagery. Feminist scholarship on the early 2000s- mid 2010s culture of sharing illicit and leaked images of female celebrities have highlighted how misogynistic targeting of celebrity women is constantly evolving: “the distaste for the public woman is not new, but what is new is the capacity for and techniques used to humiliate and punish her (p.45)” (Patrick, 2019). Recent work by Robinson et al. (2024) suggests that deepfake technology is a form of infrastructural injustice, where the harms of this technology are created by many people and things operating within established norms and practices that we take-for-granted. We have already documented how deepfake infrastructures place an onus on usees to address harms directed at them. Robinson et al (2024) draw from Susan Leigh Star’s infrastructure, studying how infrastructure is used in the development of deepfake software and the hosting and distribution of deepfake content, makes visible the ways in which companies and coders create such harms. An infrastructural justice approach is also interested in spreading responsibility to those with power and privilege to address infrastructural injustices, moving focus from the Usee, back to the developers of deepfake content and perpetrators of deepfake abuse. Robinson and colleagues focus specifically on four aspects of Star’s infrastructure as applied to deepfake abuse imagery: learned as part of membership on sites such as GitHub and Reddit, links with conventions of

practice such as open-source, embodies existing standards such as anonymity, and is built on an installed base (Figure 5.1).

Figure 5.1: Summary of Robinson's and colleagues (2024) work on the infrastructure behind the creation of deepfake NSII



Firstly, deepfake software emerges from a community of practice (forums) and neutral/misogynistic attitudes towards deepfake targets/content that are learned as part of membership of that community. Secondly, these communities both perpetuate and create conventions of practice: the ethos of open-source software views the software itself as neutral and encourages making it accessible to as many people as possible. Thirdly, it embodies the standards of internet services towards popular but immoral communities, particularly those of illegal content such as piracy, which operate under systems of moderation which are not designed to remove the illegal content until forced to. Finally, and perhaps most pertinently, is the installed base behind deepfakes (the pre-existing structure/s such as trained models, software and datasets). Today some of the most popular open source deepfake technology is linked overtly to websites specializing in the production of NSII targeting celebrity women (Winter & Salter, 2020) and still come with pre-trained celebrity models (often of young women) (McCosker, 2022). These models come with a disclaimer that they do not resemble any real person, but they clearly resemble celebrities with slightly modified pseudonyms. For example, if

your name was Jamie McDonald, it would have a picture of what is clearly your face with the name Amy Mac Ronald underneath. Consumer deepfake apps try to have it both ways, e.g. they outlaw pornographic content but do allow users to swap faces onto the cover page of Playboy magazine (reface, 2024). The developers absolve themselves of responsibility for the content while they tacitly encourage the creation of non-consensual intimate imagery of women.

Though existing literature has charted various avenues for support for the victims of deepfakes such as self-reporting abuse materials, it is apparent that most solutions requiring self-report place part (or all) of the responsibility for reporting and removing content onto the target themselves. A distinct part of being a Usee of a technology is the lack of awareness, yet much of the existing recourses rely on victims being aware that this content exists before having to report it. Robinson et al.'s work (Robinson et al., 2024) on deepfakes and infrastructural justice highlights how power, privilege, interest, and ability affect the development of solutions to the fundamental issues of deepfake abuse. These same principles apply to existing solutions; legal approaches have an interest in reducing the harms of deepfake abuse but lack the power and ability to do so due to infrastructural constraints. Conversely, search engines and internet service providers may have the power and ability to remove content but do not take more proactive solutions (presumably due to lack of interest as a whole, though individuals within these organizations may be actively working to improve recourses). They privilege the abuser/user over the abusee/usee by providing the former with a platform/anonymity and the latter with the experience of having to diligently collect evidence of their own abuse (by emphasizing manual self-report as the main solution offered to victims of NSII). Social solutions also privilege the abuser by encouraging victims to remove themselves online which just facilitates the abuse (by showing that it is a successful strategy to force activists and public figures to self-censor). We do not propose further novel solutions here, but we seek to demonstrate why current recourses are inherently infrastructurally stymied by the values of Usee-centered technology. Studying the experiences of the current victims of deepfake abuse

with existing recourse helps to highlight their limited efficacy and exposes their role in contributing to the overall distressing effects of deepfake NSII.

5.4 Methods

5.4.1 Data collection

To carry out research on the experiences of celebrities with deepfake intimate imagery we took public statements and news interviews as case studies. As celebrities are most commonly the target of deepfake abuse materials we wanted to examine their experiences (Timmerman et al., 2023). As well as this, using public statements as case studies is an appropriate methodology for HCI research on distressing topics as it avoids risks of re-traumatizing interviewees (McKay & Miller, 2021). Furthermore, existing Critical Discourse Analysis literature has suggested the benefits of using a wide variety of existing sources for a corpus, especially from new forms of media communication such as social media and various forms of online content creation (KhosraviNik, 2014). We identified numerous celebrities who had been mentioned as victims of deepfake imagery using a dataset from our previous scoping review studying extracts from the academic literature on harms of deepfakes (Twomey et al., 2024). In April 2024, we carried out a google keyword search using variations of the phrase "{name} speaks out deepfakes" to locate public statements. While the terminology of our search was specifically chosen to avoid abuse websites, pornographic websites occasionally still emerged as suggested search results – further highlighting the extent to which deepfake media is platformed by our information environment. Between April and June 2024, an additional search of online news articles about deepfake NSII was also conducted to identify further victims (i.e. searching Google and X using the phrase "speaks out deepfakes" or "celebrity speaks out deepfakes" to find relevant news articles and quotes). Interviews within larger texts were excluded if they were too short (e.g. a single sentence in a larger text) or heavily editorialized (e.g. where a newspaper article has worked direct quotations directly into the text), to avoid the risk of taking quotes out of context. As expected from qualitative work on pre-existing public

statements, the texts collected were a wide range of lengths and sources. Outside of issues with brevity and editorialization, we wanted to include as large a number of celebrities as possible in this dataset as is best practice to avoid criticisms of arbitrarily selecting text (Sriwimon & Zilli, 2017). The majority of statements were between 200 -1000 words long. We included multiple statements from celebrities in cases which they had made multiple statements.

We analyzed both text and video responses. Audio was transcribed using the transcription tool in Microsoft Word, which was manually refined by the first author. The total size of this corpus was 19,202 words, from nine celebrities across 12 sources (Table 5.1) which is an appropriate sample size as discourse analysis methods acknowledge that large sizes of sample and text would make the analytic task unmanageable as opposed to improving the quality of an analysis (Bondarouk & Ruel, 2004). Regarding demographics, while the chosen interviews reflect that the majority of early targets of deepfakes were celebrity women from English speaking countries (Timmerman et al., 2023) we did attempt to find instances of other groups of celebrities (such as Bollywood actors and K-pop idols) who had spoken out about deepfake abuse imagery but could not find any such public statements at the time of data collection. One of the celebrities in the dataset is non-binary and has expressed that they do not have a specific preferred pronoun (E). There are inherently issues when studying secondary data and attributing demographic information that participants have not consented to self-disclose. Though we are not able to give demographic data we did consciously seek to represent celebrities that come from a broad range of ethnicities, gender identities, social classes and sexuality. While not appropriate for our study of secondary data, we recognize the need for work which situates experiences with deepfake NSII directly within intersecting social and cultural factors and that further considering race, gender and sexuality would offer a complementary lens for such an analysis.

Ethical approval was obtained from our institution’s ethics board. This topic requires sensitivity to the experiences of the individuals involved. While the original data is publicly accessible online, we have chosen to obscure the celebrities in our corpus. The issue of

anonymity/identification split the research team: originally, we intended on including the identity of the public interviewees because we felt that removing their names would be against the spirit of speaking out. However, when analysing the data we found that many expressed concerns about their names/brand being associated with deepfakes in perpetuity.

Table 5.1: Deepfake statements in our corpus

Title	Type of celebrity	Description and word count
A	Streamer, activist	Tweet (203)
B	Streamer	One Vlog (394), one interview (1029) and one podcast (6831)
C	Streamer	Video (1410) and tweet (51) about experience of deepfakes
D	Influencer	One TikTok post (412)
E	Streamer	Video about experience with deepfakes (691)
F	Actress	Text statement for news article (425)
G	Actress	Two extracts from same interview (318)
H	Actress, influencer	One talk show interview (348)
I	Rapper	Tweet (39)

5.4.2 Analysis

This work uses Critical Discursive Psychology as its method of qualitative analysis. Critical Discursive Psychology draws on both discursive psychology, which is interested in discursive practices i.e. what people do with language, and poststructuralist Foucauldian discourse analysis which is interested in "what kind of objects and subjects are constructed through discourses and what kinds of ways-of-being these objects and subjects make available to people." (Willig, 2013). Potter and Wetherell (2004) suggest that combining these approaches is preferable. "As CDA focuses on discourses, which position people along power lines, CDP focuses on interpretative repertoires which 'allow for a wider focus on human agency' and are 'recognizable way[s] of describing, framing or speaking about an issue that is identifiable as such (p.240)" (Locke & Budds, 2020). Outside of interpretative repertoires we also study ideological dilemmas where people construct a struggle with divergent constructions of social relationships and realities (Fairclough, 2013). By further engaging with Critical Discursive Psychology literature (Goodman, 2017) we were able to consider how identity, memory and trauma were framed through the discourse of the corpus.

We chose Critical Discursive Psychology as it draws from Critical Discourse Analysis (Fairclough, 2013) to study the role of discourse in broader social wrongs/injustices – such as those which underly technology facilitated gendered violence towards female celebrities. This allows researchers to take a critical realist approach (Sims-Schouten et al., 2007) in how the constructed discourse links to the larger material concerns of an issue. For us, this meant analysing how discursive techniques give insight into the experiences of deepfake such as framing and positioning (accepting or rejecting one of many subject-positions which fit into broader interpretative repertoires; Potter & Wetherell, 2004). While outside the analytic lens of discourse analysis, we did consider narratives in the text in how the narrative of their abuse/harassment is constructed in terms of larger narratives of technology fuelled gendered violence. We incorporated Baumer's concept of Usees into the analytic lens by specifically considering how non-consent was constructed and framed within interpretative repertoires of other non-consensual acts, how celebrities framed the experience of being targeted through broader lenses of gendered violence, and finally how they constructed awareness and the distress of becoming aware of NSII.

Before carrying out the analysis, it was necessary to familiarize ourselves with current legislation for deepfake intimate imagery. The first author compiled a document outlining the laws around deepfakes in the countries where the celebrities lived, and we drew from comparative literature outlining the variety of current legal, technical and social solutions. This helped contextualize the experiences in the corpus within existing legal and technical infrastructures. We followed Goodman's (Goodman, 2017) guide for conducting a discourse analysis for psychologists - also drawing from Carla Willig's (Willig, 2013) and Locke and Budds (Locke & Budds, 2020) similar approaches. Our analysis first involved a preliminary reading with the primary researcher identifying the genre of the piece as well as the action orientation of language and which discourses are constructed and situated in the text. Next the primary researcher identified how language and discursive techniques such as positioning and framing was used in these discourses (Appendix IV). At this stage, two further authors

expanded and commented on these discourses suggesting new codes or offering alternative interpretations to existing codes. Formalized inter-rater reliability is not appropriate for constructionist methods to qualitative research such as discourse analysis (Potter, 1996). The primary researcher then derived resulting discourses relevant to the research questions iteratively along with feedback from the research team (Appendix V). We include further examples of our qualitative coding process in the supplementary materials.

5.4.3 Positionality

This research aims to avoid adding to the outside commentary and dissecting of celebrities' private lives. None of the research team make any claims of coming from an insider perspective of celebrity women. From our perspective as deepfake researchers, it is apparent that NSII is a problem that permeates every element of the technology. Every form of deepfake software we as researchers have interacted with has in some way ignored the consent of the target. Even the use of the term deepfakes across the literature immortalizes the Reddit user and community which popularized this content and perhaps should be re-considered as the dominant terminology.

5.5 Analysis

Our analysis highlights the various ways in which victims frame the harms of deepfakes and how the infrastructures of social media and the wider internet enable abusers while challenging usees' attempts to protect themselves/others and remove harmful content. First, we present findings relating to the three main factors from Baumer's (Baumer, 2015) concept of Usees (lack of consent, lack of awareness and the experience of being targeted) and how these factors are constructed discursively with the experience of deepfake abuse. We then focus specifically on infrastructural and social factors which affect the experiences of deepfake victims.

5.5.1 How celebrity victims of deepfake abuse construct being Usees of deepfake technology

Our first research question focuses on how celebrities construct deepfake abuse as targeting them without their consent or awareness. Our analysis highlights how this experience is framed as a non-consensual manipulation of identity and how deepfakes are constructed as a form of non-consensual sex work. Repertoires of gendered violence are used to construct the targeted abuse of Usees and how they construct the distress of seeing this abuse imagery online as part of patterns of online harassment and shaming.

5.5.1.1 Experiencing non-consent (deepfakes as manipulation of identity and as non-consensual sex work)

Celebrity victims framed deepfake abuse as theft or non-consensual manipulation of their identity and constructed deepfake identities as non-consensual versions of themselves. They use possessive pronouns to highlight how despite the deepfake not being real, they own the faces in the content. "*Someone put my face onto porn and sold it*" (C). People constructed deepfake content as not only resembling them but as in some ways being them. "*I um was just shocked because this is my face, belongs to me*" (G). The action-orientation of deepfake abuse was framed in terms of stolen/manipulated identity, where deepfakes allow anonymous abusers to control their targets' identity and making it produce non-consensual sexual imagery. "*my face was stolen so men could make me into a sexual object to use for themselves*" (A). This is consistent with the role of non-consensual mimicry inherent in deepfake technology (Akkaya, 2024; Twomey et al., 2024). This quote highlights one of many instances in the corpus where deepfake abuse was framed as a form of non-consensual online sex work. "*Porn of myself I never created has been viewed, distributed and sold. It has been used to threaten, humiliate and attempt to grab power from myself*" (E). E also draws from interpretative repertoires around gendered violence in how they construct deepfakes as non-consensual online sex work. In comparison to consensual online sex work, deepfake abuse is carried out without consent and

distributed without permission (drawing parallels to having leaked videos and pirated content released).

Many statements used interpretative repertoires of stigma and reputational damage to highlight the inter-personal harms of non-consensual sex work online.

"*Imagine that you found a video of you on the Internet where you were stamping on a kitten to death and everyone thought you were a complete asshole and people hated you and your boss was considering firing you and your family don't want anything to do with you and your friends are starting to leave and you don't know if your life is ever going to be the same again and you don't feel safe walking down the street because now loads of people hate you because of something you never did. That's how it feels to be suddenly put in sex work when you never consented*" (C).

This example constructs the various different ways the reputational damage of deepfakes may affect the life of their targets. Multiple statements highlight how the fakeness of deepfake technology may not negate the harms of the technology to their reputation and personal relationships. "*My 60 year old dad if he were to see that video, I would never be able to convince him I didn't do it*" (B). The reputational harms of deepfakes are fuelled by repertoires of the believability of deepfake videos where false videos are framed as not being harmful.

5.5.1.2. The experience of being targeted: repertoires of gendered violence

The positioning of deepfake victims as non-consensual sex workers draws from repertoires around stigma and gendered violence. One statement situates the reputational damage of deepfakes as part of repertoires of sexual entitlement: the idea that expressing sexual autonomy is used as justification for non-consensual actions.

"*The minute that I seem sexually available in any way, shape or form, I'm fair game to humiliate physically assault, rape and attack. And this happens to porn stars. There are many porn stars who get murdered on their doorstep. Because a fan of their work will think that they're easy and that they will just sleep with anyone. So, there's this idea that the minute you*

are consensually sexual, everyone can help themselves to you and I've experienced this and I've seen it, and I've seen it happen to other people" (C).

This quote constructs the harms of gendered violence against sex workers as motivated by interpretative repertoires of sexual entitlement and further constructs the potential dangers of non-consensual sex work. Other celebrities further situated this sexual entitlement in blaming narratives in cases where targets were harassed if they were seen as having "sexualized themselves" online.

"[Famous streamer] *was actually one of the people who was included in the streamers who were deepfaked. There were people coming into her chat saying "look at this photo or that photo of you that you posted". Trying to justify sexualizing her as if she brought it on herself*" (C).

This quote emphasizes how blaming narratives around deepfakes evoke stigmatizing attitudes towards sex, particularly those which tie into the values of sexual entitlement underlying "slut-shaming" tropes [66], which position women as responsible for their abuse.

Sexual entitlement was only one of many interpretative repertoires relating to gendered violence and targeted abuse online. Their statements situated deepfakes within existing forms of abuse and harassment specifically targeting successful women (and particularly women of colour) "*It's really sick how y'all go out of the way to hurt me when you see me winning. Y'all going too far, Fake ass shit. Just know today was your last day playing with me and I mean it*" (I). Here, I position the perpetrators of this content as specifically targeting her because of her success. Others framed this content within the Internet practice of targeting feminist celebrities/activists.

"*My first experience of this, my first knowledge of existing, was as a weapon to humiliate and degrade women. It was* [Notable target of Gamergate] *...and a popular streamer disliked her, so he basically mobilized his entire viewership against her. People started*

photoshopping her face onto pornography, and specifically sex acts that were considered degrading" (C).

This discursive framing constructs deepfake abuse as targeted slander directed towards women who transgress or question societal norms, continuing a historical pattern of reputational damage which aims to undermine high profile women.

5.5.1.3. The traumatic experiences of becoming aware of and watching deepfake abuse content

Victims constructed social media as both providing a venue for targeted harassment and abuse, and for making targets aware of the abuse content. In particular, the malicious direct messaging of abuse content which forces streamers to see abusive imagery of themselves – often after the original case of abuse.

"*I just had my DMs open because I was looking for panel members for my award show. Yeah, so I opened it up for people to send me their stuff or whatever and I was like, huh, I'm just like sitting there, absorbing it and I click on my messages and it's just these videos*" (B).

This example constructs the social harms of deepfake abuse, where outside of the direct creators of the content, social media facilitates the targeted bullying of victims by spreading and directly messaging abuse content. This highlights that while initially existing in communities which they do not interact with, the effects of deepfake abuse content begins to negatively affect the victims' ability to use social media wholesale. The public discourse around deepfake abuse incidents was where some victims became aware of their own imagery.

"*And when you see things like that, it's very hard not to sympathize and the whole situation led me to be a bit curious and be like, "wait am I on these websites?" so I went and googled it and unfortunately I had also been deepfaked by this person and you're about to learn how this is gonna fuck with my life forever*" (C).

This quote constructs the role of publicity and search engines as spreading awareness and serving as the first point of contact with traumatic content and abuse.

For those in the corpus who relayed firsthand experiences of watching the deepfake abuse imagery, the content was framed in terms of various forms of sexual trauma and abuse. B describes the psychological reaction to it as similar to traumatic experience of sexual abuse. "*Well, every woman that has been a part of this has kind of mentioned the same feeling. Like it reminds me of, like I was molested as a child. and it reminds me of that feeling like* (B)". The traumatic experience of deepfakes similarly draws from interpretative repertoires around sexual assault:

"*In 2018, I was inebriated at a party and I was used for a man's sexual gratification without my consent. Today. I have been used by hundreds of men for sexual gratification without my consent. The world calls my 2018 experience rape. The world is debating over the validity of my experience today. This situation makes me feel disgusting, vulnerable, nauseous, and violated -and all of these feelings are far too familiar to me*" (A).

She frames the incident as directly comparable to the feelings of her traumatic experience of rape. It should be noted how the perpetrators are constructed here; while traditional assault often only has one direct perpetrator, deepfake abuse imagery is directly perpetrated by many individuals. Furthermore, the harm is exacerbated by the erasure and debate over the validity of the experiences of its victims

Another distinct element of the experience of encountering deepfake abuse imagery to note is how deepfakes evoke the harm of seeing an image or video of abuse which you do not remember happening to you, especially in cases where the content is more realistic. "*I saw my face like plastered on all these porn images and I was like Jesus I don't remember that, I don't remember that*." "*Now, I can't be sure there's been some hazy no no haha. I'm adventurous- not that adventurous*" (H). This example parallels deepfake abuse with assault under the influence of alcohol. Finding realistic illicit deepfake videos of yourself online was framed as comparable to other forms of abuse imagery, particularly those which the victim does not remember. "*As soon as you see them, your mind shifts like from saying "Ohh. That's not my body why should I care" to – "Was I just? ...Did I do that? Do people have this of me?*" (B). This description of

reconstructing her response to the content echoes commonly reported responses to other forms of gendered abuse and its impact on memory (Hardy et al., 2009): questioning first if the incident really happened and if the video is real.

5.5.2 How do victims of deepfakes negotiate infrastructural and social obstacles against seeking recourse for deepfake abuse?

Our second research question focuses on how infrastructural and social obstacles are negotiated by the victims of deepfake abuse. Firstly, we show how internet platforms are constructed as voicing narratives which blame and silence victims. We then note how deepfake infrastructure and public awareness of deepfakes are framed as further disqualifying the experiences of the abused. We then discuss how users construct the infrastructure of removing deepfake content and how it forces them to be responsible for removing NSII online. Finally, we return to the concept of public awareness and how it creates an ideological dilemma for advocacy.

5.5.2.1. Constructing the internet and the disqualification of deepfake victimhood

One fundamental part of the discourse around deepfakes and the internet was how the victims of deepfakes position themselves both as part of and outside of the internet. When positioning themselves as members of the internet, their existence in these online spaces was framed as justifying abuse as norm within this space, for example, "*yeah "What do you expect? You're part of the Internet*"" (B). The internet here is constructed as a community space where abuse and harassment are assumed as part of membership. They also position themselves as distinct from "the internet" and use speech to construct an "internet voice" which consists of the narratives being used to undermine their experiences:

"*But if you look at a whole bunch of these comments, you'll see that so many people out there don't understand what the problem with this is… "Is there really any difference between this and some random viewer sexualizing them in my mind? I hate to break it to you, but you don't own the visage of your face. Log off. You're not a victim. What's the problem? Why do you*

care? This just comes with the job. Like this is normal. This is just a sexual fantasy. It's not that deep. Why do you care?... And there's a distinct lack of empathy and a lot of people who are genuinely confused as to what's the problem with it. And I'd say consent" (C).

This quote constructs the various ways in which the action orientation of internet narratives undermine/disqualify the experience and emotional response of its victims: where a deepfake is framed as not real by comparing it to an act of imagination; victims frame themselves as having no ownership over their own image; their emotional reaction to the content is questioned; and the experience is framed as a byproduct of working as a streamer on the internet.

Victims often responded to the disqualification of deepfakes as being "not real" by situating the denial of deepfake victimhood within larger dialogues around consent. For example, G states: "*You know, we're having this gigantic conversation about consent, and I don't consent.*" She further rejects the position that labelling deepfakes as fake curtails harms associated with being deepfaked, "*so that's why it's not OK even if it's labelled as this is not actually her*". Victims further reject the disqualification of their experiences of victimhood through emphasizing how deepfake technology has impacted them psychologically. For example, B states: "*I think the really hard thing about it is like the reason I went live crying was because I get that so often, people being like, why does it affect them? I'm like, let me fucking show you*" (B). To challenge the framing of deepfakes as being a consequence of fame, celebrities construct the impact that deepfakes will have on society in general, and particularly on other vulnerable groups. F for example, states:

"*Vulnerable people like women, children and seniors must take extra care to protect their identities and personal content. That will never change no matter how strict Google makes their policies. The Internet is just another place where sex sells, and vulnerable people are preyed upon*" (F).

This example constructs the internet as one of many unsafe places which facilitate the abuse of vulnerable people.

Despite the dominant positioning of deepfake communities who develop, support and spread deepfake content, celebrities also construct support from their online followers and other deepfake victims:

"*I am relieved to finally be talking about this with some support. I was reached out to by 'myimagemychoice'. It's a group of very hardworking activists who are openly pushing for change and becoming a resource for the countless people who have been targeted and abused by others online*" (E).

This quote constructs a counter public of activists and supporters who are positioned as a protective online community space in contrast to deepfake abuse. This community support was especially vital in situations where the veracity of the deepfake video was not obvious and needed to be debunked. *"...so I appreciate everyone in the comments that were saying that, you know, it was fake and um supporting me. You guys are amazing*" (D). Some victims who were able to identify their abuser would reach out to their other victims to help other victims achieve justice: "*...she reached out to me and let me know. And she shared his identity with me. I now know his name, where he lives and where he works. And I am so ready to finally be able to pursue whatever action I can*" (E). Others organized group chats as collective spaces which they justified as emotionally supporting fellow victims. "*I want to fix stuff so I jump into action even though I'm crying, I get all these girls that were as many as I could. And I like dming them asking if they're OK*" (B). This quote highlights how in many cases the organization of communities of victims falls to victims themselves. It also constructs direct messages as a place for positive private interaction with other victims, in contrast to their use by abusers.

5.5.2.2. Financial factors and lack of responsibility in deepfake recourses

The celebrities in our corpus highlighted how the hosting of deepfakes contribute to the length and extent of the harms of NSII:

"*You know what, the weirdest thing is you can't get them down like they're just so, anyone's allowed put them up and they are starting to look more realistic. Most recent ones are pretty realistic and I- I can kind of laugh about it, but at the same time I'm like, I hope that they come down at some stage. I don't want my kids to see them*" (H).

This quote highlights the role that internet platforms (specifically pornographic websites that host NSII) play in hosting and spreading defamatory content long after the original incident. Websites are constructed as both hosting this content and actively encouraging it by presenting users with advertisements of deepfake content. "*I've had deepfake porn pop up before and you see it and you're like, Oh my God. And you exit out and that's it. And you move on*" (B). This quote highlights how exposure to non-consensual deepfake content through pop-up ads framed as a common part of browsing the internet which users attempt to avoid – which has a disproportionate impact when victims of deepfake abuse are being exposed to advertisements for deepfake intimate imagery.

Victims of deepfakes highlighted the extent to which targets are forced to take responsibility for policing their own identity online and the various financial costs of this.

"*Obviously, if a person has more resources, they may employ various forces to build a bigger wall around their digital identity. But nothing can stop someone from cutting and pasting my image or anyone else's onto a different body and making it look as eerily realistic as desired*." (F).

This quote highlights the challenges of protecting a person's online identity from manipulation, and how this is furthered by the limited legal recourses available to victims. One person in the corpus focused on the hypocrisy inherent of platforms which host deepfake content produced by anonymous users (such as Reddit and twitter) do not face accountability for the anonymity which they provide and facilitate.

"*There's a I believe it's a law called like number 203 or something like that and it protects all websites for what gets uploaded to their website. They're not responsible, the users*

are responsible, and you can't track the users because the Internet is anonymous. So yeah, hit a wall" (B).

This quote highlights the dilemma for celebrity targets in the US, where the identifiable hosts of this content are immune from legal action, while they provide anonymity and a platform for abusers (as per Section 230 of US law with some exceptions including content which breaks US federal law or copyright law [14, 28]).

Just as various anonymous individuals/groups online profit from producing and advertising deepfake content, victims also framed deepfake recourses as a profit-motivated industry, where victims have to pay companies to clear harmful content which they struggle to remove themselves. For example, B states: "*I used to pay a company that cleans - websites for photos of you and- I said to him [the perpetrator], I said 'you are paying for every single girl to get clean'... I genuinely don't care because he caused the harm he has to fix it*" (B). The harms of deepfakes evoke repertoires of controlling one's own reputation online and the industry of removing content which often requires victims to pay for it. This example is interesting because the target puts the responsibility for this back onto a perpetrator, which they elaborate on in a subsequent interview. B would later state:

"*The streamer who leaked the website has done a lot of work on the back end……he's really shown that he's trying really hard. He's partnered with a different AI company, actually, as ironic as that is, an AI company that combats, taking these deep fakes down by using AI to find them and to delete them. So, you know, sometimes your worst enemy is your biggest help*" (B).

This example is unique in positioning of AI companies and deepfake perpetrators as both instigators of deepfake abuse and actively taking a role in removing the content, as part of their attempts to publicly apologize for their contribution to deepfake abuse.

5.5.5.3. Everything you can do makes it worse: publicity, advocacy and reputational damage

Engagement-based algorithms and publicity are framed as exacerbating reputational harms and encouraging silence; for example, search engine results privileging news coverage of deepfake abuse over their own content.

"*and the reason I did a lot of the interviews is the same reason that a lot of girls didn't. A lot of girls don't want that to be the first thing you Google. Uhhh, you know, '*[B] *talks about deep fakes'. It does suck after years and years and years of building a career, that that's what comes up. And so, a lot of people opted out of doing interviews and in fact asked me or asked ahh* [C]*, she's another one who's been vocal about it, to avoid that SEO attached to them, which is really sad. But I think it was important to talk about*" (B).

Through the framing of search engine results as something which is attached to victims, B specifically notes the effect of Search Engine Optimization (SEO) in how the victims of deepfake technology find their reputation and public perception affected. "*Fuck the fucking internet. Fuck the constant exploitation and objectification of women it's exhausting. It's exhausting. Fuck* [Male streamer] *for showing it to thousands of people. Fuck the people dming me pictures of myself from that from that website. Fuck you all*" (B). This extract highlights the anthropomorphic construction of the internet as a perpetrator, other internet celebrities as platforming and raising awareness of abuse sites and the use of direct messages of the content from other internet users as a distinct harm.

Publicity and coverage of deepfake abuse were also framed as privileging the experiences of abusers or outsiders over the experiences of the victims. "*The debate over our experience as women in this is, not shockingly, amongst men. None of you should care or listen to what any male steamer's "take" is on how we feel...This is not your debate. Stop acting like it is*" (A). Public apologies by creators and consumers of these materials after being caught with deepfake abuse imagery were framed as especially harmful as they further publicize the incident without any benefit to the target. "*It's just like so. It's so angry, anger inducing cause like the first thing he did was go live and apologize to his stream. He didn't apologize to any of the women, he apologized to his fucking stream*" (B). Deepfake publicity is seen as a way for the

consumers of deepfake, not the targets, to mitigate the damage to their reputation. D encourages other internet users to engage with reporting mechanisms instead of further publicizing abuse content. "*I would really appreciate if anyone else ever see stuff like that to just report it instead of making videos about it.*" (D). This quote frames the difference between actions that reduce visibility of the content and those which publicize it.

5.5.2.4. The ideological dilemma between the good of advocacy and the harms of publicity

Often advocacy was seen as challenging the silencing of other victims. "*I know for a fact that so many others are suffering in silence. People are afraid to speak up, to talk about it, and I am so sorry that this happened to you too. I hope that I can encourage you that you are not alone*" (E). The function of this interview itself is advocacy, and it highlights various ways for victims and others to help enact legal change and to facilitate the advocacy of others. Because the organization and encouragement of advocacy is framed as a challenge to silencing narratives, many victims felt an obligation to speak up for other victims who were self-silencing. "*It's a tough issue and I had a sneaking suspicion maybe other people wouldn't want to talk about it and I feel a responsibility. I hope that we can continue conversations about this and see who it's negatively affecting. And help to change that*" (G). Many advocates continue to justify speaking out despite the harassment and publicity and see being an advocate as a responsibility.

"*I'm willing to talk about this. And other girls aren't. And so I feel like I do need to be the one that's vocal because I'm willing to be that vocal one. So I'm trying to do. It's not only for myself, it's for a lot of other women. And that's I don't, I don't like, want to be vocal, but it feels like my only option. And so that's why I'm doing it*" (B).

This highlights a fundamental ideological dilemma where targets must reluctantly carry out advocacy. Public victims of deepfake abuse imagery are positioned both as victims and as expected to be advocates/activists.

As the first public victims of deepfake technology, they are obligated to take up a role of improving recourses for deepfake abuse. This dilemma is highlighted in the disparity between B's initial reactions to the content "*And the person that made that website. I'm going to fucking sue you, I promise you, with every part of my soul, I'm going to fucking sue you" — In contrast with later discussions— "it's just like, it's so frustrating to have to be like, OK, so now I want to sue the guy. So I've talked to three different lawyers trying to figure out how to sue the guy. They're telling me not worth your time. So now the only thing that I can do is try to create legislation. So now I have to meet with my my state legislator to try to get a bill passed of all. Blah all this stuff*" (B). This demonstrates the inherent power dynamics at play and the protection of website creators within current legislation. Because these celebrities position themselves as the first people to be targeted by this technology, taking an active role in helping develop new laws and adapt existing ones is framed as an obligation. However, speaking out and advocating on this issue opens them up to more reputational damage and harassment.

"*…there's all these media outlets that reached out to all of the girls involved saying, hey, do you want to do an interview? A lot of the girls are like, I don't want to do an interview because I don't want to further the narrative that I'm a part of this. That makes sense, but also without us doing interviews without more political outcry, then you don't get bills, you don't get legislation passed because politicians don't have the motivation, right? So so it's like I'm sitting here and I'm looking at this double-edged sword…*" (B).

This also highlights how public coverage contributes to abuse and silencing which is sadly further justified in a later interview. "*But that blew up in my face too, so I don't know. I'm getting to the point, I think where I just um I just want to be more quiet, is what I'm learning. As sad as that is*" (B). There are two sides to deepfake abuse, the production and dissemination of the content and how the silencing and harassment of public-speaking victims discourages advocacy and justice.

5.6 Discussion

This study analyses how the targets of deepfakes construct their experiences of lack of consent, lack of awareness and being targeted directly by technology. These constructions align with the characteristics of "Usees" as defined by Baumer (Baumer, 2015). Deepfakes are framed by celebrities as undermining consent in various ways, with targets positioned as non-consenting sex workers. Our analysis also highlights how the victims of deepfake abuse position themselves as targeted by abuse operating without their awareness, and they frame becoming aware of this abuse and the further harassment as part of broader repertoires of gendered abuse online. Celebrities construct their abuse as a distinct harm, involving being targeted and silenced by both by the creators and members of deepfake websites and a larger population of anonymous internet users who disseminate existing NSII, harass their victims, and subject them to stigma.

Our research highlights how deepfake abuse is infrastructurally enabled - emblemized by discursive constructions of cultures of harassment/silencing and limit recourses. Social media, the users of streaming websites and misogynistic groups that operate on the internet actively facilitate the creation and distribution of abuse content, as well as exacerbating the harms. The infrastructure is used to challenge victims across every stage of their abuse. Targets of this emerging technology must navigate their position as a victim with the obligation to activism dealing with further abuse. The platforms are constructed as not responsible for user-generated content while giving anonymity to, or even shielding, the creators of said content. Deepfake content removal is an industry; constructing other stakeholders who profit from deepfakes. This work has implications for our understanding of deepfake NSII and potential future work in this area.

5.6.1 Implications for Design, Policy and Practice

Our study highlights how the burden of responsibility to address deepfake NSII is currently placed on victims of NSII abuse through analysing the language they use in public statements. Taking an alternative, infrastructural, approach to understanding responsibility

suggests that those with greatest responsibility for NSII, are those with infrastructural power and privilege (Robinson et al., 2024). In this instance, those with power and privilege who are not enacting their responsibilities are the deepfake content makers, the platforms that host and enable them, and the consumers of deepfake content. Understanding the lived experience of those who are targeted by deepfake technology is important, but it is not up to victims alone to make this system just. Based on our analysis, and to contribute to the critical discussion regarding responsibility towards deepfake content creation and dissemination, we propose several different stakeholders (see Table 5.2) as holding responsibilities for deepfake abuse and highlight potential ways HCI can challenge existing infrastructural harms.

Table 5.2: Various stakeholders identified as holding responsibility for deepfake abuse

Stakeholders	Responsibility for infrastructural justice	Potential for change
Social media platforms	Online celebrities rely on fan interactions in communities- but direct messaging is a venue for bad actors (who may not have the skill to create deepfakes themselves) to engage in targeted abuse.	Various services are now offering default blurring of videos with nudity (such as Google's safesearch and Apple's Sensitive content warning). This practice should be continued along with improved reporting and moderation policies and legal recourses. As well as interventions focusing distinctly on the targeted dissemination of deepfakes.
Search engines	Deepfake NSII is currently platformed on search engines. It is difficult to remove this content.	Deepfake abuse imagery websites should not come up in indexed search results, otherwise publicity on the topic will direct people towards abuse websites. Simply blurring the thumbnail is not sufficient; it helps people who do not want to view deepfakes but continues enabling those who do.
	Search engine optimization and engagement-based algorithms facilitate gendered violence and encourage the reputational damage of deepfake publicity by prioritizing content engagement.	Further research should be carried out on the impact of technology on reputation and how reputational damage encouraged by social media deepens the harms of gender-based violence.
	Generative AI often creates Usees of the technology because it is trained on and creates non-consensual data	Generative AI repeatedly creates Usees, undermining the values of consent and identity. This needs to be a consideration in the ethical creation, application, and regulation of generative AI.
Online and traditional media outlets	For celebrities, the publicity around deepfake abuse is often the moment when they first realize they have been targeted.	Trauma informed practice is vital in the public coverage of deepfakes and how news coverage is presented to people online in search engines and social media.
	Victims may feel forced to speak out about this topic or may not speak at all due to the cumulative effect of publicity and reputational damage.	There needs to be an increased focus on ethical journalism for victims of deepfakes to speak out anonymously and to not have their name publicly attached to abuse if they choose. Along with this, it is vitally important to protect those who do want to speak out from further targeted online abuse.

It is clear from our analysis, that current deepfake recourses encourage targets to remove their presence from social media, as opposed to removing the anonymous perpetrators, whose identities are protected (Martin, 2021), and in this way further excludes victims. This paper does not seek to tackle the various competing epistemologies around issues of anonymity,

freedom and privacy online (which have been well-discussed previously; Akerley, 2020). However, designers should at least consider how we can offer support for targets of deepfake technology to maintain and protect their own public identity in the face of harassment and reputational damage, and to prevent such harassment and reputation damage in the first place. People who are encouraged to engage with social media as part of their profession already need to balance visibility online with the various risks of having their identity mis-represented (Ekinci et al., 2025). Existing literature has highlighted the personal, social and platform vulnerabilities of being public online and the failure of social media companies to adequately protect online users (Duffy et al., 2024). Taking an infrastructural approach, it is clear that managers of social media sites and search engines need to augment current reporting mechanisms and allocate appropriate resources towards the removal and prevention of abusive content. There are numerous ways this could be facilitated: treating AI-generated pornography as pirated content (as the systems behind enforcement and detection of copyright infractions online are much more efficient) (Reis et al., 2017) or as other forms of illegal pornography, (Minnaar, 2023) and hindering its discoverability or transmissibility online; or reconsidering how the names of victims become attached to deepfake content and publicity in search results.

Outside of the implications for deepfake recourses, our work has broader implications for the design of technology that creates Usees, and the how those Usee-centered technologies erode consent. Informed consent of the targets of a technology is fundamental to whether one is designing for Usees or users. The value of consent needs to be implicit in design of deepfake technology, not an afterthought, and not something to be retroactively gathered. In the same way we consider who technology will be used by, we must also ask ourselves who technology may be used on. Otherwise, this same discussion will be had again, and again as new technologies create more non-consenting targets.

To progress the emerging infrastructural understanding of deepfake harms (Robinson et al., 2024), our study highlights the infrastructural ways an industry supports and depends on deepfake abuse. Deepfake marketplaces produce deepfakes for profit; pornographic websites

profit from the distribution of the imagery; internet search providers profit off search traffic; and websites with no direct relationship to deepfake content may host deepfake-related advertisements and thus funnel users towards these harmful sites. Publicity around deepfakes often furthers the profits of all parties involved. The experience recounted by the celebrities in the corpus highlights that recourses against deepfake abuse are repeatedly infrastructurally challenged by the industry of deepfakes. The removal of abuse content for example is a profitable industry in its own right, preying off users' inability to remove content of themselves. Victims of deepfake abuse are being encouraged under the current system to gather abuse evidence themselves, with no guarantee of any protection from future abuse, and this process of removal must be manually repeated for every major search engine and platform (Google, 2023). While it is useful that for-profit reputation management can offer (sufficiently wealthy) celebrities some element of control over their public identity, these services must be repeatedly used (and paid for) while NSII abuse continues unchecked. Infrastructural justice needs to prioritize the equitable treatment of deepfake targets and access to recourse over profit.

5.6.2 Implications for HCI: Reconsidering deepfakes through the lens of trauma informed computing

Existing work in HCI on trauma informed computing (Chen et al., 2022) has highlighted personal trauma as a lens for understanding and designing interactions with technology. Our work details the distressing first person accounts of deepfake NSII across every stage of their abuse and attempts to seek recourse. Drawing on the framework of trauma-informed computing put forward by Chen and colleagues (Chen et al., 2022), we draw from the four Rs of trauma informed care (Jane, 2016) to identify areas in which trauma informed computing may challenge deepfake abuse: realizing trauma exists, recognizing trauma, responding to traumatic content and resisting re-traumatization (see Table 5.3). There are cultural differences in the recourses for deepfake abuse available to victims and the recourses available have changed over time. In many ways our dataset reflected celebrities' response to increasing regulation and support services becoming available. Still, existing work has

highlighted how victims from different countries may not be aware of what support and recourses are available in their country (Umbach et al., 2024) – this is something which also needs to be taken into consideration when designing support services for deepfake NSII.

Table 5.3: Implications of our work for Trauma Informed Computing targeting deepfake NSII

Principle of trauma informed computing	What this means for deepfake NSII
Realizing that trauma exists	It is vital for HCI researchers to understand that deepfake videos can be traumatic. Both to design trauma informed supports services for victims of NSII and to challenge silencing narratives online.
Recognizing signs of trauma	We expand on Baumer's concept of unaware Usees by recognizing the unique distress when one becomes aware they have been targeted by deepfakes. How Usees become aware (social media, news publicity or harassing messages) is a significant point of first contact between Usees and abuse materials
Responding to the content by applying the principles of trauma informed care	Challenging the role of social media in facilitating NSII: Social media fosters abuse by connecting abusers to each other and abuse technologies and serving as a venue to directly harass and defame individuals. There needs to be work on facilitating infrastructure which supports victims of NSII. The positive infrastructure identified include support groups established by victim-survivors and the mobilization of internet users in reporting the content and hindering visibility. Mitigating the trauma of becoming aware you have been deepfaked: There is some evidence that prompt-based interventions on search engines can be successful in directing individuals towards help and support [5]. For deepfake abuse, existing social recourses and content removal options should be prompted at users whose search terms include the word deepfake along with any persons' name.
Resisting re-traumatization	The distress of deepfakes is compounded by three factors which may contribute to re-traumatization: 1. Dealing with the publicity of deepfakes and public debate 2. Increased harassment either by shaming or silencing or sending deepfake videos directly to their targets through DMs. 3. Sharing of old deepfakes after original incidents or creation of new deepfakes

Existing research on trauma informed design highlights how social media has the potential to exacerbate various forms of trauma (Scott et al., 2023). Further research is needed which considers NSII and other forms of technology facilitated gendered violence, e.g., developing communities online for victims, improving moderation online and adapting existence support communities for gendered violence to support harms from synthetic imagery. We also call for an increased focus on how the design of technology and social media has the potential to fuel trauma and hinder victims' ability to respond or find recourse to abuse. The

most significant part of this approach is that HCI researchers need to consider the role of generative AI in image-based sexual abuse, how anonymous networked misogyny can organize around furthering abuse and sharing abuse materials and how direct messaging can serve as a venue for harassment using synthetic abuse imagery. Researchers need to consider the interactions between celebrities and their fans on social media and how the promises of connection to fan communities is also allowing abusers a chance to directly harass celebrities.

There are further implications of trauma-informed approaches to how we carry out research with deepfake abuse survivors in a trauma informed way. Razi and colleagues highlight the importance of trauma informed research in carrying out interviews with teenagers on their experience with social media and the discomfort brought up by discussing harmful experiences with technology (Razi et al., 2024). Some of their recommendations must be considered in light of our own findings when carrying out research which seeks to engage with victims of deepfakes and media editing technology. For example, when anticipating knowledge gaps and anticipating emotional vulnerabilities of research participants, we must account for the possibility of victims becoming aware that deepfakes or other similar edited media may exist of them which our dataset highlights as a particularly distressing experience.

5.6.3 Implications for implementing policy and interventions

Most classifications of deepfake perpetration classify the abuse into three stages involving creation, dissemination and removal (Falduti & Tessaris, 2023). The celebrities in our corpus construct a fourth phase: debate. The final stage of deepfake abuse involves varying social and reputational harms arising from the publicity of and social aftermath of NSII. There are prescribed ethical standards for reporting on sexual assault such as including helplines and not revealing the identity of the target, though these standards are often ignored (Ali et al., 2020). We recommend that publicity that raises awareness of deepfakes should also raise awareness of existing recourses and support. It should advocate for the victims as opposed to giving publicity to the creators of deepfakes and viewers of deepfakes. Like celebrities, activists are a group who are often targeted specifically by deepfake abuse and harassment (Martin,

2021). When celebrities take on the dual role of being celebrities and also public activists against deepfakes they are doubly subject to further targeted harassment by misogynistic internet users/groups. It is important to recognize that the advocacy provided by deepfake victims is essential to challenging the blaming and silencing narratives around the abuse. But public advocacy needs to be a choice over an obligation, especially while advocacy alone cannot challenge deep rooted obstacles towards justice (Moitra et al., 2021).

A major contribution of this work is the need for intervention at every stage of deepfake abuse, targeting various stakeholders. Beyond challenging the creators of deepfake abuse imagery, work needs to directly confront the rationalization and motivations of those who participate in the production and dissemination of deepfake abuse. Existing HCI work has shown the role that technology can play in engaging with the perpetrators of gendered abuse (Bellini et al., 2020). The motivations not only for the production/consumption of deepfakes but the distribution of content and debate of the experiences of women should be challenged by future intervention work. Our research shows for the first time how online culture around the production and accessibility of deepfakes create discourses that falsely minimizes harms to justify current and future abuse (Table 5.4). The existence of these deepfake NSII myths needs to be considered as a variation of existing myths around various forms of gendered violence (Gallagher et al., 2019) which have the potential to lower rates of reporting (Gannon, 2022). The design of perpetrator interventions for NSII myths should draw from existing HCI research into automated and manual interventions for challenging rape myths (Park et al., 2022; Suvarna et al., 2020).

Table 5.4 Deepfake NSII myths from our corpus and how they were challenged

Deepfake NSII myth identified in our corpus	How do celebrities challenge these myths in our corpus
1. Victim-survivors deserve deepfakes because they are public figures/women on the internet	Abuse imagery should not be part of someone's job
2. The content is not real, so it cannot harm	The distress of seeing a deepfake is real and the social harms of the technology are also very real
3. If the content is labelled as false it is not harmful	Being labelled as fake does not rectify the non-consensual and disturbing nature of the imagery
4. Victim-survivors are pretending to be affected by deepfake content	Deepfake victim-survivors are affected by the content in multiple different ways
5. Celebrities who are perceived as producing sexualized content should expect deepfakes	Deepfakes can be made of anybody, this argument is motivated by sexual entitlement
6. Deepfake abuse imagery is just a "joke"	Deepfake abuse imagery has serious personal and social consequences
7. There is no difference between making abuse imagery and having a sexual fantasy about someone else	Sexual fantasies are not distributed on the internet, are not being used to harass celebrities, are not based off training sets designed using celebrity images
8. Deepfake victim-survivors do not have any ownership of their face/image/body	Just because someone's image may not be protected by the law does not mean their image is not theirs

5.7 Limitations

The main limitation of this work is that the sample is limited to nine western celebrities. Our justification for focusing on this sample of victims was that they are in many ways, a "canary in the coal mine" in regard to deepfake abuse. Their experiences may inform how we approach deepfake abuse as a growing problem targeting more and more victims. Just as stolen sex tapes and non-consensual imagery of celebrities online emerged alongside the growing crisis of online IBSA targeted towards the average person. It also been remarked that online culture has also brought with it more comparable experiences of public representation between the average person and celebrity and that the IBSA myths faced by celebrities will eventually target the non-celebrity (Marwick, 2017). Deepfakes are becoming more accessible, the realism and quality offered by publicly available consumer technologies has grown rapidly (Murphy et al., 2025) which may lead to an increase in perpetration. There have been an increasing number of deepfake NSII targeting members of the general population, even towards schoolchildren (Bartolozzi et al., 2025). However, we also take this opportunity to re-affirm the necessity of supporting celebrity women as victims in their own right. While celebrity women in many ways have more social privilege than non-celebrities and reflect on this in their statements, their

experience of reputational damage is particularly harmful as it challenges their livelihood. Nonetheless, it is also important to note that the infrastructures of gendered violence discourage many more marginalized groups from equal participation in online activism (Moitra et al., 2021) and so the individuals included in our sample are likely not representative of all victims of NSII. With the increasing accessibility of deepfake technology, our work serves as a call for future research to take a more holistic view of deepfake targets, exploring how the intersection of publicity, class, gender, sexuality, occupation and race may radically affect experiences of deepfake abuse.

5.8 Conclusion

Victims of deepfake NSII are an example of Usees of technology as they are targeted by a technology without consent or awareness. Our research highlights how high-profile celebrity targets construct being targeted by deepfakes and their interactions with existing legal, social and technological hurdles to finding recourse. As HCI researchers, we should challenge technology which creates Usees. We encourage critical consideration of how the manipulation of identity is baked into existing generative AI models. This work challenges the current infrastructure of deepfakes: the failures of the limited resources available have led to deepfake removal becoming a for-profit industry. As victims of a novel technology, celebrities are often obligated into advocacy which increases the abuse and reputational damage. Future research should take a trauma informed lens, and we provide recommendations on how to report and publicize deepfake abuse while challenging myths about NSII and its victims.

Chapter 6: The Development and Validation of the Deepfake Myth Acceptance Scale (DMAS)

This paper is currently under review at Sex Roles.

Twomey, J., O'Mahony, C., Linehan, C., & Murphy, G. (2025, October 16). The Development and Validation of the Deepfake Myth Acceptance Scale (DMAS). https://doi.org/10.31234/osf.io/6qn7y_v1

Study Motivation

The concept of deepfake myths emerged in the previous chapter as an interesting form of online rhetoric faced by victims of deepfake abuse. I took a close reading of existing literature on both rape myths and IBSA myths to understand how acceptance to abuse myths may indicate future perpetration. While initially designing the perpetration intervention in the next chapter, I decided to measure these myths as an outcome measure. My supervisor Gillian Murphy suggested that the list of questions may be useful as a psychometric measure. While existing work had measured the behaviours of deepfake abuse and attitudes towards the harm, there had been no research which directly tackled victim-blaming and harm minimising rhetoric.

I chose to submit Chapter 6 to a journal which focused more on criminology and gender studies as these are the audience who I think will find the scale useful. A scale for deepfake myth acceptance has a number of benefits. It improves the measurement of the harms of deepfakes and work which aims to profile potential abusers. It also provides a tool for psychological interventions which aim to target NSII -supportive cognition (which is the aim of Chapter 7), and it allows for a clearer understanding of the myths/rhetoric which support technology facilitated gendered violence.

Author Contributions

This study was conceived by JT and GM. The data collection and analysis were carried out by JT under the guidance of CO'M and GM. JT wrote up the paper with guidance from GM, CO'M and CL. All authors approved the final version of the manuscript.

6.1 Abstract

Since their inception, deepfake videos (i.e., AI-generated or modified videos that depict a person saying or doing something they never did) have been primarily used to create non-consensual sexual imagery of women. We argue that false beliefs about the harms of deepfake technology (e.g., the idea that fake videos are not harmful because they are fake) encourage the watching, creation and sharing of deepfake content. We report on the development and validation of a six-item deepfake myths acceptance scale (DMAS) over three studies (n = 1080) which consist of an Exploratory Factor Analysis, a Confirmatory Factor Analysis and a second Confirmatory Factor Analysis, which focused specifically on high-perpetrating populations. The scale has strong psychometric properties (Cronbach's $\alpha = .87$) and moderately correlates with self-reported past behaviours and future intentions to watch, share, and create deepfakes. The scale takes approximately two minutes to complete and is a useful tool for researchers who seek to understand the individuals and communities that produce and share deepfakes. We also encourage researchers to utilise the deepfake myth acceptance measure to assess the effectiveness of interventions at changing the attitudes/beliefs that underpin deepfake abuse.

6.2 Introduction

With the technological and material developments of the 21st century, a wide variety of novel forms of gendered violence have emerged, including various forms of technology-facilitated gendered violence (Henry & Powell, 2015). Technology-facilitated gendered violence most commonly involves distributing explicit and erroneous information about women online (Hedling, 2024). This can include cyberstalking and online sexual harassment (Henry & Powell, 2018) but also Image-Based Sexual Abuse (IBSA), which involves the creation/distribution of non-consensual intimate imagery. There has also been a recent increase in forms of Nonconsensual *Synthetic* Intimate Imagery (NSII) (Umbach et al., 2024), which involves digitally created or altered media. This paper focuses on the most realistic current form of false sexual videos: deepfakes. Deepfake technology is generally used to modify existing pornographic videos and swap in the face of another person. Since their emergence in 2018, deepfakes have become increasingly accessible (Murphy et al., 2025; Twomey et al., 2024) with a number of open-source software and applications dedicated to generating NSII (Robinson et al., 2024; Winter & Salter, 2020). Because of the growing accessibility of deepfake abuse imagery, there is a need for high-quality research to understand perpetrators of this form deepfake IBSA.

In the study of deepfake perpetration, one avenue of research that may have immediate practical implications is in studying the "myths" that people believe about gendered violence (in this case technology-facilitated gendered violence). "Myths" here refer to arguments, attitudes and beliefs that function to minimise the harm of perpetration behaviours and are often visible as victim-blaming rhetoric (Edwards et al., 2011). There is a long history of previous research in gendered violence using rape myth acceptance measures in modelling perpetrator behaviour and designing tailored interventions targeting perpetrators (Lonsway & Fitzgerald, 1994; Persson & Dhingra, 2022). In recent years, novel measures have been developed to assess false attitudes towards non-consensual intimate imagery (Powell et al., 2019). Acceptance of myths that blame victims of IBSA and minimise the harms of the abuse, tends to correlate with self-reported

perpetration (Flynn et al., 2023). However, there are no existing measures which assess attitudes towards *false* AI-generated non-consensual intimate imagery.

The motivation of the present work is to understand the unique myths which minimise the experiences of victim-survivors of NSII and justify abuse and harassment. We consider deepfake myths as harmful victim-blaming rhetoric directed towards victim-survivors and attitudes/beliefs which potentially motivate perpetrator behaviour. We argue that there is a need to develop a method for measuring acceptance of false assumptions about deepfakes for a number of reasons (drawing from Persson & Dhingra (2022):

1) A deepfake myth acceptance measure would provide a measurement tool for the empirical study of the social norms behind deepfake abuse.
2) The measurement of deepfake myths allows researchers to map prevalence of these myths in the general population or in specific groups (potential perpetrators, law enforcement etc).
3) The measurement of myths also furthers intersectional feminist theory by exploring the connection between myth acceptance and various other constructs (race, class, gender etc).
4) Deepfake myths give insight into critical issues around stigma in cases where victim-blaming beliefs are held by victim-survivors (Flynn et al., 2025). Furthermore, it would allow for easier identification of deepfake victim-blaming rhetoric online by moderators or websites.
5) Developing measurement tools for deepfake myths may allow for the subsequent development of interventions to reduce these false beliefs.

We first provide a review of the existing literature around rape myths and myths underlying image-based sexual abuse (IBSA). We then consider how deepfake IBSA may be associated with distinctive myths.

6.2.1 Myths underlying image-based sexual abuse

Victim-survivors of sexual assault often encounter epistemic injustices, i.e., falsehoods that reduce the ability of victim-survivors and broader society to understand and recognise sexual violence (Jenkins, 2017). These falsehoods include rape myths such as blaming victims for assault, absolving perpetrators of guilt, and the broader minimization/ rationalization of women's experience of sexual violence (Edwards et al., 2011). Rape myths are not only rhetoric targeted towards victim-survivors of sexual assault but also social norms and attitudes which encourage sexual assault of women by minimising sexual assault.

When discussing the link between rape myths and perpetration behaviour, it is useful to draw from the Theory of Planned Behaviour (Ajzen, 1991). The Theory of Planned Behaviour advocates that a behaviour is caused by related attitudes and subjective norms (Montano & Kasprzyk, 2015). Rape myths consist of attitudes and beliefs which support the perpetration of gendered violence by a) making perpetration behaviour more acceptable and b) minimising the consequences of the behaviour. To date, a number of studies have shown that rape myths positively correlate with belief in gendered stereotypes, adversarial sexual beliefs and acceptance/perpetration of interpersonal sexual violence (Bohner et al., 2005; Burt, 1980; Trottier et al., 2021; Xue & Lin, 2022).

While initially, the focus of gendered violence literature exclusively focused on rape myths, the advancement of gendered abuse using imagery required empirical work studying the unique myths surrounding novel forms of sexual abuse. Image-based sexual abuse (IBSA) involves the distribution of (or threat to distribute) nude or sexual imagery (often called non-consensual intimate imagery) that were shared privately or recorded without consent (McGlynn et al., 2021). The distribution of the illicit imagery can be between social groups but also can involve social media or pornographic websites. Image-based sexual abuse has increased along with the mass proliferation of image-based websites and the accessibility of digital cameras in smartphones. One in three people between the ages of 18 and 65 in Australia, New Zealand and the UK report having experienced IBSA (Powell et al., 2020). Victim-survivors of IBSA

encounter distinct myths (often in schools, online or among police) which focus on blaming them for sending consensual sexual imagery or by minimising the harm (McGlynn et al., 2021; Rackley et al., 2021), and these myths are most often endorsed and normalised by men (Flynn et al., 2023).

Drawing from existing rape myth measures, Powell and colleagues developed a measure (Sexual Image-based Abuse Myth Acceptance (SIAMA) scale) to measure attitudes towards IBSA (Powell et al., 2019). There were two components to SIAMA, minimizing/excusing the harms (e.g. "*A woman shouldn't get upset if her partner sends nude pics of her to others*") and blaming the victims of IBSA (e.g. "*People should know better than to take nude selfies in the first place, even if they never send them to anyone*"). They found that participants who accepted IBSA myths and blamed victims of IBSA (as measured by the SIAMA) also self-reported higher rates of IBSA perpetration. Further research using the SIAMA found that it predicted self-reported IBSA proclivity above other theoretically meaningful predictors including dark personality traits, gender and sexual orientation. These findings demonstrate that IBSA myth acceptance may contribute to the attitudes/motivations of online perpetrators. Furthermore, the measurement of IBSA myths may benefit work aimed at designing attitudinal interventions to reduce perpetration (Adair & Senn, 2025). However, novel types of IBSA using false or manipulated imagery (such as deepfakes) requires further study and specialised measures.

6.2.2 Deepfakes: Image-based sexual abuse utilising AI

Deepfake abuse consists of the targeted creation and sharing of intimate imagery generated using deepfake technology (Flynn et al., 2025). Deepfakes are a form of synthetic media which use artificial intelligence and deep learning technology to create false audio-visual representations of people undertaking actions or saying words which they may never have said or done (Whittaker et al., 2023). Deepfake videos originated in 2017 on the platform reddit where one user (u/deepfake) took the existing practice of non-consensually editing celebrity women's faces onto pornographic content using photoshop (r/celebfakes) and used deep-

learning tools to similarly edit videos (Twomey et al., 2024). While the original forum was banned from reddit after widespread outrage, there have been numerous alternative websites which have emerged that provide access to deepfake technology and deepfake-generated intimate imagery. Deepfake NSII has remained the dominant use-case of deepfakes, with reports indicating that deepfake NSII makes up over 90% of deepfakes in circulation online (Birrer & Just, 2024). Deepfake abuse is baked into the design of deepfake technology; as well as emerging specifically from a reddit community of practice dedicated towards deepfake abuse materials (Robinson et al., 2024), many of the software repositories of the technology link to infamous deepfake abuse websites including mr.deepfakes (Winter & Salter, 2020). While the mr.deepfakes website was shut down in 2025 (Wise, 2025), deepfake software repositories online today often come with, or link to, pre-trained models of young celebrity women (Twomey, Foley, et al., 2025).

Currently, only a small number of individuals report having watched a deepfake NSII (approximately 7% in the U.S), and the number who report creating it is even lower (1% in the U.S.) (Umbach et al., 2024). Reported exposure is higher in younger and male demographics (Sippy et al., 2024). In existing empirical work on the perpetrators of deepfake NSII, it has been found that greater levels of both the interpersonal and antisocial facets of psychopathy predict desire to watch and share deepfake abuse imagery (Fido et al., 2022). Beyond that, we have no clear understanding of the attitudes and beliefs which underly the perpetration of deepfake abuse, nor do we have any validated tools to measure these beliefs.

Like all forms of gendered violence, the concern with deepfake abuse is not only the acts of abuse themselves but associated shaming, stigma and harassment which harm victims and reduce reporting intentions (Griffin et al., 2022; Reich et al., 2022). Existing work focusing on the experiences of victims of deepfake abuse has highlighted how the experience of deepfake abuse draws on existing stigmatisation and stereotypes which hinder a victim's ability to seek help (Rousay, 2023). In a qualitative study of the public statements of nine celebrities, Twomey and colleagues identified a number of false beliefs associated with deepfake abuse imagery.

These contribute to the social norms around deepfake NSII and may encourage further acts of abuse and harassment towards victim-survivors (Twomey, Foley, et al., 2025) . Outside of the direct harms of seeing deepfake content, celebrity victim-survivors described unique shaming and dismissing narratives (see Table 6.1). Celebrities describe being told that deepfake abuse "comes with the job", encountering doubt as to whether victim-survivors are actually harmed and hearing opinions that deepfakes are okay when they are labelled. Overall, these beliefs tend to minimise the harms/consequences of deepfakes and to blame victim-survivors for their own abuse. We refer to these as deepfake myths in line with previous work (Twomey, Foley, et al., 2025).

Table 6.1: Deepfake NSII myths expressed by celebrity victim-survivors (Twomey, Foley, et al., 2025)

Deepfake NSII myths
Victim-survivors deserve deepfakes because they are public figures/women on the internet
The content is not real, so it cannot harm
If the content is labelled as false it is not harmful
Victim-survivors are pretending to be affected by deepfake content
Celebrities who are perceived as producing sexualized content should expect deepfakes
Deepfake abuse imagery is just a "joke"
There is no difference between making abuse imagery and having a sexual fantasy about someone else
Deepfake victim-survivors do not have any ownership of their face/image/body

At this point, it is necessary to describe briefly how these beliefs can be considered myths. For example, one identified myth is that deepfake videos are not real, so they cannot have "real harms". The belief that deepfake videos are not harmful is a false assumption: deepfakes harm through falsely representing individuals in a manner they may find upsetting (de Ruiter, 2021). Existing qualitative interviews of deepfake abuse victim-survivors highlights that despite the falsity of the content, the harms of false abuse imagery are significant and lead to experiences of PTSD and suicidal thoughts (Rousay, 2023). If the reader considers any other example of deepfakes, such as misinformation or fraud, the harms are so significant *because* it is not real. If a deepfake is made of a male politician making bigoted or untrue claims, the damage comes from the falsity and incongruency between the actions of a real person versus

that of a fake representation. It is only in the case of deepfake abuse imagery that falsity is used as grounds to dismiss the harm caused to the target. The false dichotomy between "real" and "virtual" harms is a hallmark of false attitudes towards technology-fuelled gendered violence (Henry & Powell, 2015). Timmerman and colleagues have previously identified this as one of many distinct justifications for deepfake abuse imagery shared within deepfake community forums (Timmerman et al., 2023).

We assert that the acceptance of false beliefs relating to deepfake abuse may predict perpetration intentions. If a measurement of these deepfake myths is indeed a predictor of behaviour, it may prove a useful tool for the design of interventions which target deepfake abuse perpetration. Existing research has shown that the rhetorical minimising of deepfake harms and blaming of victim-survivors is prevalent in both perpetrators and victim-survivors of deepfake NSII (Flynn et al., 2025). According to the Theory of Planned Behaviour (Ajzen, 1991), while broader factors such as personality may predict behaviour, past behaviour and subjective norms tends to play a more marked role in predicting future behaviour. Both demographic factors and personality factors (such as psychopathy) predict deepfake abuse perpetration (Fido et al., 2022) but there is little work done on how attitudes may also contribute to perpetration. We know that false beliefs relating to sexual assault are directly associated with higher past and future perpetration (Trottier et al., 2021) and that measuring these false beliefs plays a vital role in the design of interventions (Lonsway & Fitzgerald, 1994; Persson & Dhingra, 2022). Understanding and measuring the justification behind perpetration should similarly benefit researchers who seek to study deepfake NSII perpetration. Specifically measuring deepfake myths may provide researchers with a metric for perpetrator attitudes which can be targeted by interventions

The current study

False beliefs around the perpetration of gendered violence legitimise attempts to harm women by minimising the consequences of perpetration (Edwards et al., 2011). Similarly, deepfake myths may legitimise deepfake abuse and victim-blaming rhetoric towards victim-

survivors (Twomey, Foley, et al., 2025). Developing measures for deepfake myths allows us to design perpetrator interventions and may improve our theoretical understanding of the attitudes underlying NSII perpetration. It allows us to better characterise deepfake abuse perpetration in relation to other forms of online and offline gendered violence. The purpose of this work is to create and validate a measure for deepfake abuse myths. We have two research aims. Firstly, to test whether Deepfake Myths Acceptance Scale is reliable and to explore the factor structure of the scale. Secondly, to assess whether our measure of deepfake myths correlates with problematic deepfake behaviours (the creation, sharing and watching of deepfake content) as well as other measures of gendered false beliefs.

6.3 Study 1: Exploratory Factor Analysis

The first stage of our study consisted of an exploratory factor analysis (EFA) on our Deepfake Myth Acceptance Scale to understand the factor structure of the measure. All the datasets in this paper are available at https://osf.io/4h52g/?view_only=18e50c9077af4aaeb9c328a6006c8e7d.

6.3.1 Method

6.3.1.1. Participants

For an EFA, sample sizes of around 300 are considered acceptable and the ratio of participants to variables should also exceed 10:1 (MacCallum et al., 1999). Thus, 280 participants were recruited and compensated using the platform Prolific from Ireland (n = 3), the United Kingdom (n = 143) and the United States (n = 131). The most reported ethnicity in our dataset was White (n = 200), followed by Black (n = 54), Asian (n = 13), Mixed (n = 9) and Other (n =1). The participants were aged between 19 and 81 years old (mean = 42.27, SD = 13.38). Most participants had received a postgraduate degree (n =77), graduated from college (n =129) or completed high school (n = 70). In the dataset, 150 participants identified as men, 128 identified as women and 2 identified as non-binary or a third gender.

6.3.1.2. Deepfake Myth Acceptance Scale (DMAS)

We initially developed the Deepfake Myth Acceptance Scale (DMAS) as a seven-item scale designed to measure online rhetoric which dismiss and invalidate the targets of deepfake abuse imagery. It was developed from qualitative research on the experiences of celebrity women who have been targeted by deepfake abuse (Twomey, Foley, et al., 2025). The scale was derived from the eight examples of myths which were developed from analysing the lived experiences of celebrity victim-survivors. These were lightly modified, drawing from a close reading of existing research on the attitudes and motivations of deepfake perpetrators and non-celebrity victims of deepfakes (Fido et al., 2022; Rousay, 2023; Timmerman et al., 2023). We also drew from best practice in the development of measures of gendered violence, particularly avoiding turns of phrases which may not have cross-cultural validity (Lonsway & Fitzgerald, 1994). As the items are answered on a Likert scale where 1 is Strongly Disagree and 5 is Strongly Agree, the minimum value was 7 and the maximum value was 35.

6.3.1.3. Procedure

Participants were first shown a brief neutral explainer of what deepfakes are ("*Deepfakes are a type of media generated using artificial intelligence, whereby images are combined to create new footage that depicts events, statements or actions that may never have actually happened"*) alongside an example image showing a before and after of a deepfake of the actor Tom Cruise. They were then asked to complete the deepfake myths scale. After the study they were debriefed and given the opportunity to withdraw their consent.

6.3.1.4. Ethics

This research received ethical approval from the departmental ethics committee of (Anonymised for peer review). Informed consent was received from the participants and participants were offered the chance to withdraw their consent after the study was complete. Mindful of research ethics which aim to avoid spreading misinformation when studying

misinformation (Greene et al., 2023; Murphy & Greene, 2023) we debriefed participants carefully, explaining the concept of deepfake myths and how they harm. We also provided links to NSII support services.

6.3.2 Results

6.3.2.1. Descriptive statistics

Overall, participants reported low agreement with our scale items (average agreement = 2.09 on a 1-5 scale). Out of a possible range between 7 and 35, the lowest scoring participant was 7 and the highest was 31. The average test value was 14.64 (SD = 6.31).

6.3.2.2. Exploratory factor analysis

For the exploratory factor analysis, we used the Principal Components method (PCA) in SPSS Version 29.0.1.0. The Kaiser-Meyer-Olkin Measure of Sampling Adequacy was .904 and Bartlett's test of sphericity was significant ($p < 0.01$).

The one factor solution was identified using the scree plot and accounted for 61.58% of the variance in the dataset. The measure demonstrated strong reliability (Cronbach's $\alpha = .89$). There was no need for rotation as our data indicated a one-factor solution. Each myth reported high factor loading (see Table 6.2)

Table 6.2: Means, standard deviation, and component loadings of the initial draft of the DMAS.

Myth	Mean	Std. Dev	Component
When deepfake videos are just jokes, they aren't harmful	2.15	1.17	.812
Deepfake videos are only harmful when someone believes they're real	2.27	1.26	.720
If a deepfake video is clearly labelled as false or fake it isn't harmful	2.51	1.24	.807
It is okay to make deepfakes of celebrities because they are public figures	1.76	0.97	.818
Some people are more deserving of being targeted by deepfakes than others	1.89	1.19	.723
It's okay to make deepfakes when there are no legal consequences	1.93	1.12	.842
Deepfakes aren't harmful if they're not shared online	2.11	1.124	.762

6.4 Study 2: Confirmatory Factor Analyses and Perpetration Behaviours

The second phase of the work sought to employ confirmatory factor analysis to further understand the factor structure and model fit of the measure. During the second phases of the study, we also sought to test both the convergent and criterion validity of our measure by comparing the findings to similar scales of gendered myths and testing the correlation of the measure with deepfake NSII perpetration (the study was preregistered at https://aspredicted.org/5fxt-yzdf.pdf).

6.4.1 Method

6.4.1.1. Participants

As consistent with our pre-registration, 300 participants were recruited and compensated again using the platform Prolific from Ireland (*n* = 2) the United Kingdom (*n* = 185) and the United States (*n* = 100) (demographic information was unavailable for some participants where the given prolific ID number did not match the dataset or the data had expired). We chose this number as it is considered a good sample size in CFA (MacCallum et al., 1999). The most reported ethnicity in our dataset was White (n- 218), followed by Black (*n* = 38), Asian (*n* = 19), Mixed (*n* = 6) and Other (*n* = 6). Ages of participant ranged between 18 and 80 (mean = 43.56, SD = 6.53). 158 participants identified as men and 142 identified as women. Most participants had a postgraduate degree (*n* = 69), graduated from college (*n* = 131) or completed high school (*n* = 92).

6.4.1.2. Materials

The second study used the seven item DMAS but also used additional measures of perpetration and of other false beliefs around gendered abuse.

Acceptance of Modern Myths about Sexual Aggression Scale

We used the AMMSA - Acceptance of Modern Myths about Sexual Aggression Scale (Gerger et al., 2007) which is distinct from other existing measures for rape myths as it measures more subtle and modern myths. Previous measures such as the Illinois Rape Myth Acceptance Scale (IRSMA) (Payne et al., 1999) reflect values which were much more commonplace at the time before its inception and modern rape myths tend to be subtler and covert. The AMMSA is a 30-item measure using a seven-point Likert scale (where 1 = Strongly Disagree and 7 = Strongly Agree) that measures false attitudes around rape and gender stereotypes which serve to justify sexual aggression (e.g. “Many women tend to misinterpret a well-meant gesture as a ‘‘sexual assault’” and “Women like to play coy. This does not mean that they do not want sex”). The possible averaged scores range from 30 to 210. In our sample, the average was 88.8 and the reliability was excellent (Cronbach's Alpha = .97), though a Cronbach's Alpha over .90 may reflect that some items are redundant (Streiner, 2003).

Misogyny Scale

We chose to measure misogyny as a related but distinct false cultural belief that falsely assumes generalised motivations and beliefs on the gender roles and the actions and behaviours of women. These cultural myths include generalisations on the manipulative and exploitative nature of women, distrust of women as a group and devaluation of women (Rottweiler & Gill, 2021).To measure misogyny, we used the ten-item Misogyny Scale (Rottweiler & Gill, 2021) which is measured on a seven-point Likert scales (1 = Strongly Disagree and 7 = Strongly Agree). The average score was 27.69 from possible values ranging from 10 to 70. The Cronbach's Alpha value was high (.96), again demonstrating high reliability but potentially redundant items.

Perpetration behaviours

We created novel measures of self-reported deepfake NSII behaviour, drawing from previous literature (Kugler & Pace, 2021; Umbach et al., 2024). The questions used five-item Likert scales (1 = Strongly Disagree and 5 = Strongly Agree). First, we asked if participants had

intentionally watched, shared or created deepfakes and if they would watch, share or create NSII deepfakes in future. In addition, we posed these questions relating to comedic or political deepfakes, as well as NSII, but these were not considered for analysis (though they are available in the full data set in supplemental materials). We also asked them how likely they were to watch share or create deepfake NSII of four different types of victim-survivors: a friend, an ex-partner, a stranger, or a celebrity.

6.4.1.3. Procedure

Participants completed – in order - the deepfake myths scale, followed by measures of deepfake behaviours. Then participants were given the Misogyny Scale and the Acceptance of Modern Myths about Sexual Aggression Scale in a random order. The ethics procedures were the same as Study 1.

6.4.2 Results

6.4.2.1. Descriptive statistics

Participants again reported low agreement with DMAS items (average agreement = 2.19 on a 1-5 scale). Out of a possible range between 7 and 35, the lowest scoring participant was 7 and the highest was 35. The average test value was 15.30 (SD = 6.03). Participants also reported quite low past perpetration behaviour (watched ($\bar{x}$ = 1.67), shared ($\bar{x}$ = 1.22) or created ($\bar{x}$ = 1.14) and future perpetration intentions to watch ($\bar{x}$ = 1.62), share ($\bar{x}$ = 1.22) or create ($\bar{x}$ = 1.16). Past behaviours and future intentions were significantly correlated (see Table 6.3).

Table 6.3: Correlations between different deepfake NSII behaviours

	Have you intentionally watched NSII before?	Have you intentionally shared NSII before?	Have you intentionally created NSII before?	Would you intentionally watch NSII in future?	Would you intentionally share NSII in future?	Would you intentionally create NSII in future?
Have you intentionally watched NSII before?	-					
Have you intentionally shared NSII before?	.52**	-				
Have you intentionally created NSII before?	.43**	.71**	-			
Would you intentionally watch NSII in future?	.69**	.36**	.39**	-		
Would you intentionally share NSII in future?	.46**	.66**	.74**	.60**	-	
Would you intentionally create NSII in future?	.41**	.63**	.84**	.52**	.76**	-

**. Correlation is significant at the 0.01 level (2-tailed).

In terms of perpetration intentions by victim type, watching intentions were the most common and participants reported greater intentions to commit NSII against celebrity victims versus any other group (see Table 6.4).

Table 6.4: Perpetration intentions for different types of victim-survivors (1 = Strongly Disagree and 5 = Strongly Agree)

	A friend	An ex-partner	A stranger	A celebrity
Watch	1.38 (SD = 0.94)	1.42 (SD = 0.96)	1.56 (SD = 1.06)	1.74 (SD = 1.22)
Share	1.17 (SD = 0.65)	1.19 (SD = 0.67)	1.23 (SD = 0.68)	1.26 (SD = 0.77)
Create	1.17 (SD = 0.69)	1.17 (SD = 0.68)	1.15 (SD = 0.59)	1.19 (SD = 0.67)

It should be noted that the scores for these measures across the dataset are very low. These floor effects are consistent with other work on reporting of illegal (Conrad et al., 2010) and immoral behaviours (Tang & Tang, 2010) and is perhaps exacerbated by the general lack of awareness of and competency with deepfake technology in the general population (Wasi et al., 2025).

6.4.2.2. Confirmatory Factor Analysis

The nature of gendered violence as a research area poses challenges to the use of Confirmatory Factor Analysis to validate psychometric measures. Namely that the data often violates the assumptions of multivariate normality (Yang & Liang, 2013). Specifically, a major concern of existing measures of gendered violence and misogyny is the likelihood of severely non-normal data (Geldenhuis, 2023; Mou et al., 2025; Rottweiler & Gill, 2021). That non-normality may be further exacerbated when we are studying a novel technology which already tends to independently provoke strong responses of scepticism and concern (Sippy et al., 2024). Our dataset evidenced extreme levels of kurtosis and skew according to Mardia's measure of multivariate normality (skewness = 525.85, $p < .001$; kurtosis = 13.55, $p < .001$).

For the confirmatory factor analysis, we used a diagonal weighted least squares estimator (DWLS), as the data was ordinal (five item Likert responses). We followed established procedure for using DWLS estimators on ordinal data using Lavaan in behavioural

sciences (Rogers, 2024). We used Lavaan version 0.6-18 and also used Jasp (which also uses Lavaan) for the tests of normality and for graphing. We chose to use the DWLS estimator over the usually more robust WLSMV estimator as previous literature has indicated that WLSMV performs poorly in severely non-normal data with five or more categories (answer options) per variable, especially with smaller samples of N = 200 (DiStefano & Morgan, 2014). That work also showed that WLSMV was more likely in these conditions to produce higher than normal RMSEA/Chi Squared, which is consistent with what our data was producing under that estimator (DMAS : $\chi^2 = 101.29$, $df = 14$, $p < .001$, RMSEA = .14, CFI = .97, SRMR =.05; six-item DMAS : $\chi^2 = 38.17$, $df = 9$, $p < .001$, RMSEA = .10, CFI = .99, SRMR = .03). For our fit statistics we used existing guidelines to determine the appropriate values for good fit (Hu & Bentler, 1999).

Using the DWLS model, our initial Chi-squared test was significant ($\chi^2 = 42.77$, $df = 14$, $p < .001$). However, other measures of fit indicated decent model fit (RMSEA = 0.083, CFI = .995, SRMR = .045). At this point we cut one question from our scale ("*If a deepfake video is clearly labelled as false or fake it isn't harmful"*). The question was interacting with question 2 which was similarly worded ("*Deepfake videos are only harmful when someone believes they're real"*). Furthermore, modification indices showed that it's removal would improve the overall performance of the scale. After removing the item, the dataset was still significantly non-normal (Skewness = 445.88, $p < .001$; Kurtosis = 13.02, $p < .001$). Our final results demonstrated good model fit. The chi-squared statistic was not significant ($\chi^2 = 14.42$, $df = 9$, $p = 0.108$). Other measures of fit also fell within acceptable ranges (RMSEA = 0.045, CFI = .999, SRMR = .031).

Returning to the EFA dataset from Study 1, the final six-item scale had excellent reliability (Cronbach's $\alpha = .87$). And a one factor solution was still appropriate based on the scree plot and accounted for 54.74% of variance in the dataset. The myths still reported high factor loadings (see Table 6.5)

Table 6.5: Component Matrix of the final six item scale

Myth	Mean	Std. Dev	Component
When deepfake videos are just jokes, they aren't harmful	2.15	1.17	.810
Deepfake videos are only harmful when someone believes they're real	2.27	1.26	.710
It is okay to make deepfakes of celebrities because they are public figures	1.76	0.97	.833
Some people are more deserving of being targeted by deepfakes than others	1.89	1.19	.756
It's okay to make deepfakes when there are no legal consequences	1.93	1.12	.842
Deepfakes aren't harmful if they're not shared online	2.11	1.12	.767

Confirmatory factor analysis on perpetrating and non-perpetrating populations

In an exploratory analysis, we decided to separate the perpetrating and non-perpetrating samples to reduce the noise in the dataset. Our justification for analysing these two groups was that the distribution of responses for perpetrators would be less skewed than non-perpetrators. We expected a small number of perpetrators to have a higher agreement with deepfake myths and for their responses to more closely resemble normal distribution. To generate these two samples, we separated participants based on the question "Have you intentionally watched pornographic deepfakes before?" (this question was considered the most likely to include a usable number of participants versus more obscure perpetration behaviours such as creation). Participants who scored possibly/definitely yes (n = 43) and possibly/definitely no (n = 242) were separated and those who reported "unsure" (n =15) were excluded.

We found that separating perpetrating and non-perpetrating populations improved the model fit for both groups. For the group of non-watchers n = 242 (χ^2 = 10.19, df = 9, p = .335, RMSEA = .023, CFI = 1.00, SRMR = .033. Mardia's test of multivariate normality indicated that for non-watchers, our data was still significantly non-normal (Skewness = 461.58, p < .001;

Kurtosis = 16, $p < .001$. For the group of watchers $n = 43$ ($\chi^2 = 5.31$, $df = 9$, $p = .806$, RMSEA = .000, CFI = 1.00, SRMR = .058). Using Mardia's test of multivariate normality for watchers indicates that the data was normally distributed (Skewness = 73.29, $p = .06$; Kurtosis = -.55, $p = .583$). While the sample size was too small to draw significant conclusions from, it motivated our third study.

Second-Order Confirmatory Factor Analysis

Based on our analyses, we wanted to further consider the structure of our scale, so we decided to take a second-order CFA approach informed by a close reading of the data. We identified that the false beliefs in our items consisted of one of three major ideas:

1. Deepfake media is false, so it is not harmful (questions one and two)
2. Some people deserve deepfakes (questions three and four)
3. Deepfake media has no consequences (questions five and six)

These three categories serve as first order factors, under the second order factor: deepfake myths (Figure 6.1). Using the DWLS estimator, our chi squared was not significant ($\chi^2 = 4.243$, $df = 6$, $p = .644$). Similarly, we had strong values for the other measures of fit (RMSEA = .000, CFI = 1, SRMR = .017). These fit measures indicate a much better model fit and shows that the second order factor solution is appropriate in the general population.

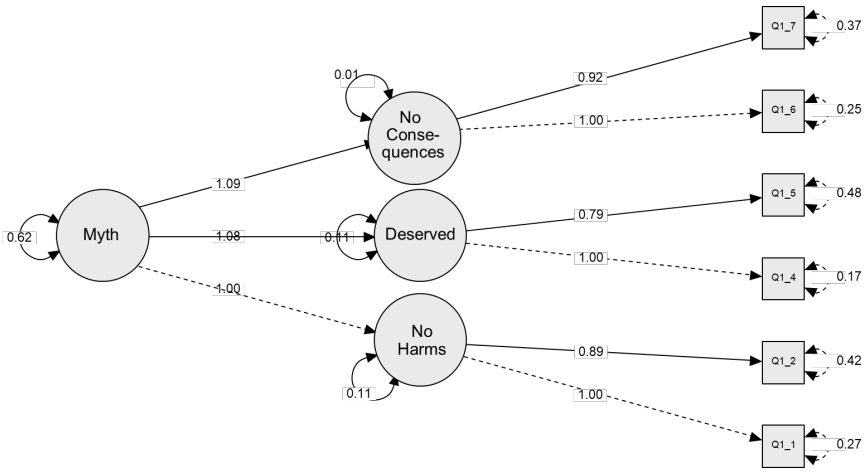


Figure 6.1: Second order factor analysis where the first order factors are deepfakes: i) having no consequences, ii) being deserved and iii) not being harmful. Deepfake myths are the second order factor.

6.4.2.3. Cross validation

The 6-item DMAS had a significant medium positive correlation to both the AMMSA (r = .454, p < .001) and the Misogyny Scale (r = .419, p < .001) which indicates convergent validity. Misogyny and rape myth acceptance had a very strong correlation (r = .823, p < .001) which indicated that these concepts are more closely related. In terms of criterion validity, we found that the measure of deepfake myths moderately correlated with every one of the measured deepfake NSII behaviours (Table 6.6) and was a stronger predictor of each one of the measured behaviours than either AMMSA or Misogyny.

Table 6.6: Pearson correlations of deepfake myths (DMAS), Acceptance of Modern Myths about Sexual Aggression Scale (AMMSA), and Misogyny with deepfake NSII intentions by victim type.

	DMAS	AMMSA	Misogyny
Watch NSII of Friend	.41**	.37**	.37**
Watch NSII of ex	.37**	.37**	.33**
Watch NSII of Stranger	.44**	.41**	.39**
Watch NSII of Celebrity	.46**	.38**	.34**
Share NSII of Friend	.44**	.37**	.33**
Share NSII of ex	.46**	.40**	.38**
Share NSII of Stranger	.49**	.41**	.42**
Share NSII of Celebrity	.50**	.33**	.35**
Make NSII of Friend	.42**	.34**	.30**
Make NSII of ex	.42**	.33**	.28**
Make NSII of Stranger	.42**	.33**	.32**
Make NSII of Celebrity	.43**	.32**	.32**

**. Correlation is significant at the 0.01 level (2-tailed).

6.5 Study 3: Confirmatory factor analysis in a higher perpetrating sample

Based on our exploratory analysis in Study 2, we decided to specifically study a higher-perpetrating population. We wanted to increase the number of perpetrators to reduce the extent to which our data violates the assumption of multivariate normality required by a CFA. Focusing on perpetrating populations was also appropriate in terms of the intended target for the DMAS – deepfake perpetrators. Study 3 was preregistered at https://aspredicted.org/vgqg-5f3d.pdf.

6.5.1 Methods

6.5.1.1. Participants

For the high-perpetrating sample, we recruited men aged between 18 - 45 who had expressed that they had watched online pornography within the last 6 months. The age profile was defined by existing literature and the age profile of perpetrators in the previous studies. As consistent with our pre-registration, 500 participants were recruited and compensated again using the platform Prolific from Ireland (*n* = 2) the United Kingdom (*n* = 270) and the United States (*n* = 193) (this information was again unavailable for some of the Prolific demographics dataset). The most reported ethnicity in our dataset was White (*n* = 334), followed by Black (*n* = 64), Mixed (*n* = 29), Asian (*n* = 27) and Other (*n* = 10). Ages of participant ranged between 18 and 80 (mean = 34.46, SD = 6.53). 499 participants identified as men and one self-identified as transmasculine. Generally, participants had completed a postgraduate degree (*n* = 106), graduated from college (*n* = 247) or completed high school (*n* = 131). 110 participants reported either probably or definitely watching deepfake NSII before. Study 3 participants reported a rate of 22% perpetration, which is higher than the 14% in the general population recruited to Study 2.

6.5.1.2. Measures

The same measures were used as Study 2, including the six-item version of the DMAS.

6.5.1.3. Procedure

The procedure followed the same structure as Study 2.

6.5.1.4. Ethics

The ethical procedures for the project remained the same.

6.5.2 Results

6.5.2.1. Descriptive statistics

For the higher-perpetrating sample, participants still reported low agreement with DMAS items (average agreement = 2.17 on a 1-5 scale). For the 6-item scale, the range was between 6 and 30 and the lowest scoring participant was 6 and the highest scored 30. The average test value was 13.03 (SD = 5.22). Participants again reported quite low in past perpetration behaviour (watched ($\bar{x}$ = 2.00), shared ($\bar{x}$ = 1.18) or created ($\bar{x}$ = 1.16) and future perpetration intentions (watch ($\bar{x}$ = 2.15), share ($\bar{x}$ = 1.26) or create ($\bar{x}$ = 1.23) (full data is available on the OSF page).

6.5.2.2. Confirmatory factor analysis

While the data was still significantly non-normal, it was less extreme than the previous iterations (Skewness = 6.59, $p < .001$; Kurtosis = 19.89, $p < .001$. The Chi square results were significant ($\chi^2 = 14.90$, $df = 9$, $p < .094$). The other measures of fit also indicated good fit (RMSEA = .04, CFI = .99, SRMR =.03). For the 110 members of the dataset who directly reported perpetration we had similarly positive results. The chi squared was significant ($\chi^2 = 12.27$, $df = 9$, $p < .198$). Other measures of fit demonstrated good fit (RMSEA = .06, CFI = .998, SRMR =.05). Again, the data was still significantly non-normal but less extreme than previous iterations (Skewness = 5.67, $p < .001$; Kurtosis = 3.60, $p < .001$. We also attempted to run a second-order CFA on the data, but the model was unable to converge. Lavaan reported that some estimated lv variances are negative, assumedly because of high correlations between the questions. These results potentially indicate poor model fit for the second-order CFA in higher-perpetrating populations.

6.5.2.3. Cross validation

For high perpetrating populations, the DMAS still significantly correlated with AMMSA ($r = .400$, $p < .001$) and the Misogyny Scale ($r = .332$, $p < .001$) and the AMMSA and the Misogyny Scale correlated strongly with each other ($r = .820$, $p < .001$). There were

significant correlations between deepfake myths and past NSII watching ($r = .258$, $p < .001$), sharing ($r = .329$, $p < .001$) and creation ($r = .309$, $p < .001$). There were also significant correlations between deepfake myths and future NSII watching ($r = .331$, $p < .001$), sharing ($r = 366$, $p < .001$) and creation ($r = .365$, $p < .001$). We also found positive correlations when comparing future intention by victim type (Table 6.7), where again the DMAS was generally more predictive of perpetration behaviours than either the AMMSA or the Misogyny scale.

Table 6.7: Pearson correlations between DMAS scores (as well as the AMMSA and misogyny scales) and intended perpetration behaviours in a higher-perpetration sample

NSII Behaviour	DMAS	AMMSA	Misogyny
Watch NSII of Friend	.32**	.32**	.26**
Watch NSII of ex	.31**	.34**	.27**
Watch NSII of Stranger	.32**	.31**	.25**
Watch NSII of Celebrity	.30**	.26**	.21**
Share NSII of Friend	.38**	.27**	.22**
Share NSII of ex	.31**	.33**	.26**
Share NSII of Stranger	.34**	.34**	.27**
Share NSII of Celebrity	.32**	.34**	.25**
Make NSII of Friend	.37**	.23**	.17**
Make NSII of ex	.38**	.26**	.21**
Make NSII of Stranger	.36**	.25**	.20**
Make NSII of Celebrity	.35**	.25**	.19**

**. Correlation is significant at the 0.01 level (2-tailed).

6.6 Discussion

As novel forms of online gendered violence emerge, so too have novel myths which seek to minimise the harms and blame victim-survivors. Acceptance of myths around technology-facilitated gendered violence such as Image Based Sexual Abuse (IBSA) impact how victim-survivors understand their experience and serve to delegitimise/minimise the offense in the public sphere (Powell et al., 2019). Deepfake abuse myths may similarly undermine the experience of victims and minimise the harms of deepfake Nonconsensual Synthetic Intimate Imagery (NSII). Outside of the damage of these myths as rhetoric, acceptance of myths may also demonstrate beliefs which encourage perpetration, in line with

the Theory of Planned Behaviour (Ajzen, 1991). The current study demonstrates that the six-item Deepfake Myths Acceptance Scale (DMAS) is psychometrically robust and is moderately correlated with expected related attitudes such as misogyny (Rottweiler & Gill, 2021) and modern myths around sexual aggression (Gerger et al., 2007). The relationship between the DMAS and other gendered false beliefs demonstrates a broader set of social norms which seek to justify gendered violence. Importantly, the DMAS also correlates with self-reported past behaviours and future behavioural intentions (specifically regarding the watching, sharing, and creation of deepfake abuse imagery). The current study also found that the correlation between the DMAS and behavioural intentions was consistent across different groups of targets (e.g. ex-partners, celebrities). These findings suggest the DMAS is an appropriate measure of deepfake myths that underpin deepfake perpetration across a variety of contexts.

An effective measure of deepfake abuse myths offers similar contributions to research and practice as measures of rape myths: contribution to theory, improved understanding of victim-blaming stigma and a tool for future research and interventions (Persson & Dhingra, 2022). The majority of deepfakes online are NSII (estimates suggest approximately 95%; Ajder et al., 2019) and the DMAS contributes to work which aims to challenge or reduce the increasing amount of NSII on the internet. Namely, the DMAS contributes to our theoretical understanding of deepfake perpetration by demonstrating that deepfake NSII myths correlate with self-reported perpetration behaviours/intention. We consider the correlation between myth acceptance and behaviour in line with the Theory of Planned Behaviour (Ajzen, 1991): deepfake myth acceptance is an attitude towards perpetration behaviour which may partially motivate deepfake abuse. Of course, there are a number of other potential factors (including personality factors like psychopathy and demographic factors like age/gender) which correlate with perpetration, which should be considered along with deepfake myths for future work assessing perpetration (Fido et al., 2022). The DMAS allows for future work which aims to reduce the actions and beliefs associated with deepfake NSII and also opens up the potential for work exploring how these myths and beliefs change within different perpetrator groups. Current

deepfake interventions mainly focus on the use of the technology in disinformation campaigns, often trying to teach individuals to identify synthetic videos (Twomey, O'Mahony, et al., 2025a). The DMAS opens up the possibility for interventions which seek to test if reducing deepfake myths also reduces NSII perpetration.

The profile of victims is a larger point raised by our work. At present, research shows clearly that public women make up the overwhelming amount of deepfake targets and attitudes towards celebrities and politicians are used to minimise the harms of deepfake abuse imagery (Birrer & Just, 2024; Fido et al., 2022). Across our results, self-reported past and future deepfake abuse was more likely for celebrity women than any other victim group. While our work draws mainly from previous research on the experiences of celebrity women (Twomey, Foley, et al., 2025), we have made an effort to write the DMAS myths in a sufficiently broad manner so as to not exclude other victims, such as children or non-celebrity women. As deepfake technology continues to evolve, further study of different kinds of perpetrators and victim-survivors is also warranted. Another important point to consider is the prevalence of deepfake myth acceptance among victim-survivors themselves. For IBSA, victims tend to report higher rates of myth acceptance as part of internalised stigma (Flynn et al., 2023). While we did not measure victimisation in this study, it would be worth measuring the prevalence of deepfake myth acceptance among victim-survivors in future.

As with all measures of gendered violence myths, there is a need to be aware of the various demographic, generational and sociocultural differences which contribute to acceptance of these myths (Persson & Dhingra, 2022). Future work should consider other marginalised groups who may experience distinct variations of deepfake myths. For example, LGBTQ+ individuals report higher rates both of perpetration and targeting by deepfakes (Flynn et al., 2022). The deepfake myths for this group may be informed by the various other myths and stereotypes they experience and deepfake myth acceptance may vary in complex ways among different LGBTQ+ identities. Deepfakes fulfil a dangerous purpose in allowing online abusers to create abuse content which may seek to erase or falsely represent a victim's sexuality and

gender. At least one non-binary celebrity has already spoken out about their experiences with deepfake NSII (Twomey, Foley, et al., 2025). Studying how deepfake abuse towards and within the LGBTQ+ community relates to existing literature on intimate imagery abuse is a vital area of future research.

There are some methodological limitations to our work, some of which apply to other psychometric measures of misogynistic attitudes and behaviours. As previously mentioned, the issues with non-normality of data in measures of extreme and novel acts of technology facilitated gendered violence has been extensively considered in the creation of this scale, especially in how this non-normality challenges the general practice of Confirmatory Factor Analysis. The challenges are further exacerbated by broader scepticism and hostility towards AI-generated and otherwise fake media (Sippy et al., 2024) which we expect contributed to floor effects within a mostly non-perpetrating dataset. However, we have gathered evidence that our scale is effective in higher-perpetrating samples and as deepfake technology becomes more widely known, this can be tested further. Another potential limitation to the present work is the risk of response bias (Schwartz, 2000). Participants may report lower deepfake perpetration as it is seen as an immoral/ illegal behaviour. Researchers in this area must also be concerned with ensuring participants understand what deepfakes are, as there is widespread definitional confusion (Twomey et al., 2024) and approximately 12% of people have never heard of a deepfake before (Sippy et al., 2024). Thus, we recommend that our scale is not deployed without first providing a neutral definition of deepfake technology, as in the current study.

In conclusion, the current study presents a novel and urgently needed psychometrically robust measure of belief in myths about deepfakes. The harms of deepfake myths are in how they may potentially legitimise deepfake perpetration (Twomey, Foley, et al., 2025) and are experienced by victim-survivors as part of victim-blaming narratives (Rousay, 2023). The DMAS demonstrates good reliability and correlates with future intentions to commit deepfake abuse. Future intervention work which seeks to reduce these myths may not only reduce direct

perpetration, in line with the Theory of Planned Behaviour (Ajzen, 1991), but may also reduce social stigma and harassment that block victims of deepfake NSII efforts to receive support.

Chapter 7: Designing an Intervention to Reduce Attitudes and Behaviours Related to Deepfake Non-Consensual Intimate Imagery

Study Motivation

The final study of the thesis aimed to make a positive contribution to the study of deepfake NSII perpetration through the design and implementation of a psychoeducational intervention. The intervention is modelled after existing interventions for gendered violence and aimed to reduce the previously identified deepfake myths and self-reported perpetration behaviours/intentions. I was motivated to carry out this work by the lack of education on deepfake abuse imagery and over-focus on poorly implemented technological solutions. I felt that by using an intervention to educate people about the harms of this technology, we may be able to reduce deepfake perpetration at the source.

As someone from a background in cognitive psychology and misinformation research, I am interested in the design and implementation of interventions. Specifically, I am aware of some critical concerns of interventions and if they actually can successfully influence attitudes and behaviours over a long period of time (as outlined in Chapter 4). Across interventions for gendered violence (both technology-facilitated gendered violence and more broadly), there is a lack of critical work evaluating if and how psychoeducational interventions actually work over long periods of time. I felt that as a psychologist, I could make a significant contribution developing a psychoeducational intervention modelled off of theories of behavioural and attitudinal change and critically evaluating how the effects of the intervention are maintained over time.

Author Contributions

This study was conceived by JT, DC, CL and GM. The data collection and analysis were carried by JT, EL and ALG under the guidance of GM. JT wrote up the paper with guidance from GM and CL. All authors approved the final version of the manuscript

7.1 Abstract

Deepfake abuse involves the use of generative AI to create non-consensual synthetic intimate imagery (NSII) consisting of pornographic videos with someone's identity swapped in without their consent. Originally deepfake videos were used as a way to harass and defame celebrity women and activists but there is a growing number of cases of deepfakes targeting ordinary people. While there are existing legal and technological efforts to mitigate perpetration, there are no evidence-based interventions aimed at preventing perpetration behaviour. We designed a short-form online intervention for deepfake NSII. The Deepfake/Real Harms intervention consists of three vignettes and takes approximately 10 minutes to complete, with the aim of increasing empathy towards victims and educating participants about deepfake myths. Over three pre-registered experimental studies (N = 1628), we provide evidence of the efficacy and acceptability of the intervention, including tests in high-perpetrating populations. The intervention lowered belief in myths about deepfakes (e.g. that they are not harmful because they are not real) and partly reduced intentions to perpetrate (e.g. to watch, share, or create deepfake NSII) with the effects on watching behavioural intention still evident at follow-ups ranging from one week to one month. We have made the current intervention freely available online and recommend follow-up work testing adapted versions of the resource with other vulnerable groups such as schoolchildren.

7.2 Introduction

Deepfake technology involves the use of AI to edit or generate false media (usually videos) which represent someone doing or saying something which may never have happened. While there has been a general focus on the use of deepfakes in AI-generated disinformation, the primary use case is the generation of non-consensual synthetic intimate imagery (NSII) (Umbach et al., 2024). Deepfake NSII usually involve the manipulation of existing pornographic material to replace the faces in a video with those of non-consenting individuals. Deepfake NSII is an increasing problem. Between the first news coverage of deepfake NSII in December 2018 until June 2020, the number of deepfakes online doubled every six months (Kugler & Pace, 2021). Since then, the prevalence of both deepfake abuse imagery and deepfake research have continued to grow exponentially (Acim et al., 2025; Domenteanu et al., 2024). Already, 14% of British, Australian and New Zealand adults report having been a victim of some form of NSII (Flynn et al., 2025). Ordinary people are increasingly encountering deepfake abuse imagery of themselves or others online, whether they want to see the material or not. The harms of deepfake NSII extends to children: 10% of teenage boys and 2% of teenage girls in the UK have encountered deepfake NSII online (Internet Matters, 2024). In terms of perpetration, 8% of British, Australian and New Zealand adults reported that they have either created, distributed or threatened to distribute NSII (Flynn et al., 2022). The prevalence of deepfake NSII perpetration is a global concern, though currently the highest self-reported rates of watching celebrity NSII (6-8%) tend to occur in the USA, Mexico, Australia and South Korea (Umbach et al., 2024). However, there is a lack of empirical research which focuses specifically on deepfake NSII and this is an emerging field of research (Ching, Twomey, Aylett, et al., 2025).

There are a number of current legal and technological approaches to stopping NSII perpetration. There have been a growing number of laws enacted specifically to criminalise deepfake NSII perpetration, but these laws are often hindered by a lack of enforcement due to the anonymous and transnational nature of perpetration (Armesto-Larson, 2020; Ma'arif et al., 2025). Technological solutions often tend to rely on a variety of reporting modalities where

social media users can report NSII to moderators. However, these are extremely inconsistent. In an audit of X's NSII content removal policies, it was found that reporting AI-generated nudity under NSII guidelines had a zero percent success rate over three weeks (Qiwei et al., 2024). While better than having no recourse, these solutions often put unnecessary work on the victims of NSII, with little guarantees of any justice or protection from future abuse (Martin, 2021; Twomey, Foley, et al., 2025). The limited number of technical interventions which are not focused on reporting modalities relies on the usage of image forensics and hashing to detect and track NSII content online (Falduti & Tessaris, 2023). Academics have argued that outside of technical and legal solutions aimed towards victim-survivors there is a distinct need for public education on the harms of deepfake abuse imagery with the expressed aim of reducing perpetration (Ma'arif et al., 2025). Both technical and legal solutions tend to be post-factum (i.e. after the event has occurred) (Falduti & Tessaris, 2023): these solutions have to be supplemented by educational policy and interventions which seek to reduce people's desire to perpetrate pre-factum (before the event has occurred) (Rizzica, 2021). In this paper we argue that simply raising general awareness of deepfake abuse through public education may not be sufficient to actively stop perpetration. Education needs to specifically target motivations and intention behind deepfake NSII behaviour, necessitating the design of a novel deepfake NSII intervention.

7.2.1 How can we reduce the perpetration of deepfake abuse?

Existing interventions for deepfakes have tended to focus on misinformation, consistent with an over-focus on the political harms of deepfake misinformation in public discourse and academic literature (Gosse & Burkell, 2020; Twomey et al., 2024). Deepfake disinformation interventions have focused on the ability to educate the public on deepfake detection (Twomey, O'Mahony, et al., 2025a). For deepfake NSII, the harmful content and use in harassment are more significant concerns than the ability of the general public to detect deepfakes. Instead of focusing on identifying deepfakes (or solutions which require victims to do the bulk of the work), deepfake NSII interventions could instead aim to reduce perpetration behaviours. Indeed,

we argue that ideepfake abuse interventions should model themselves after existing programs for other forms of gendered violence such as programs for sexual assault and image-based sexual abuse.

There are an extremely limited number of active interventions targeting technology-facilitated gendered violence and none which aim to reduce deepfake abuse. Instead, we will briefly summarise the broader literature on gendered violence interventions to understand what works. Gendered violence interventions usually aim to change false beliefs and perpetration behaviours through psychoeducational interventions and bystander interventions. Early perpetrator-focused interventions towards gendered violence consisted of psychoeducational interventions which targeted beliefs that condoned, minimized, and excused perpetration (Chung, 2014). While these forms of interventions are often referred to as psychoeducational, they share little in common with the psychoeducational interventions in clinical or health sciences (which can involve therapeutic elements such as cognitive behavioural therapy; Jones et al., 2018). Psychoeducational interventions for gendered violence perpetrators often consist of programmes where educators offer role-playing scenarios as form of persuasive communication (Gilbert et al., 1991) and tend to focus on psychological mechanisms of behaviour change (changing attitudes and beliefs about behaviour). These attitudes often include forms of rape myth acceptance, which are a number of false beliefs and rhetoric about sexual assault. Rape myth acceptance is directly related to a number of self-reported perpetration behavioural outcomes (Trottier et al., 2021). Across review work there is mixed evidence as to the efficacy of rape myth interventions (Yeater & O'Donohue, 1999), though generally interventions may increase positive attitudes toward the victims of sexual violence and create more negative attitudes towards perpetration (Lonsway & Fitzgerald, 1994). Rape myth interventions tend to work most effectively when they present educational components, increase empathy and are presented as part of bystander intervention programs (Hudspith et al., 2023). The targets of psychoeducational rape-myth interventions may include current perpetrators, potential perpetrators (as determined by demographic or abuse-supportive attitudes; Flores & Hartlaub,

1998), or non-perpetrating bystanders. While bystander interventions are often not focused specifically on perpetrators, there is some evidence that bystander intervention programs actually reduce perpetration behaviours (Coker et al., 2016, 2017). In fact, the benefit of bystander interventions is that they allow researchers to challenge the values and beliefs of perpetrators without directly accusing participants of being perpetrators themselves (Katz, 2014).

Review work has highlighted the importance of interventions having a theoretical basis for attitudinal change, encouraging gendered violence interventions to employ a strong theoretical underpinning relating to how they are meant to change attitudes (Hudspith et al., 2023). As previously, mentioned, attitudinal change in interventions tends to focus on so called myth acceptance, the belief in rhetoric which seeks to blame victims, minimise harm and absolve perpetrators of guilt. Just as the perpetration of gendered violence is linked with rape myth acceptance, the perpetration of deepfake abuse is potentially linked with belief in deepfake myths. Deepfake myths are prominent in the harassment of celebrity victim-survivors and the minimising of deepfake abuse (Twomey, Foley, et al., 2025). Twomey and colleagues developed the Deepfake Myth Acceptance Scale (DMAS) which consisted of six myths around deepfake abuse which were subsequently grouped around three themes (Twomey, O'Mahony, et al., 2025b):

1. *Deepfakes are not harmful because they are not real.*

   Attempts to minimise the harms of deepfake abuse focuses on the idea that false media cannot have real harms. Deepfakes by definition involve the manipulation of a piece of media to depict an individual falsely and in fact, the non-consenting manipulation of identity is the primary harm of deepfake technology (de Ruiter, 2021). Celebrity victim-survivors report the personal damage of deepfake videos which are believed to be real as well as the harmful experience of seeing footage of themselves in sexual acts which they do not remember engaging in (Twomey, Foley, et al., 2025). Labelling a deepfake NSII video as false is not associated with lower perceived criminality and harm (Kugler & Pace, 2021).

In contrast to this myth, there is in fact some evidence that young girls see deepfake abuse as being *more* dangerous than having real images shared online because you can have no idea who made or shared the imagery, and it may already exist without your knowledge (Internet Matters, 2024).

2. *Some people deserve deepfakes*

Part of the victim-blaming rhetoric around deepfake abuse is the idea that some women "deserve" deepfakes. The myth mainly focuses on celebrity women as deserving of deepfakes due to being in the public eye and related claims that invasions of their sexual privacy are "part of the job" (Twomey, Foley, et al., 2025). Though most deepfake NSII online targets celebrities, there are a variety of other public figures who may be similarly targeted. Feminist activists are often targeted for deepfake abuse by misogynists online because they are public figures and part of efforts to reduce the harms of IBSA and NSII (Martin, 2021). Often deepfake NSII of politicians are seen as warranted or justified both towards conservative (Gozzi, 2024) and liberal women (Gutierrez, 2023). Another form of NSII content directed towards politicians includes putting misogynistic/homophobic male politicians into NSII with other men (Brieger, 2024; Twomey et al., 2023). The general justification for NSII online towards public figures focuses on deepfake abuse being expected when one has public information on social media (Kugler & Pace, 2021). The belief that public figures deserve deepfake abuse is more concerning when you consider how broad and ill-defined the term "public figure" is (Cover et al., 2025). With the rise of social media, there are ever increasing requirements for individuals to have a public presence and imagery online. It is possible to make deepfakes with even a single image or video online and victims have reported that perpetrators may access photos through befriending mutual followers on social media (Martin, 2021). Blaming victims for having an online presence draws parallels with existing myths around Image Based Sexual Abuse that blame victims for taking or sharing sexual imagery (Flynn et al., 2023).

3. *Deepfakes have no consequences*

Other myths about deepfakes seek to absolve perpetrators with the belief that there are no consequences of deepfakes and claims that perpetrators should not/do not face any consequences for their behaviour (Twomey, Foley, et al., 2025). There are two main myths focusing on the lack of consequences for perpetrators. Firstly, the idea that the creation of NSII for your own viewing is like a private sexual fantasy because no-one else sees it: the so-called "pervert's dilemma" (Öhman, 2020). Labelled deepfake NSII videos made for private viewing are not considered a crime by about 24% of people (Kugler & Pace, 2021). However, unshared sexual footage of a third party still violates their sexual privacy as deepfakes are fundamentally based off of real footage which has been doctored without a victim-survivors consent (Citron, 2019; Kugler & Pace, 2021). It is equivalent to other forms of video voyeurism or IBSA where the victim is unaware. The second form of the "no consequences" myth claims that there is a complete the lack of legal consequences for perpetrators. However, while it has taken a while for the development of deepfake-specific legislation, there are now a number of laws targeting deepfake abuse in many countries worldwide (Mania, 2024).

Twomey and colleagues found that found that deepfake myth acceptance strongly correlated with self-reported past deepfake NSII perpetration behaviours, as well as future perpetration intentions (Twomey, O'Mahony, et al., 2025b). Just as psychoeducational interventions targeting rape myths aim to reduce the perpetration of sexual assault, psychoeducational interventions targeting deepfake myth acceptance may reduce NSII perpetration intentions.

7.2.2 Deepfakes/Real Harms Intervention

Drawing from the theory of planned behaviour and from social learning theory, we have designed an online intervention (the Deepfakes/Real Harms Intervention) to increase empathy towards victims and challenge false social norms relating to deepfake NSII. In the current paper, we report three pre-registered experimental studies, with longitudinal follow-ups, to evaluate the

effectiveness and acceptability of the intervention. We targeted two main outcomes: to reduce belief in deepfake myths and to reduce self-reported perpetration intentions.

Noting the importance of a strong theoretical rationale for interventions to change attitudes towards gendered violence (Hudspith et al., 2023), our approach is grounded in the Theory of Planned Behaviour (Ajzen, 1991). The theory of planned behaviour is a well-established psychological model of behavioural change which focuses on how psychological factors influence behavioural intentions and thus behaviour (Ajzen, 1991). The three factors as applied to deepfake NSII perpetration are attitudes towards behaviour, subjective norms and perceived behavioural control. In the development of the Deepfake Myths Acceptance Scale, Twomey and colleagues note that these myths may contribute towards behavioural intentions (Twomey, O'Mahony, et al., 2025b). Deepfake myths may interact in two ways with the Theory of Planned Behaviour. Firstly, deepfake myths may play a role in attitudes towards the behaviour as the belief that deepfakes are not harmful and that perpetrators do not/should not face consequences may constitute attitudes which make perpetration behaviours more permissible. And secondly, myths may also influence subjective norms around the behaviour, as norms around public figures deserving deepfakes may also encourage NSII perpetration.

Gendered violence interventions should draw from a wide theoretical backing because the goal of an intervention is not to advocate the benefits of one theory or another but for the intervention to measurably work (Chung, 2014). While the present research mainly draws from the Theory of Planned Behaviour, it is also important to consider other theoretical models such as Bandura's Social Learning Theory, which focuses on how perpetrators model their behaviours off influential others (Chung, 2014). When applied to rape myths, social learning theory encourages interventions that alter rape-supportive cognitions, decrease problematic outcome behaviours, and/or increase empathy towards the victim (Schewe & O'Donohue, 1996). It should be noted here that raising empathy is an important goal in successful interventions and it may play a particularly important role in their effectiveness (E. Kim & Muralidharan, 2020). While the role of social learning in deepfake abuse is unclear, deepfake

abuse is often connected to networked misogyny (online groups of men organising the targeted violence and harassment of women online) (Banet-Weiser & Miltner, 2016; Robinson et al., 2024). A successful intervention should aim to make deepfakes appear socially undesirable as means of decreasing NSII-supportive cognition and perpetration behaviours as well as increasing empathy towards victims-survivors of deepfake NSII.

While in general, existing psychoeducational educations tend to report reductions in perpetration intentions, we sought to avoid common pitfalls of interventions for gendered violence. Firstly, we wanted to ensure experimental rigour; only a small minority of studies aimed at reducing intimate partner violence report the use of a control group and a comparative research design (Flood, 2014). Not using a control group is common in perpetrator-focused interventions where participants are perpetrators and where researchers have a duty of care towards potential victims. Similarly, many gendered violence interventions do not use longitudinal follow-ups and thus it is unclear if the effects of the intervention are short-lived or due to priming or demand effects (Evans et al., 2019). Most interventions are evaluated only once, rather than iteratively improved – just 4% of sexual violence prevention programs use multi-study designs (DeGue et al., 2014). Secondly, we were mindful of the risk that we would inadvertently increase intentions to perpetrate deepfake abuse. There are concerns that some sexual violence interventions may lead counter-intuitively to an increase in perpetrator behaviours (DeGue et al., 2014), in particular when an intervention may normalise gendered abuse by asking participants to imagine themselves as perpetrators. As research has shown that many members of the public have little to no experience with deepfake technology (Sippy et al., 2024) and may be introduced to the technology via our intervention materials, we set out to design an intervention that focuses on the experience of victims, rather than encouraging identification with perpetrators. Thirdly, there are always limitations in understanding if perpetrator-based interventions have worked because self-reported behavioural intentions may not reflect actual change and the high rates of non-completion in serious perpetrators (L. Kelly & Westmarland, 2014). For these reasons, we aimed to develop a brief, online intervention that

may offer a more accessible alternative to in-person interventions and thus reduce attrition (Hines & Reed, 2017; Kleinsasser et al., 2015).

The aims of the intervention are as follows:

1. To successfully reduce deepfake myth adherence in the intervention group versus a control group.
2. To reduce perpetration intentions significantly in the intervention group versus a control group.
3. To measure how long the benefits of the intervention last over time (these results will indicate if longer form or multi-stage interventions are potentially needed)

7.3 Pilot Study

7.3.1 Intervention design

Gendered violence interventions can be broken down into different types based on timing: primary and secondary (Cooper et al., 2013). Primary interventions aim to stop perpetration before it happens through education and public outreach. Secondary interventions target perpetrators after the harm has occurred. Secondary interventions also usually focus on directly treating perpetrators using an intervention program or therapy. However, the Deepfake/Real Harms intervention is a case where the two types may overlap as the intervention is designed both to stop future perpetration in a general population and also to encourage existing perpetrators to reconsider their actions.

We chose to design a brief online intervention as it is practical to deliver at scale, and accessible to broad populations. There's no evidence that any one modality of rape-myth intervention is more effective than the others (Flores & Hartlaub, 1998) and short-form interventions seem to be effective in reducing rape myth acceptance, at least in the short term (Hudspith et al., 2023). We chose an online format as deepfake abuse is a crime which exists online and is derived from online communities partaking in online targeted harassment. Existing research on bystander intervention also shows the positive potentials for online intervention

programs as a way of improving the accessibility of the interventions and targeting a broader range of groups than college students (Kleinsasser et al., 2015).

We drew from Bandura's social learning theory in using vignettes to depict the harms of deepfakes in social contexts. Vignettes are research stimuli describing a person or situation to which participants are invited to respond (Alexander & Becker, 1978; Hughes & Huby, 2004). Vignette approaches are used in 36% of all bystander intervention programmes (Mujal et al., 2021). Our approach followed best practice for the conception, design, and administration of vignettes as previously identified in the broader literature (Skilling & Stylianides, 2020). Specifically, we considered the length of the vignettes in the intervention and the benefits of using both open and closed questions as interactive elements. To improve the quality of vignettes, other methods of attitudinal change and behavioural change were also employed such as perspective taking and behavioural change techniques.

Perspective taking

Drawing from existing literature, we decided to use narrative writing as a way to induce perspective taking and empathy (Shaffer et al., 2019). Each intervention vignette involved a perspective taking task where to participants were asked to critique the perpetrators or help victim-survivors. In the perspective taking exercise, we did not want to directly put participants into the role of the direct perpetrator. We were worried that perpetrator-focused roleplay may reduce the efficacy of the measure by normalising perpetration behaviours. Failed rape myth interventions note the dangers of putting participants into a position where they directly identify with a perpetrator (they usually respond by developing arguments to exonerate themselves) (Lonsway & Fitzgerald, 1994). Instead, we asked participants to take up a bystander role and to identify with the victims' experience. In cases where participants were asked to imagine a perpetrators justification, we followed this directly up with a critique of the perpetrator.

Behavioural change techniques

We chose to incorporate existing evidence-based literature on behavioural change techniques into our intervention. We specifically drew from the behaviour change technique (BCT) taxonomy v1 (Michie et al., 2013) with a particular interest in four techniques mainly associated with reducing subjective norms.

1. Credible source (including information from an authority figure that explains the harms of the behaviour)
2. Information about others' approval (using examples from research illustrating how unpopular a behaviour is)
3. Natural consequences (including information about social and environmental consequences, anticipated regret and information about emotional consequences)
4. Imaginary punishment (asking participants to imagine negative consequences to an act)

We sought out these techniques as they are generally utilised in clinical interventions (Dugdale et al., 2016) and thus strengthen the evidence base and likely effectiveness of the work. *Design of the intervention*

We drew from existing deepfake vignettes (Kugler & Pace, 2021) as inspiration but focused on tailoring each vignette to specific myths identified in the Deepfake Myth Acceptance Scale (DMAS) (Twomey, O'Mahony, et al., 2025b) and drawing from existing literature on perpetrators and real-world examples when possible. The primary researcher piloted the intervention among members of the research team to refine the vignettes. Initially, the draft vignettes contained NSII and non-NSII examples (Figure 7.1). Non-NSII examples of deepfakes were originally included in early pilot testing as a way to challenge the myths without engaging participants with disturbing content. However, we eventually chose to mainly focus on NSII for two reasons: firstly, more general examples run the risk of stigmatising deepfakes wholesale and secondly, there was not sufficient theoretical justification to believe we could effectively reduce deepfake NSII without talking directly about deepfake NSII. These brief vignettes used a short prompt to encourage reflection on the experience of the victim.

Vignette: Some people are more deserving of being targeted by deepfakes than others	**Vignette: Some people deserve deepfakes (NSII)**
There's a guy, James, who hangs around your friend group who no one really likes. You have only really talked to him once or twice and the conversations were quite strange. A friend texts you a video clip from a Frankenstein movie, except James is the monster, being chased by villagers with pitchforks. The text underneath says, "what we want to do to James every time he shows up at the party". It is clearly a deepfake video, a funny one, so you like the message. *How would you react if you found out your friends made a similar video of you?*	There's a person in your family, Aurora, who everyone dislikes for many valid reasons. While she attends most of the social events your family organises, everyone treats her coldly. Recently, she was a victim of non-consensual deepfake images, which were shared by an ex-partner to "prove" she was cheating on him. While many of your family publicly offered words of support, privately they make jokes about it. One of your cousins is particularly dubious and she has started spreading rumour that the images were real and Aurora only called them deepfakes because she was embarrassed. Your mother now refuses to talk to this cousin because she is only making things worse and "real or fake, nobody deserves to have their name dragged through the mud and images like that shared". *How would you explain to your cousin why spreading these rumours is wrong?*

Figure 7.1: Early drafts of the one-part vignettes including a non-NSII example

It was eventually decided to focus on the three broad categories of deepfake myths instead of addressing each of the six DMAS myth individually and that longer form vignettes would be more suitable. Each of the vignettes in the final intervention had three parts within it (note that all versions of the intervention are available in Supplemental Materials on OSF). The first of these explained a social situation involving deepfake perpetration and asked participants to reflect on the myths/harm of the technology e.g. "*You have an old family friend named Doug. He has had a crush on a cousin of yours, named Zoe, for years. They used to be very good friends, but they've grown apart. One night when Doug has been drinking, he starts talking about Zoe, he tells you that he's been editing adult videos and putting images of her into them. He says this is okay to do and amounts to the same as a sexual fantasy because he is not sharing them with anyone else. If they're only for his viewing, it's not hurting anyone. Reflect for a moment. Why do you think Doug believes this? "*

The second part required the participant to directly write a short piece where they took the perspective of a bystander either apologising to the victim for not doing enough to stop

deepfake abuse or writing to the perpetrator to explain the wrongs of his action e.g. "*You are not the only person he has talked about this to while drinking. It gets back to Zoe, and she writes a long post on social media detailing how upset she is. She asks you to write a message to Doug explaining why this is making her so upset. What would you write? (15 words min)*"

The final part of the intervention relied on educating the participant about the deepfake myth and information from a credible source or information about others' social approval (as per the behavioural change techniques literature) "*Research has shown that most people find creating non-consensual deepfake videos to be morally wrong, especially if they are of people you know.*

There is increasing legal focus on the use of deepfakes to create abuse imagery. ***These acts are not without consequences: recently a man went to prison in the United Kingdom for using deepfakes to create abuse imagery of children for his own personal use.*** *Do you agree that it is wrong to make deepfake videos of people you know, even if you don't share them?* Participants were asked to fill out a four-item Likert scale from Strongly Disagree to Strongly Agree, the aim of this question was not as a measurement but to force participants to engage with this social norm information and to consider their stance.

Five friends of the research team were asked to run through the intervention to ensure the length was appropriate and the prompts were clear and easy to implement. As the intervention aimed to use a multi-study design there was no need for a more complex piloting process as the intervention would be improved after every iteration.

7.4 Study 1: Intervention Evaluation

We ran the first study of the intervention to test if the first iteration would successfully reduce attitudes and perpetration intentions. We preregistered the study at https://aspredicted.org/bwzd-tc9p.pdf. While we noted in the preregistration that we would combine the data with university data, we chose not to do this because the data could not be followed up on after time and we only received 133 responses while over 150 participants did

not complete the survey. The materials for all the studies and supplementary analyses are available at https://osf.io/xvukt/overview?view_only=a7ad34ab296847bf82db93b437bb847c

7.4.1 Method

7.4.1.1. Participants

Sample size was pre-determined using GPower Version 3.1.9.7 for an independent samples t-test with an effect size d of 0.2. For the first study stage the target population was kept as broad as possible. We recruited all the participants using Prolific, a paid survey taking platform across Ireland, the United Kingdom and the United States of America. Participants were screened to have fluency in English. The study was advertised as a "Deepfake Attitudes and Behaviour Study" which would focus on participants experiences of the positive and negative uses of deepfakes. Participants were divided into a control group (n = 280) and an experimental group (n = 264). In the dataset 311 participants identified as Female,226 identified as Male, 2 self-described as non-binary and 3 preferred not to say. The average age was 42.80 (SD = 13.86) with a range from 18-78. Most participants had completed a college degree (n = 359).

7.4.1.2. Measures

Interventions are often hampered by untested measures with dubious reliability (Mujal et al., 2021). For our research we drew on psychometrically-sound behavioural/attitudinal measures from existing literature.

Deepfake Myth Acceptance Scale (DMAS)

The DMAS is a six-item scale measured using five-point Likert scales (1 = Strongly Disagree, 5 = Strongly Agree) (Twomey, O'Mahony, et al., 2025b).

1. When deepfake videos are just jokes, they aren't harmful
2. Deepfake videos are only harmful when someone believes they're real
3. It is okay to make deepfakes of celebrities because they are public figures
4. Some people are more deserving of being targeted by deepfakes than others
5. It is okay to make deepfakes when there are no legal consequences
6. Deepfakes aren't harmful if they're not shared online

For transparency, it is important to note that the DMAS was developed in tandem with the intervention but only one part of the datasets overlap (the control group in Study 3 of the present paper is the same as Study 1 in the DMAS paper. As a result, a number of the pre-registrations are inconsistent about whether the scale is six or seven items as the scale was reduced to six items during its development and the items of the scale also went through minor changes before Study 3 (e.g. item 6 used the word "posted" instead of "shared" and item 5 used "because" instead of "where"). While all participants in Study 2 received the same scale, there also unfortunately was a typographical error on the item points in this iteration of the intervention where the two positive items were listed as "possibly agree" and "agree", as opposed to "agree" and "strongly agree". While technically still equally weighed in terms of positives and negative items, this error was rectified in Studies 3 and 4.

Behavioural Interventions

The behavioural measure was five-point Likert scale focused on three salient aspects of past deepfake NSII perpetration behaviours: watching, sharing and creating drawing from previous literature (Umbach et al., 2024). Participants were also asked about future NSII perpetration intentions. While we disagree with the use of the term deepfake pornography, we chose to use the term in the behavioural measurement as it is the term that the average person

would use. These questions were modelled off existing work which measured perpetration behaviours (Umbach et al., 2024). To reduce response bias and include measures of less stigmatised engagement with deepfake media, we also asked participants about past and future non-NSII deepfake behaviours (e.g. engaging with comedic or political deepfakes) but as per our pre-registration, we did not incorporate these into our analysis. For a full outline of all the questions used, please see Supplemental Materials.

7.4.1.3. Procedure

Participants first completed a consent form and demographic information. They were then introduced to the concept of deepfakes through a visual example (an image showing the before and after of an entertainment deepfake of Tom Cruise). One group was tasked with completing the intervention. All participants then filled out the DMAS and the behavioural measures. Finally, all participants were debriefed (described below in Ethics) and given the option to provide general feedback, as well as the option to withdraw their consent. We also asked participants to complete a follow up study two weeks after the initial data collection to assess whether any effects of the intervention would change over time. The measures taken were the same, but participants did not carry out the intervention. Participants were again debriefed on the harms of deepfakes. We also took more general qualitative feedback from participants after the follow up.

7.4.1.4. Ethics

Ethical approval was received from our departmental ethics committee at (*anonymised for review*). As well as the risk of participant distress, we were mindful that completing the myths measure might inadvertently spready misinformation amongst the control group (i.e. that participants might think the DMAS items were true and reflect a scientific or legal consensus around deepfakes). Thus, we followed best-practice guidelines from misinformation research and designed debriefing materials that explicitly challenged these myths and made it clear to participants that we were not endorsing them (Murphy & Greene, 2023).We also took their

informed consent into account by offering participants the chance to withdraw consent after revealing the purpose of the survey (Greene et al., 2023). We made sure to link to existing support services for deepfake abuse (such as STOPNCII.org) at the end of the study.

7.4.2 Results

The average deepfake myth score was 9.82 out of a possible 36 (SD = 4.27). Across all the behavioural items participants scored close to the minimum value of one (on a scale from 1--5, indicating floor effects in the dataset (Table 7.1). This is consistent with the generally low levels of perpetration across the general population globally (Sippy et al., 2024; Umbach et al., 2024).

Table 7.1: Means and standard deviations for self-reported perpetration behaviours and intentions in Study 1

Self-reported perpetration behaviours/intentions	M	SD
Have you watched a deepfake pornographic video before?	1.43	0.95
Have you shared a deepfake pornographic video before?	1.06	0.38
Have you created a deepfake pornographic video before?	1.03	0.26
Would you watch a deepfake pornographic video in future?	1.44	0.89
Would you share a deepfake pornographic video in future?	1.07	0.42
Would you create a deepfake pornographic video in future?	1.04	0.32

7.4.2.1. Post-Intervention

Deepfake myth acceptance was higher in the control group (M = 10.38, SD = 4.45), than the intervention group (M = 9.23, SD = 3.99) (fig. 7.1), which was statistically significant

with a small effect size ($t(540) = 3.15$, $p = .002$, $d = 0.27$). As expected, for past perpetration there was no significant difference between groups ($t(534) = 1.15$, $p = .253$, $d = 0.10$) between the control ($M = 1.47$, SD = 1.03) and the intervention group ($M = 1.38$, SD = 0.87).

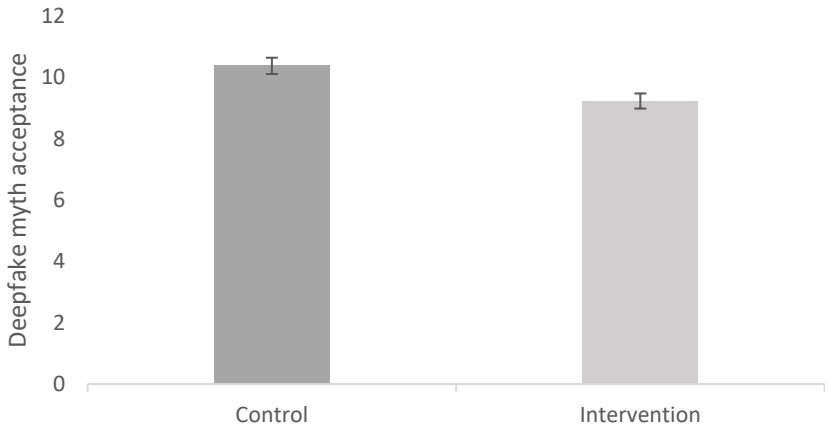


Figure 7.1: Deepfake Myth acceptance by group in study 1

There was some evidence indicating that the intervention successfully reduced future intentions to engage in deepfake abuse perpetration with small effect sizes. Future intention of watching NSII were significantly ($t(519.34) = 2.79$, $p = .006$, $d = 0.24$) higher in the control group ($M = 1.54$, $SD = 0.99$) compared to the intervention group ($M = 1.33$, $SD = 0.76$) (Figure 7.2). While future sharing intentions were higher in the control group ($M = 1.10$, $SD = 0.52$) than the intervention group ($M = 1.03$, $SD = 0.28$), the difference was not quite significant ($t(437.95) = 1.958$, $p = .051$, $d = .17$). Finally, regarding NSII creation, intentions were significantly higher in the control group ($M = 1.07$, $SD = 0.43$) than the intervention group ($M = 1.03$, $SD = 0.28$) ($t(328.08) = 2.12$, $p = .035$, $d = .18$).

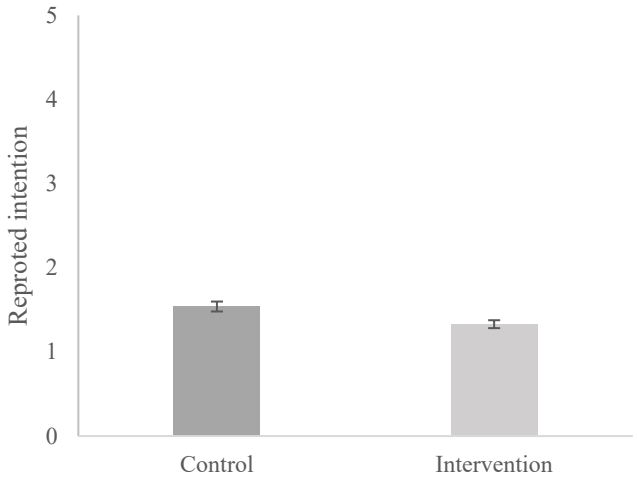


Figure 7.2: NSII watching intentions by group in study 1

7.4.2.2. Follow-Up

At follow up two weeks later, the study retained 212 participants from the control group and 194 from the intervention group. We first conducted exploratory t-tests to see if the significant effects from the first iteration remained. There were no longer significant differences across any of the conditions (Table 7.2).

Table 7.2: Exploratory analysis of differences between groups at follow up

	Control T2 (SD)	Intervention T2 (SD)	t(df) =	Significance (two-sided)
Deep fake myths	10.52 (4.57)	10.26 (4.43)	t(404) = 0.60	.552
NSII Watch Future	1.44 (0.90)	1.42 (0.86)	t(404) = 0.18	.855
NSII Share Future	1.07 (0.40)	1.07 (0.33)	t(404) = -0.04	.969
NSII Create Future	1.03 (0.19)	1.05 (0.34)	t(404) = -0.66	.507

We conducted mixed between-within subjects ANOVAs to test if there was a difference between the two groups over time and interaction effects between time and group (see Supplemental Materials). While we found mostly null effects, the few significant effects

suggested both decreases in the control group and increases in the intervention group over time (Figure 7.3).

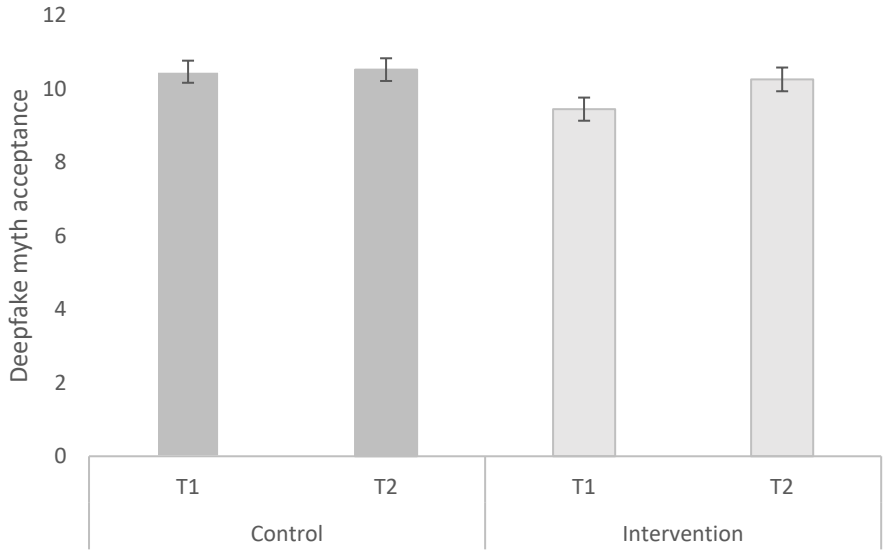


Figure 7.3: DMAS scores across the control and intervention group at times 1 and 2 for study 1.

On reflection, we felt these results may have been partly due to the extensive debriefing provided to all participants at the end of the first survey, which may have acted as an intervention for those in the control group. This mini intervention for the control group, coupled with the already small effects observed within the intervention group, would explain the null effects at follow-up.

7.4.2.3. Qualitative feedback

We have chosen here to report the qualitative feedback along with age, gender and reported NSII watching perpetration to contextualise responses. The feedback was overwhelmingly positive for the benefits of the study. Participants focused on the importance of the work in general "*I think it's good that you are working to show that deep fakes are harmful and educate people about why the myths are untrue*" (63, F, No past or future perpetration reported) or how the intervention educated them on the topic "*Thank you for opening my eyes to something I'd not really thought about"*[62, M, Unsure about past watching, no other indicated perpetration]. One participant also noted the need for education and interventions aimed at

younger individuals*: "More education is needed around these deep fake images, especially for teenagers who think things are a joke but can be really distressing*" (57, F, No past or future perpetration reported). Some participants noted at follow up that they felt the benefits of the intervention and how it has changed their views towards NSII "*interesting and I have got a more negative view of them now after recalling the previous study*" (43, F, No past or future perpetration reported). This also seems to have affected one respondents interactions with a peer: "*Deceit is wrong especially deep fakes - my respect for a friend who does this is now non-existent"* (47, F, No past or future perpetration reported) which indicates the success of the intervention at challenging social norms. However, one participant continued to justify deepfake perpetration behaviour in their feedback noting "*its just fun*" (44, M, Reported unsure about past perpetration but reported future intentions which reduced slightly at follow up). While the feedback here was mainly positive, we wanted to gather more critical feedback for the benefit of improving the intervention which led to use changing the feedback question in the subsequent studies.

7.5 Study 2: Revised intervention evaluation

Overall, while we found significant effects in terms of deepfake myth acceptance in Study 2 we felt that the intervention could be stronger and that this might maintain the effects at follow up. We also decided to take more granular measures of behaviours varied by victim-type as existing research has shown deepfake abuse as more socially acceptable when directed at celebrity women (Fido et al., 2022; Kugler & Pace, 2021). We hoped this would allow a closer look at the intervention effectiveness, as well as possibly reducing floor effects in reported perpetration behaviours by providing more detailed data on more common NSII watching intentions.

7.5.1 Method

7.5.1.1. Participants

Just as with Study one, we used Prolific to recruit participants advertised in the same way as study 1 and with the same power analysis. There were 544 participants in this iteration as there was a Prolific error which added two extra participants. The average age was 42.44 years old (SD = 13.38), and the ages ranged between 19 and 81 years old. In the dataset, 285 people identified as male, 256 identified as female and 3 identified as Non-binary/Third gender. In terms of education, six participants reported some schooling, 152 graduated high school, another 251 reported graduating with a college degree and 135 reported having attained a postgraduate degree.

7.5.1.2. Materials

We wanted to improve the intervention after the first round of feedback and we specifically targeted the wording of the final vignette, concerning the idea that deepfakes are deserved by some people as we felt the deserved deepfakes myth was not as obvious to readers. We changed the wording of the reflective task from "How do you feel the harms of fake videos compare to real videos?" to "How do you think it feels for a public figure to see fake videos of themselves online" with the aim of better promoting empathy towards victims. We also changed the wording of a sentence in the final intervention vignette from "*Just because someone is in the public eye, it doesn't mean that they deserve to have their consent overridden*" to "*Just because someone is in the public eye, it doesn't mean that they deserve to have to deal with non-consensual intimate imagery and defamatory videos*" to make the myth easier to understand for the reader.

The measures were mostly the same as Study 1, with the exception that we added the word intentionally into our behavioural measures as there are a number of ways in which an individual can accidentally be exposed to deepfake NSII online. However, we also added

questions which specifically measured future perpetration intentions by victim-type such as a friend, ex, stranger or celebrity drawing from existing work in the field (Kugler & Pace, 2021; Twomey, O'Mahony, et al., 2025b)

7.5.1.3. Procedure

The procedure was the same as the previous study, except that we requested both general feedback and critical feedback (where participants were asked how to improve the intervention) for the intervention and we specifically asked participants how they felt about the use of the intervention in schools. The follow up was conducted one week after the initial study following the same procedure as the Study 1 follow-up.

7.5.2 Results

In Study 2, the reported deepfake myths adherence was higher than in Study 1 (potentially due to the changes to the scale) with a mean of 11.31 and a standard deviation of 5.38. While there were similar floor effects in the victim-type behavioural measures (Table 7.3), NSII intentions towards celebrity women was the highest across all forms of perpetration measures. This is consistent with work which highlights how deepfake abuse tends to be mainly targeted/justified towards celebrities (Fido et al., 2022; Twomey, Foley, et al., 2025).

Table 7.3: Self -reported perpetration behaviours and intentions

Perpetration behaviours/intentions	M	SD
Have you intentionally watched a deepfake pornographic video?	1.69	1.20
Have you intentionally shared a deepfake pornographic video?	1.21	.67
Have you intentionally created a deepfake pornographic video?	1.16	.62
Would you intentionally watch a deepfake pornographic video?	1.66	1.11
Would you intentionally share a deepfake pornographic video?	1.22	.67
Would you intentionally create a deepfake pornographic video?	1.18	.64
How likely are you to watch a pornographic deepfake of a friend?	1.33	.87
How likely are you to watch a pornographic deepfake of an ex romantic partner?	1.46	1.00
How likely are you to watch a pornographic deepfake of a stranger?	1.65	1.14
How likely are you to watch a pornographic deepfake of a celebrity?	1.78	1.25
How likely are you to share a pornographic deepfake of a friend?	1.16	.64
How likely are you to share a pornographic deepfake of an ex romantic partner?	1.17	.68
How likely are you to share a pornographic deepfake of a stranger?	1.21	.69
How likely are you to share a pornographic deepfake of a celebrity?	1.25	.77
How likely are you to make a pornographic deepfake of a friend?	1.15	.59
How likely are you to make a pornographic deepfake of an ex romantic partner?	1.15	.61
How likely are you to make a pornographic deepfake of a stranger?	1.16	.63
How likely are you to make a pornographic deepfake of a celebrity?	1.18	.65

7.5.2.1. Post-intervention

There was a significant difference in DMAS scores between the two groups ($t(542) = 3.69$, $p < .001$, $d = .32$), where deepfake myth acceptance was higher in the control group ($M = 12.13$, $SD = 5.35$) than in the experimental group ($M = 10.45$, $SD = 5.23$) (Figure 7.4). While the intervention group reported slightly lower past NSII watching behaviour ($M = 1.63$, $SD = 1.160$) than the control group ($M = 1.75$, $SD = 1.24$), the results were not significant ($t(542) = 1.18$, $p = .240$, $d = .10$).

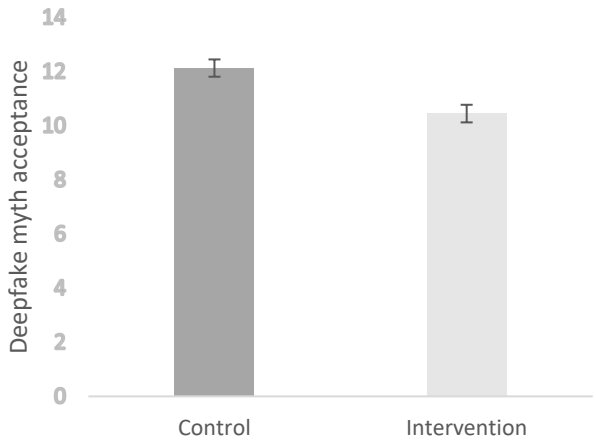


Figure 7.4: Deepfake Myth acceptance by group in study 2

There was a significant difference in future intentions to watch NSII ($t(538.87) = 2.86$, $p = .004$, $d = .24$) between the control ($M = 1.79$, $SD = 1.18$) and intervention group ($M = 1.52$, $p = 1.03$) (Figure 7.5). For future NSII sharing intention, the control group ($M = 1.24$, $SD = 0.68$) scored similarly to the intervention group ($M = 1.19$, $SD = 0.65$) and so the difference between groups was not significant ($t(542) = .93$, $p = .351$, $d = .08$). Similarly, regarding future creation of NSI, the control group again reported similar future intentions ($M = 1.20$, $SD = .67$) than the intervention group ($M = 1.16$, $SD = .61$) and the results again were not significant ($t(542) = .74$, $p = .460$, $d = .06$).

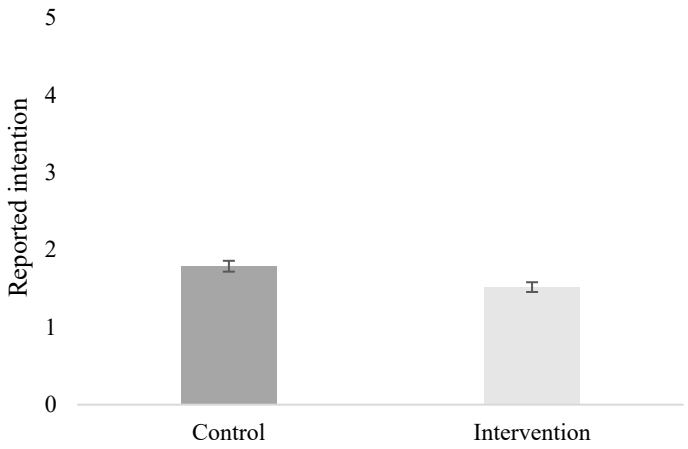


Figure 7.5: NSII watching intentions by group in study 2.

For future NSII perpetration, none of the sharing or creating behaviours were significantly different between groups, though the floor effects were quite notable for both perpetration behaviour categories (see Table 7.4). The control group were however more likely (t(532.24) = 2.18, p = .030, d = .19) to report future intention to watch NSII of an ex (M = 1.55, SD = 1.09) versus the intervention group (M = 1.36, SD = 0.90). Similarly, there was a significant difference between the control group (M = 1.90, SD =1.30) and the intervention group (M = 1.66, SD = 1.18) for future NSII watching of a celebrity (t(541.72) = 2.297, p = .040, d = .20).

Table 7.4: Comparisons between experimental and control group across measures

	Control group mean (SD)	Experimental group mean (SD)	t(df) =	Significance (two sided)	Cohen's D
Future Watch Intentions					
Friend	1.35 (.87)	1.31 (.87)	t(542) = 0.43	.668	.04
Ex	1.55 (1.09)	1.36 (.90)	t(532.24) = 2.18	.030	.19
Stranger	1.73 (1.18)	1.56 (1.09)	t(541.72) = 1.76	.079	.15
Celebrity	1.90 (1.30)	1.66 (1.18)	t(541.23) = 2.30	.040	.20
Future Share Intentions					
Friend	1.14 (.61)	1.18 (.67)	t(542) = -.71	.476	-.06
Ex	1.18 (.69)	1.17 (.66)	t(542) = .08	.938	.01
Stranger	1.21 (.70)	1.21 (.67)	t(542) = .04	.971	.00
Celebrity	1.29 (.83)	1.21 (.70)	t(535.61) = 1.17	.243	.10
Future Create Intentions					
Friend	1.14 (.59)	1.15 (.60)	t(542) = -.10	.924	-.01
Ex	1.15 (.64)	1.14 (.58)	t(542) = .26	.797	.02
Stranger	1.16 (.62)	1.17 (.65)	t(542) = -.18	.861	-.02
Celebrity	1.19 (.68)	1.16 (.63)	t(542) =.47	.637	.04

7.5.2.2. Follow-up

An exploratory analysis demonstrated that though the control group reported higher DMAS scores at follow-up (M = 11.75, SD = 5.20) compared to the intervention group (M = 10.87, SD = 4.67), this did not quite reach statistical significance (t(476) = 1.94, p = .053, d = .18). Across the dataset (with the exception of future watching for ex romantic partners) there were no significant differences between the groups (see supplementary materials).

As in Study 1, we ran mixed between-within subjects ANOVAs to test if there was a difference between the two groups over time (see supplemental materials). The means in the control group again decreased again due potentially to exposure to the debrief (Figure 7.6).

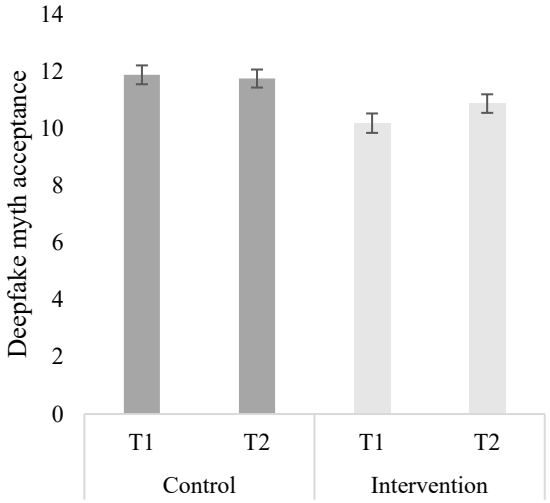


Figure 7.6: DMAS scores across the control and intervention group at times 1 and 2 for study 2.

To account for the effect of the debrief in reducing the control group, we ran an exploratory analysis comparing the intervention after 1 week to the original control group values. We carried out a number of independent t-tests comparing the two groups (see supplementary materials) and found that there was still significant differences between the Deepfake myths ($t(508.53) = 2.84$, $p = .005$, $d = .25$) between the control ($M = 12.13$, SD = 5.35) and intervention groups ($M = 10.87$, SD = 4.67) We also found a difference in future NSII watching behaviours ($t(509.49) = 2.93$, $p = .003$, $d = .26$) between the control ($M = 1.79$, SD = 1.18) and intervention groups ($M = 1.50$, SD = 1.01). Finally, across NSII watching intentions by victim type we noticed a difference for ex romantic partner, stranger and celebrity but not for a friend (Figure 7.7).

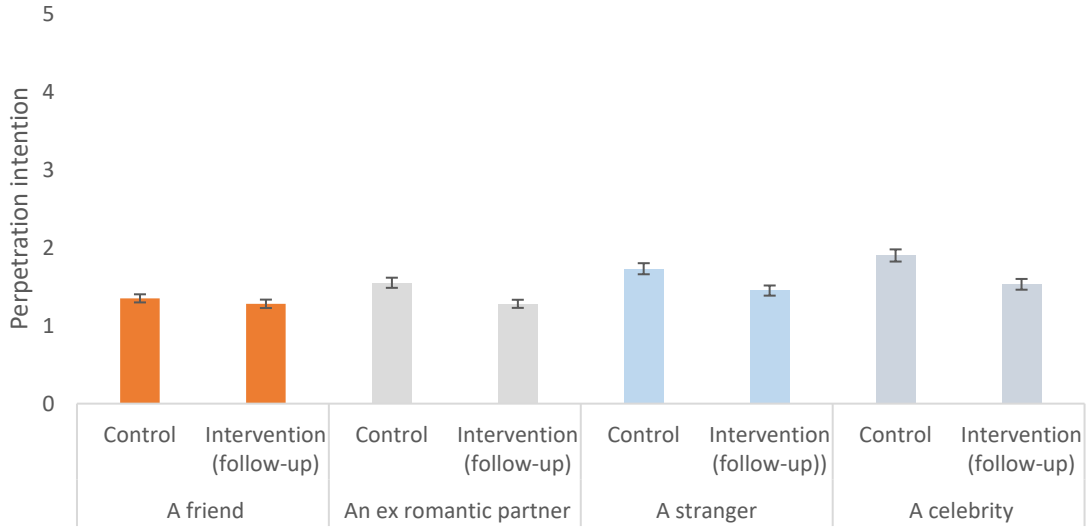


Figure 7.7: Intended NSII watching perpetration by victim group in study 2 comparing the original control group with the intervention group at follow up one month afterwards

7.5.2.3. Qualitative feedback

General feedback from the intervention highlighted a change in values or recognising the harm, especially after follow up: "*I gained a better understanding of the various situations in which deepfakes could be employed thanks to the enlightening vignettes. I've been considering it for a week, and now I'm more conscious of the moral dilemmas and possible risks. This type of intervention, in my opinion, aids in perspective shifting by personalizing the repercussions*" (63, F, Both past and future perpetration reported with no change in follow up). One participant noted that they would be interested in incorporating the intervention into education "*Great study, I work in education and will incorporate this subject into a lesson*" (57, M, Reported past perpetration but not future intentions). We also saw a focus on peer-behaviours, but this was in the control group: "*I have a friend who told me that there are not victims when looking a porn online. I tend to agree with him. Looking at porn may seem harmless, but the picture or video may have been created against one's will, or when someone was vulnerable. There are always victims, whether the images are real, or fake. This is my take*"(68, M, Reported past watching behaviour and future watching intentions).

As we wanted to further improve the intervention for the next study, we specifically asked participants for critical feedback as to how the intervention could be improved in contrast to the first intervention which just asked for feedback. Critical qualitative feedback demonstrated that participants wanted more detail on the consequences of deepfake abuse. "I *would get an interview with someone who has been personally affected by deepfake videos so that people can understand the damage they can do*" (28, F, No past or future perpetration reported). The desire for more detail and more real examples came up a number of times, with one participant noting "*You've made something fake to say that we should not do the same*" (27, M, No past or future perpetration reported). In contrast to our initial aims to keep the intervention as non-explicit and non-distressing as we could, many participants recommended more graphic elements, with an especial focus on interactive and evocative elements. "*I would include more real-world examples of the harm caused by deepfakes, such as legal outcomes or emotional stories from victims. Also, adding interactive elements or short videos could help make the experience more personal and memorable, especially for students*" (62, M, No past or future perpetration reported). In general, the suggestions for improvement were all broadly similar and related to more detail and more engaging content: "*Enhance the intervention with real testimonies, ethical visuals, interactive elements, legal context, and actionable steps*" (30, M, Reported past watching behaviour but not future watching intentions).

We also asked the participants if they believed that the intervention should be run in schools on a five-point Likert scale (M = 4.21, SD = 1.03). Again, in qualitative feedback participants expressed the need for interventions targeting teenagers: "*the scenarios mentioned in the survey happens mostly in schools. i would really recommend this intervention to be introduced in schools*" (25, F, Reported past watching behaviour but not future watching intentions). Across the qualitative feedback, participants further reflected on how to improve the efficacy of this type of intervention for younger adults and teenagers by changing the targets of the vignettes "*Make them more relatable to teenagers like using teenagers in the scenarios*" (19, M, No past or future perpetration reported). Another participant wanted the vignettes to

specifically focus on the consequences for students "*Say more about the consequences. For example, a student died by suicide after a video of him in an intimate situation was posted online. This emphazises that it's not a joke and how harmful it can be regardless of the intention*" (57, F, No past or future perpetration reported). We reflect briefly on the potentials for ethically designing a version of this intervention for schools in the discussion section.

7.6 Study 3: Evaluating a further-revised intervention in higher perpetrating populations

The final iteration of the study was aimed specifically at studying higher perpetrating populations and using improved vignettes. This iteration was preregistered at https://aspredicted.org/zhbb-bjxt.pdf

7.6.1 Methods

7.6.1.1. Materials

The intervention was broadly the same as Study 2, with a number of changes aimed at improving the emotional saliency. To account for feedback regarding multi-modal materials and direct victim-survivors experience, we included a video from an Australian news organisation which interviewed a number of ordinary women who had experienced deepfake abuse perpetrated by one man (a link to the video is available on the OSF page). The video was 53 seconds long and contained three women making statements about the perpetrator such as "I searched for almost a year to feel safe in my body again" and "You shattered the sense of security I used to take for granted". We chose this video as it highlighted the various consequences of deepfake abuse and the perspectives of victim-survivors but was not gratuitous in the content which it showed, and it did not use clickbait. We also asked participants an open question on the video as an attention check. Feedback had indicated that multi-modal content would be more captivating for the viewer. We included images as part of the vignettes, two of these represent the text conversations mentioned in the vignettes and the other depicted a man at

a computer representing a creator of the deepfakes. We also noted based on feedback that the vignettes were developed from real world examples.

7.6.1.2. Participants

To compensate for the floor effects in previous iterations and the literature as a whole, we decided to study a high-preparation sample. To accommodate this, we decided to profile the average deepfake perpetrator and use this profile to filter our Prolific recruitment. Based on our previous data (Twomey, O'Mahony, et al., 2025b) and existing literature (Fido et al., 2022; Flynn et al., 2022; Kugler & Pace, 2021) we chose to specifically recruit men between the ages of 18 and 45. We also chose to filter Prolific participants by those who had previously expressed (via Prolific demographic questionnaire) watching online pornography in the past six months. We again recruited 542 participants. As expected, the entirety of the sample were men between the ages of 18 and 45. The average age was 33.19 (SD = 6.43). In terms of education level, nine reported having had some schooling, 170 reported having graduated high school, 267 reported having a college degree and a further 96 reported having a postgraduate degree. There were 301 participants in the control group and 241 in the intervention group.

7.6.1.3. Measures

We changed the wording of the feedback in the survey to have both positive and negative feedback/ We were concerned that the highly critical focus of our feedback questions (while useful for design) may backfire as interventions which encourage participants to question and doubt the intervention's efficacy are associated with worse outcomes (Roozenbeek et al., 2025). To resolve this, we also asked participants to describe what they liked in the intervention. And instead of asking what participants *would* change about the intervention, we asked them *if* there was anything about the intervention they would change.

7.6.1.4. Procedure and ethics

The procedure and ethics were the same as Study 2 with the addition of the news video as before the vignettes and asking participants for positive and negative feedback. For the

follow up, we again took follow up data after one week. However, in this study we also followed up again after one month to measure effectiveness over a longer period. We recruited over a week for this follow up to try and gather a larger amount of participants to make up for the expected drop off after the month. We pre-registered the month-long follow up separately at https://aspredicted.org/7ry6av.pdf.

7.6.2 Results

For this sample, the average deepfake myths score was 12.30 with a standard deviation of 5.66. Across the board, the perpetration behaviours and intentions ranked higher than in previous iterations, though most participants still did not report perpetration (Table 7.5).

Table 7.5: Means and standard deviations across perpetration measures

	M	SD
Have you intentionally watched a deepfake pornographic video?	2.14	1.50
Have you intentionally shared a deepfake pornographic video?	1.13	.54
Have you intentionally created a deepfake pornographic video?	1.13	.62
Would you intentionally watch a deepfake pornographic video?	2.09	1.30
Would you intentionally share a deepfake pornographic video?	1.19	.62
Would you intentionally create a deepfake pornographic video?	1.22	.70
How likely are you to watch a pornographic deepfake of a friend?	1.52	1.07
How likely are you to watch a pornographic deepfake of an ex romantic partner?	1.58	1.12
How likely are you to watch a pornographic deepfake of a stranger?	1.81	1.23
How likely are you to watch a pornographic deepfake of a celebrity?	2.13	1.39
How likely are you to share a pornographic deepfake of a friend?	1.10	.48
How likely are you to share a pornographic deepfake of an ex romantic partner?	1.12	.56
How likely are you to share a pornographic deepfake of a stranger?	1.17	.61
How likely are you to share a pornographic deepfake of a celebrity?	1.23	.71
How likely are you to make a pornographic deepfake of a friend?	1.15	.60
How likely are you to make a pornographic deepfake of an ex romantic partner?	1.19	.70
How likely are you to make a pornographic deepfake of a stranger?	1.19	.68
How likely are you to make a pornographic deepfake of a celebrity?	1.24	.74

7.6.2.1. Post-intervention

For deepfake myth acceptance, the control group reported higher levels of myth acceptance (M = 13.14, SD = 5.76) than the intervention group (M = 11.24, SD = 5.36) and this was significant ($t(540) = 3.93$, $p < .001$, $d = .34$) (Figure 7.8). Future NSII watching intentions were also significantly different between the groups ($t(536.88) = 2.80$, $p < .001$, $d = .34$) with the control group reported a higher likelihood (M = 2.28, SD = 2.37) than the intervention (M =

1.11, *SD* = 0.44). However, past watching behaviour was also significantly different between the two groups ($t(535.32) = 2.80$, $p = .005$, $d = .24$) where the control group reported more past experience (*M* = 2.30, *SD* = 1.56) than the intervention (*M* = 1.95, *SD* = 1.37), which potentially demonstrates perpetrator dropout before completion, or response bias. For NSII sharing intentions, the results were again significantly different between groups ($t(539.05) = 2.94$, $p = .003$, $d = .24$) with the control group (*M* = 1.26, *SD* = 0.72) reporting higher intentions than the intervention group (*M* = 1.11, *SD* = 0.44). Self-reported creation intentions were not significantly different between groups ($t(539.05) = 1.50$, $p = .133$, $d = .13$) though rates for both groups were very low and the control group reported slightly higher perpetration (*M* = 1.26, *SD* = 0.74) than the intervention (*M* = 1.17, *SD* = 0.62)

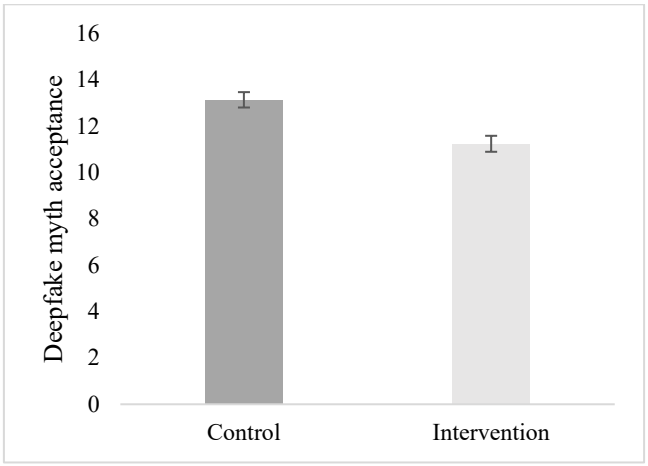


Figure 7.8: Deepfake Myth acceptance by group in study 3

By perpetration group, the intervention group were less likely to report intentions to watch, share or create deepfake NSII of celebrities. The intervention group also reported less likelihood to watch and share deepfake NSII of strangers and to share deepfake NSII of ex partners. All the other results were non-significant though the intervention group scored lower in each (see Table 7.6).

Table 7.6: Difference between the control group and intervention by victim-type

	Control group mean (SD)	Intervention group mean (SD)	*t*(df) =	Significance (two sided)	Cohen's D
Future Watch Intentions					
Friend	1.58 (1.15)	1.44 (.95)	*t*(539.23) = 1.49	.137	.13
Ex	1.64 (1.17)	1.50 (1.05)	*t*(532.11) = 1.47	.143	.13
Stranger	1.97 (1.30)	1.61 (1.11)	*t*(537.36) = 3.52	<.001	.30
Celebrity	2.34 (1.46)	1.87 (1.26)	*t*(537.38) = 4.01	<.001	.34
Future Share Intentions					
Friend	1.13 (.56)	1.07 (.35)	*t*(509.57) = 1.69	.092	.14
Ex	1.16 (.67)	1.07 (.38)	*t*(492.88) = 2.12	.035	.17
Stranger	1.23 (.70)	1.10 (.47)	*t*(524.19) = 2.43	.015	.20
Celebrity	1.32 (.85)	1.10 (.46)	*t*(477.99) = 3.81	<.001	.31
Future Create Intentions					
Friend	1.18 (.67)	1.12 (.51)	*t*(538.62) = 1.18	.239	.10
Ex	1.23 (.80)	1.13 (.54)	*t*(524.77) = 1.81	.072	.15
Stranger	1.22 (.74)	1.15 (.59)	*t*(534) = 1.28	.202	.11
Celebrity	1.29 (.82)	1.17 (.62)	*t*(538.82) = 2.04	.042	.17

7.6.2.2. Follow-up

We report the findings of the one week follow up in Appendix VI and note they were broadly similar to the findings in the previous iteration. In the previous studies' follow ups, we observed that the control group's NSII supportive attitudes and behavioural intentions decreased over time. We believe that the change in attitudes is a result of the effective debriefing. The process by which we debriefed control participants may also be considered as a form of educational intervention. To account for the effect of debriefing, we decided to follow up with the intervention group only at one month and compare these results to the original control group and experimental group results. For myth scores, there was no difference between the original control group (M = 13.14, SD = 5.76) and the intervention group after one month (M = 12.60, SD = 6.04) (Figure 7.9).

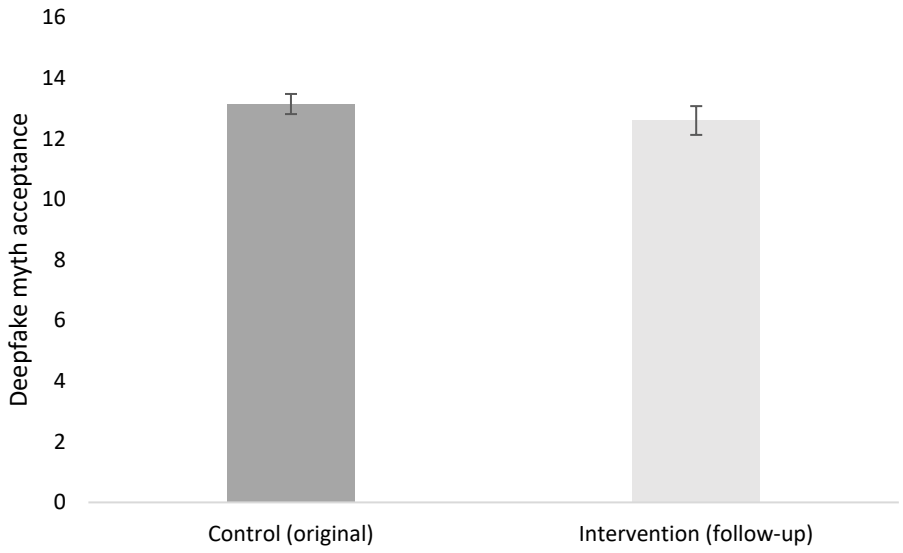


Figure 7.9: DMAS scores between control group at initial data collection and intervention group at the follow-up one month afterwards.

As expected from a month long follow up, the values tended to return closer to baseline, though there were still significant differences between groups in relation to watching behaviours (Table 7.7). We ran an paired samples t-test on the data and found that the only significant difference between the groups ($t(160) = -3.60$, $p < .001$, $d = .28$) was in the deepfake myths condition where the initial intervention group ($M = 11.32$, SD = 5.42) scored lower than the follow-up ($M = 12.64$, SD = 6.04). This indicates that while attitudes returned close to baseline since the intervention, they are not significantly different from the post-intervention results and future watching intentions also remained lower than the control group.

Table 7.7: Comparison of the means in the month long follow up versus the original control group

	Control group mean (SD)	Intervention group mean after month (SD)	*t*(df) =	P (two sided)	Cohen's D
Deepfake myths	13.14 (5.76)	12.60 (6.04)	*t*(461) = .95	.340	.09
NSII Watch Past	2.30 (1.56)	1.90 (1.36)	*t*(370.63) = 2.85	.005	.27
NSII Share Past	1.16 (0.56)	1.13 (0.56)	*t*(461) = .47	.640	.05
NSII Create Past	1.14 (0.62)	1.14 (0.67)	*t*(461) = .01	.995	.00
NSII Watch Future	2.28 (1.37)	1.87 (1.26)	*t*(461) = 3.16	.002	.31
NSII Share Future	1.26 (0.72)	1.20 (0.66)	*t*(461) = .86	.392	.08
NSII Create Future	1.26 (0.74)	1.23 (0.73)	*t*(461) = .47	.636	.05
NSII Future Watch Intentions					
Friend	1.58 (1.15)	1.44 (0.97)	*t*(378.37) = 1.32	.188	.13
Ex	1.64 (1.17)	1.43 (0.96)	*t*(387.67) = 1.92	.550	.19
Stranger	1.97 (1.30)	1.68 (1.15)	*t*(364.82) = 2.48	.014	.23
Celebrity	2.34 (1.46)	1.80 (1.23)	*t*(380.69) = 4.16	<.001	.39
NSII Future Share Intentions					
Friend	1.13 (0.56)	1.10 (0.50)	*t*(461) = .65	.518	.06
Ex	1.16 (0.67)	1.10 (0.530	*t*(461) = .96	.340	.09
Stranger	1.23 (0.67)	1.14 (0.55)	*t*(399.42) = 1.52	.129	.14
Celebrity	1.32 (0.85)	1.16 (0.62)	*t*(421.23) = 2.34	.020	.21
NSII Future Create Intentions					
Friend	1.18 (0.67)	1.17 (0.70)	*t*(461) = .14	.887	.01
Ex	1.23 (0.80)	1.17 (0.69)	*t*(461) = .80	.423	.08
Stranger	1.22 (0.74)	1.20 (0.67)	*t*(461) = .36	.722	.04
Celebrity	1.29 (0.81)	1.20 (0.75)	*t*(356.16) = 1.26	.208	.12

There is still a larger concern to be raised here about how the debrief itself may have been a minorly successful intervention.

7.6.2.3. Qualitative feedback

As previously stated, we asked participants to specifically provide both positive and negative feedback. There was a large amount of positive feedback for the intervention:

- "It was thoughtful and non judgemental, it has the right level of self reflection to it" (41, M, Past or future perpetration reported).

- “It was straight to the point, no sugar coating addressing serious issues” (23, M, No past or future perpetration reported).
- “It wasn't enjoyable but did make me think” (33, M, No past or future perpetration reported)
- “It was effective, especially asking me to draft an apology.” (29, M, No past or future perpetration reported).
- “It presented facts, and made me rethink about the topic” (23, M, Unsure about past or future perpetration reported)
- “it was ok, gets the message across without taking up too much time or being too boring” (29, M, No past or future perpetration reported)
- “I think it was very useful to show that deepfakes can be damaging even if people know they aren't real. Too much of the deepfake discourse focuses on people being unable to tell them apart from reality when that's only part of the issue.” (36, M, No past or future perpetration reported)

In terms of constructive feedback, participants again highlighted the potential of more interactive elements within the intervention. *“maybe include a bit more context or background before each scenario so people understand the bigger picture. A short video or voice clip might also help bring it to life more”* (29, M, No past or future perpetration reported) and specifically visual elements: *“More videos!”* (36, M, No past or future perpetration reported). In particular, the addition of the video was seen as very positive *“N/A - I wouldn't change anything but I would attest that the video of the women sharing their personal accounts was most effective to me”* (32, M, Past watching reported but no future intention reported).

In this iteration of the study we intentionally sought out a higher perpetrating population of younger men, and that may explain the notable increase in negative feedback we observed. A number of participants disagreed with the premise of the intervention and encouraged victim-blaming myths. *“I would rethink the entire thing. Be more specific, compare it to other things we already consider illegal. Make a much stronger case other than "it made someone sad." The*

video at the beginning was probably the most effective part of the intervention, but even that made me roll my eyes at how incredibly dramatic the claims were. If their claims are genuine, at least two of those people have preexisting mental health issues that the deepfakes may have exacerbated but certainly did not cause. In order for me to have empathy for their lives being destroyed, I need concrete statements that the deepfakes made their way through their social circles and left them ostracized by their peers". (40, M, No past perpetration reported but reported unsure about future intention). In the control group a perpetrator noted their remorse but a lack of behavioural control: "*Deepfakes are terrible, regardless of what they were created for. Yet there's an awful side of me that wants to view them all the same*" (40, M, Reported past perpetration but also reported unsure about future intention).

Despite our efforts to encourage the perspective of non-perpetrating bystanders, some survey takers still found the wording of the intervention accusatory. "*less accusing of the survey taker. I know i really didnt need an intervention for it as i do not like deepfakes personally and dont engage with them"* (34, M, No past or future perpetration reported). Some participants disagreed with framing this concern as mainly targeting women (despite the majority of deepfakes posted online targeting women). *"I disagree with particularly women, stats may show that but they go off of reports, women are more likely to report than men are remember that"* (26, M, Reported unsure on past perpetration and no future perpetration reported). There were also examples of victim blaming rhetoric within the feedback "*Reduce emphasis on celebrity figures. Ultimately they have traded a reduced right to privacy compared to the general public. If they were concerned about a library of media being out there, they wouldn't post so routinely on Instagram"* (30, M, Both past and future perpetration reported). There was also concerns that using examples that privileged the perspectives of women would reduce the efficacy of this intervention. *"Maybe a bit of diversity in showing how it can affect men as well would help some people to understand more"* (36, M, No past or future perpetration reported). While one of the examples did involve deepfake NSII being generated of a man and a woman, it did mainly focus on the harms to women as there are no examples of qualitative literature from male

victim-perspectives. *"If I'm honest I felt the intervention wasn't very effective because of the accusatory nature of the scenarios. Being told "You find it funny and share it" is not something I would do and not something I'd find funny to be honest. I also think this focused entirely on how women are hurt by this type of content with the scenario in particular with the fake scenario of the friends hurting two people. To diversify and represent a wider group of both men and women as well as different ethnicities and nationalities might show the broader scope. This could also included people of both right/left political alignments because someone right wing might simply say "Ahh see it's the liberals!" to dismiss the entire intervention and someone left leaning might say "It's the right!" to dismiss it. Showing a wider scope could bridge the gap* (31, M, No past or future perpetration reported). For the most part, the interventions intentionally did not make the ethnicities of any of the vignettes obvious (though the photo of one of the perpetrators and the photo of the celebrity victim in one vignette was shown to be white).

Similarly, participants also believed that the intervention came across as shaming young men "*If your target audience is young men, you need to be careful about coming across as shaming. Young men are so used to being shamed for toxic masculinity that they become desensitized to shame as a persuasive technique and will check out. Empathy is perhaps a better tool"* (36, M, No past or future perpetration reported). In contrast some participants believed that empathy was not a sufficient mechanism: *"I would include less of the victim's stories, at least in a direct manner. I think most Gen Z people that would already engage with deepfake stuff would only double down if they knew how much it hurt people. It's better to make them uncomfortably confront their own actions, than to try to make them feel sympathy for people"* (34, M, No past or future perpetration reported). One participant disagreed entirely with the intervention believing that the voluntary educational vignettes were tantamount to censorship: *"If I made any changes it would be to make this less effective because I fundamentally disagree with censorship"* (29, M, Both past and future perpetration reported). While our quantitative findings in general show a decrease in perpetration, our qualitative

feedback indicates that far more work needs to be done with higher perpetrating populations in future.

7.7 Discussion

The Deepfake/Real Harms intervention is a short, online intervention which successfully reduces deepfake perpetration intentions in the general population and in high-perpetrating samples. For gendered abuse interventions, a successful intervention is defined by three characteristics: a theoretical rationale, evidence of implementation and evidence of effectiveness (Flood, 2014). We have developed a novel intervention based on existing literature and the Theory of Planned Behaviour. We have implemented it online and demonstrated its effectiveness through empirical study. Similarly, the intervention meets DeGue and colleagues (2014) criteria of effective sexual violence intervention programmes as there is a demonstrably positive impact on perpetration behaviours. All research in the study of deepfake abuse online is impacted by significant floor effects in the data as most individuals are concerned about the technology and not interested in creating/sharing deepfakes (Sippy et al., 2024; Twomey, O'Mahony, et al., 2025b). These floor effects meant that there were few significant differences between groups for more extreme behaviours (such as creation of deepfake NSII of friends). However, we did consistently notice a significant drop in the most common forms of perpetration (watching NSII, particularly of celebrities) and we noted a significant drop in the myths at the time of intervention, though that had limited evidence at follow-up.

A strength of the Deepfake/ Real Harms intervention is that it empirically demonstrates the value of psychoeducational interventions targeting technology facilitated gendered violence. Reviews of current bystander interventions in colleges report that only a small number of papers used theoretical models as the basis of their intervention (Evans et al., 2019; Mujal et al., 2021). Similarly, intervention work has been limited by the lack of multi-study designs and follow up data. The samples used were quite large in comparison to those employed by traditional in-

person interventions and the online and short form nature of the work means the intervention is easy to access and complete. For the final intervention iteration, we received very positive feedback from most participants.

In the design of interventions, the theory behind the intervention should be inherently pragmatic: researchers should focus on an intervention working above all else (Chung, 2014). In these terms, the intervention has been a success, though with some limitations. Across the intervention studies, participants tended to score quite low in the various measures of attitudes and behaviours. While studying higher perpetrating samples attempted to reduce these floor effects, they are still of some concern. Similarly, though the difference was not large, there was a significant difference between past watching behaviours between the control group and intervention group across all studies. There are a number of alternative possible reasons for this, including social desirability bias (specifically impression management). Impression management is intentional deception of a research team by participants who use unrealistic positive responses. Impression management is associated with lower self-reports of undesirable behaviours (Perinelli & Gremigni, 2016). Existing work has suggested that social desirability bias may be avoided by presenting information as if there was no clearly correct viewpoint (Labhardt et al., 2017) but it was deemed inappropriate for our work as the mechanism of behavioural change is inherently linked to changing the subjective norms around the social desirability of the behaviour. The significant difference in past behaviour could also be linked to higher perpetrating men or those concerned with the legality of their own behaviours not completing the initial survey.

Similarly, our findings suggest that any reduction in perpetration behaviour intentions start to return closer to baseline over time (though they were still higher than the control group's original values). Decreasing efficacy over time is to be expected from a short form online intervention. There is an extremely limited number of research papers that used randomised groups which use a follow up to measure the decrease in effect over time (DeGue et al., 2014) and these follow ups often find a decrease in the reported attitudinal and behavioural change.

The lack of long-term efficacy is a significant concern in online interventions for gendered violence, which is a trade-off with improving the accessibility of the intervention. It also appears that the debrief for the control group succeeded in reducing intentions and attitudes over time, which while a wrinkle in our evaluation plans, is a positive sign that even very brief educational interventions can have some small effect on attitudes towards deepfake abuse. Drawing from the existing literature on misinformation interventions, a way to improve the long-term efficacy of this intervention may be to follow up with a second "booster" intervention to maintain the effects (Maertens et al., 2021). Similarly, it may make sense to test these follow up effects on a more general population as Prolific participants are essentially professional survey takers and may not engage in the same way as the general population (Ford, 2017; Ternovski & Orr, 2022). Participants who complete a lot of surveys, may also be at risk of taking the intervention less seriously than participants who are directly educated by an authoritative source.

We specifically used a multi-study design to test the intervention both on a general population and a high-perpetrating sample. It is vitally important to study high-perpetrating populations so a broad decrease in the general population does not disguise a small number of people increasing perpetration attitudes and behaviours. It is also necessary to demarcate those who have accidentally been exposed to deepfake NSII online from those who have intentionally chose to watch NSII materials. It should also be noted that this intervention (and deepfake NSII research in general) does not distinguish how prolific different perpetrators are. There is evidence from the mr.deepfakes forum, that the majority of deepfakes online come from 0.1% of deepfake NSII website users (Han et al., 2025). Qualitative feedback from the final round of the intervention on the high-perpetrating sample, highlights how much more needs to be done to challenge extreme perpetrators with this intervention. Work on highly prolific perpetrators is challenged by their anonymity and low prevalence (Twomey, Foley, et al., 2025), but future intervention work which focuses on these perpetrators should be prioritised. An intersection of a number of factors may impact efficacy of gendered violence intervention for perpetrating

populations (McMahon et al., 2020). One such factor which our study did not explore was behavioural control and self-efficacy. One perpetrator noted in the qualitative feedback that they had no control over their own behaviours. Though part of the Theory of Planned Behaviour, behavioural control was not a target of the intervention as there is no identified connection to Deepfake Myth Acceptance. However, for work which aims to specifically study and intervene perpetrating samples, it may be useful to draw from existing literature around problematic pornography use. In particular the encouragement of self-efficacy through cognitive behavioural therapy as a way to decrease perpetration (Antons et al., 2022).

The risk of increasing perpetration behaviours is always a concern with interventions (DeGue et al., 2014). Specifically, previous work has highlighted that educating individuals on rape myths may actually increase participants acceptance of these beliefs where they misremember the myths as facts (Krahé, 2025). Interventions may also potentially be a population's first interaction with novel forms of gendered abuse. Any backfiring effects may not be a consequence of the intervention, but of being made aware of deepfake technology in general. Obviously, participants who were unaware deepfake abuse existed before an intervention may choose to engage in more perpetration behaviours afterwards. Previously unaware participants cannot reduce perpetration which is currently zero. While there was no evidence that this intervention increased perpetration behaviours or deepfake myth acceptance, future research must study the dangers of increasing awareness of deepfake abuse. This study provides some useful insight in that regard, as we report follow-ups ranging from one week to one month, with no evidence of a backfire effect. This is in contrast to interventions aimed specifically at deepfake misinformation, where being made aware of novel media-manipulation technology broadly increases scepticism towards real information (Twomey, O'Mahony, et al., 2025a). However, this may be caused by a significant percentage of the population not knowing what a deepfake is (Internet Matters, 2024; Umbach et al., 2024). As popular understanding of deepfakes and generative AI grows, backfire effects in deepfake interventions will be important to monitor.

7.7.1 Implications

The intervention serves as an example of educational materials on deepfake abuse that may be utilised as part of other intervention programs offered by victims-rights groups. The Deepfakes/ Real harms intervention is open source and can be a model for deepfake related content/components in larger interventions for technology facilitated gendered violence. Perhaps, future interventions could also seek to discuss the common functions of myths across different types of technology facilitated gendered violence. While there are a number of novel myths towards different forms of gendered violence in the literature (Flynn et al., 2023; Harper et al., 2023), they all have the same aims. At their root, myth acceptance is about minimising perpetration, blaming victims and justifying future perpetration. Novel measures and interventions may always fall behind the newest and latest form of gendered abuse and associated victim blaming rhetoric and beliefs.

While we have demonstrated the effectiveness of the intervention in the general population, there are significant outstanding concerns over the need for education and interventions targeting younger school children. 86% of parents and 92% of teenagers in the UK believe that children should be educated about deepfake NSII in schools (Internet Matters, 2024). There have already been serious high-profile incidences of deepfake NSII targeting children. A large deepfake Telegram group in South Korea with 227,000 subscribers was recently uncovered (Bartolozzi et al., 2025). The group featured a large number of deepfake videos targeting ordinary women and girls, including schoolchildren and college aged students. In the South Korean context, 36% of reported deepfakes are of underage victims (Bartolozzi et al., 2025). In the United States of America, there are growing reports of male students producing AI generated NSII of their classmates and sharing them in person and online (Singer, 2024). While only a small number of UK teenagers have reported direct exposure to deepfake NSII (about 5%), exposure is significantly higher for more vulnerable children (16%) (Internet Matters, 2024). The Deepfake/ Real Harms intervention could be changed to an in-person intervention for implementation in schools. The intervention vignettes would have to be

changed to be more appropriate for schoolchildren as currently the vignettes draw from adult victim-survivors. However, this contrasts with the general feedback from participants, who outlined repeatedly in qualitative feedback that the intervention should be more emotive and contain real world examples of consequences towards perpetrators. It would also be necessary to design vignettes focusing on teachers and fellow students and it may be necessary to focus more broadly on the use of AI generated materials in cyberbullying.

Future work should also test how the Deepfake/Real harms intervention interacts with jury decision making process. While there is currently no work that explores the role of deepfake myth acceptance in legal contexts, deepfake myths may similarly correlate with more lenient jury decisions towards perpetrators. Higher rape myth acceptance is associated with more lenient outcomes in juries in rape trials (Lonsway & Fitzgerald, 1994). The design of rape myth intervention for juries has been encouraged in existing literature (Hudspith et al., 2024). In trials, rape myth interventions may be employed before the trial or after jurors have been sworn in (Willmott & Hudspith, 2024) though there is little work testing the efficacy of these interventions. As legal cases involving deepfakes become more common, it may be necessary to design interventions and measures to measure and challenge the damage of deepfake myths in a legal context. The deepfake/Real harms intervention may be especially suited to this as it already follows some of the principles of best practice for rape myth acceptance interventions in jury trials: it is short in length, specifically addresses myths, has an interactive component, and uses an attitude change theory (Willmott & Hudspith, 2024). There is a critical role that forensic/legal psychology and criminology can play in understanding victim-blaming attitudes towards technology facilitated gendered violence.

7.8 Conclusion

The aim of this paper was to develop a short form online intervention that could successfully reduce deepfake myth acceptance and perpetration intentions. The Deepfake/ Real Harms intervention used vignettes, educational materials and perspective taking exercises as a way to reduce NSII-supportive cognition and increase empathy towards victims. We found that the intervention was able to reduce deepfake myth acceptance and common perpetration behavioural intentions (watching deepfake NSII and watching/creating/sharing celebrity NSII). We found that while these effects waned over time as expected, there was still some limited evidence of the intervention's efficacy in follow up samples. The Deepfake/Real harms intervention can be found online and can be downloaded as a handout for offline use.

Chapter 8: Discussion

8.1 Contribution of the research

The contributions of the work can be broken down into contributions of the research itself as well as wider contributions to theory. In terms of the contribution of the research, I had three goals of the thesis: to understand the academic consensus around deepfakes and their novel harms, to provide real world evidence for the harms of deepfake technology in misinformation and abuse imagery, and to contribute towards the future study of perpetration of deepfake NSII. Through a number of papers, I have contributed to furthering my understanding of all three of these goals and created an empirical base for future study.

8.1.1 Scoping current academic consensus on deepfakes

The thesis contributed a broad scale multi-disciplinary scoping review which aimed to critically analyse and make sense of existing issues in the deepfake literature. Firstly, the work critically analysed existing definitions across academic literature and highlighted the importance of mimicked identity to deepfakes, which benefits future researchers who may be similarly confused as to what exactly distinguishes deepfakes from other technology. Secondly, the review analysed what exactly differentiates deepfakes from pre-existing technology. We highlighted that the uses/harms of deepfakes are not new but that the greater accessibility and believability of deepfake videos may be a cause for concern. Finally, the review identified a number of areas of potential work which the thesis was able to directly follow up on. The first of these areas was the novel epistemic harms of deepfakes which we identified as an “epistemic pollution” (Ryan, 2018) and further studied in Chapter 3 in the context of the Russian invasion of Ukraine. We also followed up on that study with a specific commentary piece on how these epistemic harms may lead to increased scepticism in the well-intentioned design of deepfake literacy education and interventions.

I also critiqued the notion of a purely positive use of deepfakes in the thesis which is an approach that influenced my choice not to include positive case studies in the thesis. Taking a critical approach to the positive uses of deepfakes influenced my contribution to a number of papers included in the appendices which critically study the supposedly positive uses of deepfake technology (Ching, Twomey, Linehan, et al., 2025; Murphy et al., 2023; Twomey, Foley, et al., 2025). One final contribution of the review which should be noted is the critique of the lack of focus on abuse imagery outside of legal literature in comparison to the study of misinformation. One of the most significant implications of my review paper is critiquing the idea of deepfakes as having "personal" harms against women and "political" harms against men. We followed up on these findings in Chapter 5 by directly focusing on how the harms of abuse imagery tie into a larger pattern of online harassment of public women (celebrities, politicians and activists)

8.1.2 Real world case studies of the harms of deepfakes

Following up on the lack of empirical evidence in the scoping review, and the broader literature on the potential moral panic of future harms of deepfake technology (Broinowski, 2022), I decided to look empirically at the present harms of deepfake technology. The next two papers aimed to evidence these real-world harms of deepfake technology using case studies. The first of these two case studies aimed to analyse the current online information environment for the effects of deepfake misinformation and epistemic harms. Existing experimental work had identified that deepfakes are as capable as any other modality in inducing false memories and affecting political beliefs/voting intentions (Hameleers et al., 2022; Hameleers & Marquart, 2023; Murphy & Flynn, 2022). However, no qualitative work had aimed to show the damaging effect of false deepfake information to information environments. The Ukraine case study provided a unique insight into the use of deepfakes in conflict and the epistemic harms of the technology. Namely the work contributed to a critique of positive educational deepfakes which were contributing to scepticism around real world videos. The work found more evidence for the misattribution of real evidence as deepfake than people being fooled by deepfakes. In

particular, there were also some concerns from the dataset that technical artifacts in real videos were being identified as evidence of falsity. The increasing scepticism to real videos has major implications for the education of the general public about false information. I coauthored a short follow up piece shortly (Chapter 4) afterwards reflecting on the implications for the design of deepfake interventions and noting the current lack of focus in existing misinformation interventions on increasing false positives (i.e. participants becoming overly sceptical of real media after an intervention).

Chapter 5 sought to analyse deepfake abuse imagery as a growing harm. While the general focus of academic and media discourse has been on deepfake misinformation, scholars have previously pointed out that the most common political harm was in the targeted abuse and harassment of women online using deepfakes (Broinowski, 2022; Maddocks, 2020). While there had been some limited existing qualitative research on the victims of deepfake abuse in the fields of criminology (Flynn et al., 2025; Rousay, 2023), ours was the first to take a HCI approach, focusing specifically on how the traumatic experience of deepfake abuse is linked to the design of the technology and furthered by infrastructures of recourse. I also specifically focused on celebrities who are overwhelmingly the most common targets of deepfake abuse, though the profile of these celebrity targets has changed in recent years to focus more on non-western celebrities such as Korean pop stars (K-pop idols). The paper analysed the public statements of a number of celebrity victims with the aim of understanding their experiences of being “usees” of the technology and their interaction with existing recourses. There were a number of interesting findings with implications across a variety of stakeholders. Mainly, I critiqued the dominant focus on techno-solutionism, legislation and advocacy as recourses. Celebrities identifying experiences with existing time-intensive technical recourses which required payment, legislation which they had to help create and advocacy which led to increased harassment that dissuaded survivors from speaking out. The findings of the paper also illustrated the extent of harassment which celebrity women face online. I noted a number of

victim-blaming and harm-minimising myths they encountered online. These findings directly motivated the final two studies.

8.1.3 Improving psychological solutions and measures for technology facilitated gendered violence

The final two studies in my thesis aimed to make an active contribution to reducing deepfake NSII perpetration through designing a measure and intervention for deepfake abuse perpetration. While previous work had profiled abuser characteristics (such as psychopathy and demographic characteristics) (Fido et al., 2022; Umbach et al., 2024), there was a lack of measures and interventions targeting the beliefs which support perpetration. The contribution of Chapter 6 was in the design of a measure for deepfake abuse myths. The concept of deepfake myths came from the qualitative research in Chapter 5, where victims of deepfake abuse identified online rhetoric which aimed at diminishing their experience and justifying deepfake abuse. We likened the rhetoric to that which is often levied against victims of gendered violence and drew from a large literature which measures acceptance of this rhetoric as attitudes. However, we wanted to distinguish deepfake myths from broader myths around IBSA and gendered abuse by discussing specifically how the falsity of the content itself fuelled dismissive rhetoric and how beliefs towards celebrity women also motivated deepfake abuse.

While Chapters 6 and 7 significantly contribute to the development of measures and interventions aimed at perpetrators of TFGV, the two papers also ultimately levy significant critique at the methodological approach towards measuring and reducing TGFV. The limitations raised in the two chapters are not a weakness of the thesis itself but a weakness in the broader literature concerning technology facilitated gendered abuse and deepfake abuse. As part of the development of the DMAS, I wanted to really consider larger issues with measurements in the literature, namely the lack of psychometrically validated measures, the lack of discussion of floor effects/ multivariate normality and the use of larger scales.

In chapter 7, we developed and tested the Deepfakes/ Real harms intervention. The intervention was short form and online and drew from established theories of behavioural change. I also drew from existing critiques of gendered violence interventions and misinformation interventions as way to test the efficacy of the Deepfake/ Real Harms intervention over time. Overall, I demonstrated that the intervention was effective at reducing these outcome measures in the general and high-perpetrating populations. Though for the most part we found the intervention was effective at reducing self-reported NSII-supportive beliefs and perpetration intention/behaviour, I also levied further similar methodological criticism at the development of interventions. Changing attitudes and behaviours over long periods with short-form interventions is a challenging task. As I based the idea of the intervention off existing rape myth interventions, I was operating under the naïve assumption that these interventions would generally work. I knew also however, that across cognitive interventions targeting misinformation and false beliefs, there are concerns as to the efficacies of interventions, given the potential for backfire effects and beliefs returning to baseline (O'Mahony et al., 2023). I emphasise a healthy critique of the significant issues with measuring attitudes, behaviours and behavioural change relating to deepfake abuse perpetration.

8.2 Contribution of the thesis to theory

There are also a number of contributions of the thesis to the theoretical understanding of the harms of AI technology. Firstly, the work contributes an improved understanding of the harms of technology to people's online identity. The work also reconsiders how classic HCI literature on online identity can tie into the field of technology-facilitated gendered violence. Secondly, the work identifies the benefits of augmenting technological solutions with psychoeducational interventions. Existing deepfake interventions are remedial, offering a possibility of justice and support to victims of deepfake abuse. There also needs to be a focus on preventative interventions and education which aim to reduce the acceptability and perpetration of deepfake abuse imagery among the general public and perpetrators. Finally, we aim to re-evaluate the current focus on the novelty of AI technologies. Instead of focusing on

the “next big thing”, researchers should consider technological innovations along with existing harms which underlie all forms of audiovisual misinformation.

8.2.1 Online identity and moral infraction

> “*When we step through the screen into virtual communities, we reconstruct our identities on the other side of the looking glass.”*
>
> (Turkle, 1995)

The present thesis encourages the application of HCI theory on online identity to the study of gendered violence online and for HCI to reconsider how a user-centred focus on technology ignores other stakeholders targeted by technology. In Chapter 2, I highlighted the importance of mimicked online identity to the definition of deepfakes where the modification and theft of identity is the dominant concern of deepfakes. The novelist Kurt Vonnegut once wrote “*We are what we pretend to be”* (Vonnegut, 1961). In the case of online identity, a person may not be in control of how their online self is portrayed and when the manipulation of identity is put into someone else’s control, it may be of significant concern. The manipulation and targeting of digital identities with deepfake technology constitutes a significant moral infraction. De Ruiter critiqued deepfakes as morally suspect depending on the consent of their target to having their identity manipulated (de Ruiter, 2021). Looking at the practical experience of this loss of control of identity, Chapter 5 highlighted how victims of deepfake abuse felt like they had been put non-consensually into sex work online. In Roussay’s qualitative work, victims also noted a feeling of “becoming pornography” (Rousay, 2023). Both of these examples represent the harms of manipulated online identity, where individuals report the harm of seeing an online version of their identity forced into sexual content without their consent. The concept of the manipulation of identity has motivated the focus of my thesis and I believe informs a theoretical contribution to both HCI and TFGV studies.

HCI researchers should consider how AI manipulations of media fit into long established patterns of novel technology being used to harm women and various marginalised

peoples through attacks on online versions of self. For example, in an older and much less technologically sophisticated context than deepfakes, Sherry Turkle described the experienced of forced sexual interactions with online avatars and the distinct harm of having someone else force a representation of yourself into sexual contexts (Turkle, 1995). This in many ways can be likened to deepfake abuse, where a person finds their online self has been inserted into a sexual encounter without consent. Turkle even describes how "one made light of the offender's actions by saying that the episode was just words", highlighting the emergence of victim blaming myths towards victims of TGFV. Similarly, in Baumer's work, the prototypical Usee is described as a victim of stalkerware, where someone's online activity and identity are monitored by a stalker (Baumer, 2015). Again here, the harm of the technology is exemplified in how a woman's online self is targeted and harassed. In the end these approaches all tie towards a larger theory of Technology Facilitated Gendered Violence and how online selves are targeted by nefarious misogynistic abusers. The study of infractions against online identity has implications for criminological, forensic and feminist psychology which study online harms. Similarly, future HCI work should prioritise post-userism work which focuses on the experience of those who AI technology is used on instead of the creators and "users" of deepfake technology. The present thesis aims to critique the focus in HCI from users as agents of technological change (Oudshoorn & Pinch, 2005). A usee-centred approach to deepfakes is less concerned with the creation and use of tools to create and detect deepfakes, and instead focusing on the development of educational, legislative and technological tools to contribute materially to the safety and privacy of women online.

8.2.2 Augmenting technological solutions with psychoeducational interventions

I believe that socio-technical issues require both technological and social intervention. The existing solutions for deepfake facilitated harms are characterised by an over-focus on detection and removal as opposed to hindering the creation and sharing of abuse imagery online. Existing technological and legal solutions have also been critiqued for their efficacy (Falduti & Tessaris, 2023). Audits of existing reporting systems on X (formerly Twitter) show that abuse

reporting modalities are often widely ineffective (Qiwei et al., 2024). Chapter 5 demonstrated the failure of existing recourses in removing abuse content online and the use of paid services to delegate the constant reporting of abuse imagery online to another party.

There are also emerging concerns that any security vulnerabilities may open up the users of technological solutions to immense harm. For example, image hashing is a commonly used method of reducing the impact of known deepfakes and IBSA online by storing videos as a "hash" and then testing new videos or images to see if they match a hashed example. The benefits of hashing are that they may stop the spread of abuse imagery once identified and existing deepfake abuse charities employ such techniques. However, recent work by Sophie Hawkes and colleagues has demonstrated that current image hashing technology may be vulnerable to cyberattacks which can generate the original image based solely off the hash (Hawkes et al., 2024). If there was a leak of the original hashed dataset, hackers could potentially generate a dataset of decrypted abuse imagery.

While technological solutions, reporting modalities and legislation have a vital place in recourses and justice for deepfake abuse victims, they do not necessarily play a role in reducing perpetration. Instead, targeting perpetrators with educational work can contribute to reducing NSII before the abuse happens and without putting the onus of recourse on victims. Education and intervention are alternative means towards improving the internet and reducing the perpetration of TFGV (Henry et al., 2020). The Deepfake/ Real Harms intervention aimed at reducing deepfake abuse perpetration and perpetration supportive rhetoric/attitudes. As an educational tool, we demonstrated its potential to reduce self-reported perpetration intentions and NSII supportive myths. While the tool is not perfect, it is a much-needed start to designing psychoeducational intervention and education aimed at reducing NSII perpetration. From the previous two chapters, I have aimed at contributing to how we understand the motivation and beliefs which underly NSII perpetration. As a psychologist, I feel that the design of psychoeducational interventions is the best way for the discipline to provide a strong positive contribution to stopping deepfake NSII.

8.2.3 Re-evaluating the focus on “novel” technologies versus established harms

Outside of the impracticalities of how large technological solutions can be implemented across all platforms, a general technological solution to media-editing technologies only works if there is no further technological innovation. The final contribution to theory is the need to understand the common values and harms behind technologies as opposed to focusing on the novelty of technology. It is worth to link back here to how Social Construction of Technology scholars conceive technological development not in the traditional linear model of development but as something which is actively shaped by current and past technologies’ interactions with a number of different groups and problems/solutions they have with technology (Pinch & Bijker, 1984). In Chapter2, I specifically focused on analysing how novel deepfakes are by analysing a variety of academic disciplines and how they approached the problems and solutions of deepfakes. That scoping review was motivated by a lack of academic research on similar photo editing technologies such as photoshop which had overlapping problems and harms. The review made the argument that while deepfakes are indeed more technologically sophisticated and realistic, they are also just a new way of causing old problems.

One case in point is the emergence of purely AI-generated media such as that produced by Large Language Models (LLM) and Stable Diffusion models. From the beginning of my research there was debate as to if these forms of media should be considered as deepfakes, despite often not involving the interpolation of real footage. However, the harms of both have significant overlap in how they may be used to portray individuals authentically. Recently, the X LLM (Grok) was found to be able to generate completely AI-generated nude imagery of celebrity women (BBC, 2025). Currently, these other forms of AI-generated content are far behind the quality of deepfakes. Researchers question whether generative AI will be able to produce AI abuse imagery of an average person or if they mainly rely on the prevalence of data on celebrities and pornographic materials in X’s dataset. However, again, the Grok case shows how novel audio-visual manipulations have the same harms. Similarly, the rise of Large Language Models (LLMs) has led to an increase in chatbots impersonating celebrities and

sending lewd and flirtatious messages to users. Meta (who own Facebook and Instagram) have faced significant controversy over their own LLM impersonating celebrity women and sending AI-generated intimate imagery of both adult and child celebrities (Horwitz, 2025).

Again and again, novel technology is aimed at harming women (often public women) and appropriating their identity to feed whatever a misogynistic online audience requires. As media-manipulation technology develops, we constantly have to ask ourselves how it differs from technology that existed before and whether it makes more sense to try and solve the harms of the technology or the larger social problems which underly them. Returning to a critique of techno-solutionism, is it worth trying to design and legislate for each possible new form of technology or is it more important to look at a wider picture of technology-facilitated gendered violence?

8.3 Future directions

There are a number of areas for future work which would not have been feasible under the scope of the thesis, but I would be interested to returning to in future. Across each chapter, areas of future interest emerged. Firstly, with Chapter 3 and the use of deepfakes in the Russian invasion of Ukraine, I would like to return to the study of secondary online data on conflicts in future. Surprisingly, I have seen almost no research which qualitatively analysed the use of deepfakes in other global conflicts since. Qualitative methodologies may help give more insight into misinformation online than modelling or experimental work alone. I would also like to return to the harms of AI-generated videos to trust in future through quantitative work. My colleague Didier Ching has followed up on the area of political and epistemic harms of deepfakes as part of unpublished work in his PhD thesis which I collaborated on, where he specifically studied the harms of deepfake labelling to trust. In that work we trialled numerous different educational approaches to teaching people about deepfake technology in a way that limited the unintended consequences such as increased anxiety about technology and reduced trust in all media. I think there is important work to be done to establish how we live with

deepfake technology and how we raise understanding of it, as it potentially grows in prominence in society.

While I directly followed on from the deepfake myths which emerged in the qualitative study of celebrity deepfake abuse victim-survivors in Chapter 5, there are a number of other implications of that study on which I would like to carry out future work. Mainly, I would be interested in how to design better solutions catered to influencers and public women. While outside the scope of the thesis, it is essential that future work takes a participatory approach in the design of deepfake abuse recourses. Chapter 5 highlighted the positive effects of celebrity victims of deepfake abuse coming together and organising their response to the news coverage of their abuse. The protective role of communities may also be another way HCI may positively impact solutions for deepfake abuse imagery. While historically fan interactions with celebrities have been viewed as contributing towards the problem of deepfake abuse imagery, the response to deepfake abuse imagery of celebrity women on X also showed that fans were able to reduce the spread of abuse imagery by flooding related hashtags with innocuous content (Riedl & Newell, 2024). Existing work has described how K-pop fans use educational threads, reporting systems and public condemnation as a way of reducing the harms of abuse imagery (Her & Ringland, 2025). Future study of protective and supportive practices online for deepfake NSII victims could be a vital contribution of HCI and would further contribute to augmenting existing technological solutions with social solutions.

For chapters 6 and 7, there are a number of areas of future work I would be really interested in carrying out focusing on the study of deepfake myths across various contexts. The thesis aimed to interact with as broad a number of stakeholders possible, but there are a few specific groups who I would have liked to have worked with more closely such as lawyers, police and journalists. These groups may have unique perspectives on the reduction of deepfake perpetration and myths. The study of deepfake myths in legal contexts would be invaluable for understanding if victim-blaming rhetoric might affect how police and judges enforce new deepfake abuse legislation. I am also very interested in developing psychoeducational

interventions for deepfake misinformation or deepfake cyberbullying in schools. We have seen the rise in cases of deepfakes in schools and targeted towards young people internationally (N. Kelly, 2024). As I have reached the last few months of the PhD, a number of people have personally contacted me and my supervisors to discuss cases of deepfake abuse happening in schools in Ireland. They have also mentioned cases of deepfake abuse between students being swept under the rug by staff. As the concerns of deepfakes being used ins schools grow, the Deepfake/ Real harms provide a vital starting off point for educators which should be further developed and tailored towards unique perpetrating groups.

There are a number of different groups who may benefit from a deepfake myths intervention. One example of a unique group with high perpetration and victimisation which I specifically mention when developing the measure in Chapter 6 is the LGBTQ+ community. I believe existing findings in the literature focus overly on measuring perpetration and are missing out on the wide variety of diversity and experience with deepfakes within the LGBTQ+ community. While higher rates of both victimisation and perpetration in the LGBTQ+ community have been claimed in the literature (Flynn et al., 2022), there is little qualitative work exploring the experiences of LGBTQ+ individuals with deepfakes and none which seeks to concretely understand the interaction between victimisation and perpetration. Similar findings in the literature (where men report higher rates of perpetration and higher rates of victimisation) may be related to perpetrators having more awareness of deepfakes and deepfake websites or having had the experience normalised through their own experience (Flynn et al., 2022). In chapter 5, I outline a need for more granular research which considers the diversity within the LGBTQ+ community when trying to identify why some individuals report higher victimisation/perpetration. Alone, I would not be the right person for that research as I do not have the necessary insider expertise to do that research justice. Specifically, I believe an LGBTQ+ variant of the intervention should be designed with the active participation of members of the LGBTQ+ community. While there is nothing within the Deepfake/Real Harms

intervention which specifically excluded LGBTQ+ individuals, a version of the intervention which draws specifically from LGBTQ+ experience with deepfakes may be more beneficial.

8.4 Limitations

I would like to speak briefly to the limitations of the overall thesis, though most critiques have been previously outlined in the limitation sections of the papers themselves. A broader concern with carrying out a thesis on an emerging technology is that the technology is not "finished" developing. What seemed years away at the start of my PhD (live deepfake videos, high quality AI cloning software and purely AI-generated deepfakes of people) is now reality. By focusing on the real-world harms, I aimed to avoid our papers becoming woefully outdated immediately. I wrote more broadly under the assumption that the harms will increase in future as the technology becomes more prevalent. One issue with studying rapidly developing AI technology is the extent to which people dismiss the evidence of current harms based on hypothetical future advancements. Stuart Reeves noted that technologies which were once considered AI, (such as spellcheck and automatic filtering of spam messages) are no longer considered AI (Reeves, 2022). The concept of AI as more aspirational fantasy than technology. At the start of this thesis, deepfakes and AI images were that aspirational use. Today, one could argue the focus on the potentials of Large Language Models and so-called Artificial General Intelligence (human level intelligence AI) have become the dominant focus (Witschas, 2025).

Just as the technology behind deepfakes has improved, the harmful uses of generative AI have grown massively from the inception of the thesis to the present. While the attention in the literature has partially moved on to novel forms of AI, these harms are what society is left with. If I were to recarry out the same scoping review today, there would be far too many papers to do a similarly detailed analysis and the review itself would be out of date almost immediately. Already the review was published over a year after the searches and took a significant amount of effort. For the two case studies, our research drew from secondary data analysis as there were no direct reports of victims of deepfakes in Ireland at the beginning of the thesis. If I was

starting from scratch again now, I would also have been more likely to carry out interviews with individuals who have actually been affected by deepfake abuse or misinformation.

I have already spoken at length as to the general methodological issues of deepfake perpetration measurement. The study of attitudes and behaviours associated with deepfake abuse is consistently hamstrung by floor effects and non-normal distributions. Approaches to gendered violence measures and interventions have sought to rectify the methodological issues of non-normal data by developing scales which focus on more subtle myths (Gerger et al., 2007). However, researchers have to ask whether it is more important that a measure produces a normal distribution or instead actually reflects non-normal attitudes towards abnormal behaviours. To try and balance measuring unpopular behaviours with methodological concerns around floor effects and multivariate normality, I employed approaches targeting general and highly perpetrating populations. The study of two different populations benefits both the study of general deepfake education as well as specialised perpetrator education. There are also significant concerns over the effectiveness of interventions in specific high-perpetrating individuals who resist education and interventions. I chose to report behavioural and attitudinal measures when reporting the qualitative data in the final study as a way of demonstrating the benefits of mixed methods reporting of intervention data. Unlike traditional gendered violence interventions, the reader gets an insight into certain participants who respond drastically to the intervention and whether that resistance comes from perpetrators or not.

While the broader issues with measurement in the study of deepfake abuse have already been mentioned in my thesis, the current measure of perpetration also should be critiqued. In chapters 6 and 7, we drew from granular measures of perpetration in the literature, breaking it down by watching, sharing and creating across a number of different potential victims. However, on the whole, most people scored quite low with the exception of watching celebrity NSII. There is another broader issue with the wider literature and our work, which is a lack of measurement of frequency of perpetration. It is worth considering that some forms of TFGV

may only be perpetrated by a small fraction of the adult population but that those perpetrators may be prolific in their perpetration.

8.5 Conclusion

The present thesis aimed to used mixed-methods research to broaden our understanding of the harms of deepfake technology and to benefit researchers who seek to measure and prevent these harms. I approached the thesis with three research aims and with the overarching goal of developing a clear consensus and empirical research base on the harms of deepfake technology. My first thesis aim was to understand what deepfakes are, what makes deepfake technology different from what has come before and to parse the potential threats and benefits of the technology through a broad multi-disciplinary scoping review. My second thesis aim was to provide real world evidence for the current harms of deepfake technology across two domains (misinformation and abuse imagery). The final aim of the thesis was to contribute an empirically measure and intervention to aid the psychological study of the most ubiquitous harm of deepfake technology: deepfake abuse imagery. The thesis has contributed to the empirical study of the harms of deepfakes through case studies, psychometric measures and psychoeducational interventions. There are also a number of implications to the theoretical approach to novel media-editing technology: the role of online identity in TGFV, how technological solutions for TFGV should be augmented by social solutions and the risk of focusing on the novelty of technology at the expense of studying pre-existing harms.

References

Acim, B., Boukhlif, M., Ouhnni, H., Kharmoum, N., & Ziti, S. (2025). A Decade of Deepfake Research in the Generative AI Era, 2014–2024: A Bibliometric Analysis. *Publications*, *13*(4), 50.

Act, A. I. (2021). Proposal for a regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. *EUR-Lex-52021PC0206*.

Adair, J., & Senn, C. Y. (2025). The Role Of Rape Myth Acceptance, Situational Context, And Gender In Individual's Perceptions Of Image-Based Sexual Abuse Victims And Perpetrators. *Journal of Interpersonal Violence*, 08862605251319022.

Ahmed, S. (2021a). Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism. *New Media and Society*. Scopus. https://doi.org/10.1177/14614448211019198

Ahmed, S. (2021b). Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. *TELEMATICS AND INFORMATICS*, *57*. https://doi.org/10.1016/j.tele.2020.101508

Ahmed, W., Bath, P. A., & Demartini, G. (2017). Using Twitter as a data source: An overview of ethical, legal, and methodological challenges. *The Ethics of Online Research*, *2*, 79–107.

Ajder, H., Patrini, F., & Cullen, L. (2019). *The State of DeepFakes: Landscape, Threats, and Impact. Deeptrace Labs*.

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, *50*(2), 179–211.

Akerley, S. (2020). Let's Talk about (Fake) Sex Baby: A Deep Dive into the Distributive Harms of Deepfake Pornography. *Ariz. LJ Emerging Tech.*, *4*, i.

Akkaya, F. D. (2024). Deepfake Dilemmas: Imagine Tomorrow's Surveillance Society through Three Scenarios. *Journal of Economy Culture and Society*, *70*, 121–134.

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

Alexander, C. S., & Becker, H. J. (1978). The use of vignettes in survey research. *Public Opinion Quarterly*, *42*(1), 93–104.

Ali, S., Youssef, E., & Qamar, A. (2020). Analyzing the news media practices concerning sexual exploitation of women. *Multicultural Education*, *6*(4), 90–99.

Aliman, N.-M., & Kester, L. (2022). *VR, Deepfakes and Epistemic Security*. 93–98.

Allyn, B. (2022, March 14). Telegram is the app of choice in the war in Ukraine despite experts' privacy concerns. *NPR*. https://www.npr.org/2022/03/14/1086483703/telegram-ukraine-war-russia

Altuncu, E., Franqueira, V. N., & Li, S. (2022). Deepfake: Definitions, Performance Metrics and Standards, Datasets and Benchmarks, and a Meta-Review. *arXiv Preprint arXiv:2208.10913*.

Anna, R. L. (2021). Deepfakes: What Are They, and Why Are They Dangerous? *South Carolina Lawyer*, *33*, 34–39.

Antons, S., Engel, J., Briken, P., Krüger, T. H., Brand, M., & Stark, R. (2022). Treatments and interventions for compulsive sexual behavior disorder with a focus on problematic pornography use: A preregistered systematic review. *Journal of Behavioral Addictions*, *11*(3), 643–666.

Apolo, Y., & Michael, K. (2024). Beyond A Reasonable Doubt? Audiovisual Evidence, AI Manipulation, Deepfakes, and the Law. *IEEE Transactions on Technology and Society*, *5*(2), 156–168.

Arksey, H., & O'Malley, L. (2005). Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology: Theory and Practice*, *8*(1), 19–32. https://doi.org/10.1080/1364557032000119616

Armesto-Larson, B. (2020). Nonconsensual pornography: Criminal law solutions to a worldwide problem. *Or. Rev. Int'l L.*, *21*, 177.

Asadzade, P., & Izadi, R. (2022). The reputational cost of military aggression: Evidence from the 2022 Russian invasion of Ukraine. *Research & Politics*, *9*(2), 20531680221098337.

Ascott, T. (2020). Microfake: How small-scale deepfakes can undermine society. *Journal of Digital Media and Policy*, *11*(2), 215–222. Scopus. https://doi.org/10.1386/jdmp_00018_1

Aylett, M. P., & Vazquez-Alvarez, Y. (2020). *Voice puppetry: Speech synthesis adventures in human centred ai*. 108–109.

Banet-Weiser, S., & Miltner, K. M. (2016). # MasculinitySoFragile: Culture, structure, and networked misogyny. *Feminist Media Studies*, *16*(1), 171–174.

Barnes, C., & Barraclough, T. (2020). Deepfakes and synthetic media. In *Emerging technologies and international security* (pp. 206–220). Routledge.

Baron, L. F., & Gomez, R. (2016). The associations between technologies and societies: The utility of actor-network theory. *Science, Technology and Society*, *21*(2), 129–148.

Bartolozzi, M., Corsaro, S., & Nguyen, V. A. N. (2025). Mobilising Digital Feminism in the Age of Artificial Intelligence: South Korean and Chinese Women against Pornographic Deepfakes. *Peace Human Rights Governance*, 1–40.

Baumer, E. P. (2015). *Usees*. 3295–3298.

Baumer, E. P., & Brubaker, J. R. (2017). *Post-userism*. 6291–6303.

Bazarkina, D. Yu., & Pashentsev, E. N. (2020). Malicious use of artificial intelligence: New psychological security risks in BRICS countries. *Russia in Global Affairs*, *18*(4), 154–177. Scopus. https://doi.org/10.31278/1810-6374-2020-18-4-154-177

BBC. (2022, July 28). False claims of "deepfake" President Biden go viral. *BBC News*. https://www.bbc.com/news/62338593

BBC. (2025, August 8). *Pornographic Taylor Swift deepfakes generated by Musk's Grok AI*. https://www.bbc.com/news/articles/cwye62e1ndjo

Bellini, R., Forrest, S., Westmarland, N., Jackson, D., & Smeddinck, J. D. (2020). *Choice-point: Fostering awareness and choice with perpetrators in domestic violence interventions*. 1–14.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the dangers of stochastic parrots: Can language models be too big? 🦜*. 610–623.

Birrer, A., & Just, N. (2024). What we know and don't know about deepfakes: An investigation into the state of the research and regulatory landscape. *New Media & Society*, 14614448241253138. https://doi.org/10.1177/14614448241253138

Bode, L. (2021). Deepfaking Keanu: YouTube deepfakes, platform visual effects, and the complexity of reception. *Convergence*, *27*(4), 919–934. https://doi.org/10.1177/13548565211030454

Bohner, G., Jarvis, C. I., Eyssel, F., & Siebler, F. (2005). The causal impact of rape myth acceptance on men's rape proclivity: Comparing sexually coercive and noncoercive men. *European Journal of Social Psychology*, *35*(6), 819–828.

Bondarouk, T., & Ruel, H. (2004). Discourse analysis: Making complex methodology simple. *ECIS 2004 Proceedings*, 1.

Braun, V., & Clarke, V. (2006). Qualitative Research in Psychology Using thematic analysis in psychology Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101.

Braun, V., & Clarke, V. (2012). *Thematic analysis.* American Psychological Association.

Braun, V., & Clarke, V. (2021a). Can I use TA? Should I use TA? Should I not use TA? Comparing reflexive thematic analysis and other pattern-based qualitative analytic approaches. *Counselling and Psychotherapy Research*, *21*(1), 37–47.

Braun, V., & Clarke, V. (2021b). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352.

Brieger, A. R. (2024). *Empowerment or exploitation: A qualitative analysis of online feminist communities' discussions of deepfake pornography*.

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.

Broussard, M. (2018). *Artificial unintelligence: How computers misunderstand the world*. mit Press.

Brunk, I. W., & Hakimi, M. (2022). Russia, Ukraine, and the future world order. *American Journal of International Law*, *116*(4), 687–697.

Burt, M. R. (1980). Cultural myths and supports for rape. *Journal of Personality and Social Psychology*, *38*(2), 217.

Caldera, E. (2019). Reject the evidence of your eyes and ears: Deepfakes and the law of virtual replicants. *Seton Hall L. Rev.*, *50*, 177.

Carlson, M. (2021). Skepticism and the Digital Information Environment. *SATS*, *22*(2), 149–167. Scopus. https://doi.org/10.1515/sats-2021-0008

Celli, F. (2020). Deepfakes Are Coming: Does Australia Come Prepared? *Canberra Law Review*, *17*(2), 193–204. a9h.

Chen, J. X., McDonald, A., Zou, Y., Tseng, E., Roundy, K. A., Tamersoy, A., Schaub, F., Ristenpart, T., & Dell, N. (2022). *Trauma-informed computing: Towards safer technology experiences for all*. 1–20.

Chesney, R., & Citron, D. (2019a). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*, *107*, 1753–1820.

Chesney, R., & Citron, D. (2019b). Deepfakes and the New Disinformation War The Coming Age of Post-Truth Geopolitics. *FOREIGN AFFAIRS*, *98*(1), 147–155.

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*, *CSCW2025, 9*(7), 1–23.

Chung, D. (2014). Behavior change programs for intimate partner violence abusers: A means to promote the safety of women and children? In *Critical Issues on Violence Against Women* (pp. 171–182). Routledge.

Citron, D. K. (2019). Sexual privacy. *Yale Law Journal*, *128*(7), 1870–1960. Scopus.

Coker, A. L., Bush, H. M., Cook-Craig, P. G., DeGue, S. A., Clear, E. R., Brancato, C. J., Fisher, B. S., & Recktenwald, E. A. (2017). RCT testing bystander effectiveness to reduce violence. *American Journal of Preventive Medicine*, *52*(5), 566–578.

Coker, A. L., Bush, H. M., Fisher, B. S., Swan, S. C., Williams, C. M., Clear, E. R., & DeGue, S. (2016). Multi-college bystander intervention evaluation for violence prevention. *American Journal of Preventive Medicine*, *50*(3), 295–302.

Coleman, A. (2022, February 25). Ukraine conflict: Further false images shared online. *BBC News*. https://www.bbc.com/news/60528276

Conrad, K. J., Riley, B. B., Conrad, K. M., Chan, Y.-F., & Dennis, M. L. (2010). Validation of the Crime and Violence Scale (CVS) against the Rasch measurement model including differences by gender, race, and age. *Evaluation Review*, *34*(2), 83–115.

Conte, P. (2019). Mockumentality: From hyperfaces to deepfakes. *World Literature Studies*, *11*(4), 11–25. Scopus.

Cooper, L. B., Paluck, E. L., Fletcher, E. K., Ryan, I. M., Branscombe, N., & Center, T. J. (2013). Reducing gender-based violence. *The Sage Handbook of Gender and Psychology*, 359–378.

Cover, R. (2022). Deepfake culture: The emergence of audio-video deception as an object of social anxiety and regulation. *Continuum: Journal of Media & Cultural Studies*, *36*(4), 609–621. a9h.

Cover, R., Henry, N., Gleave, J., Greenfield, S., Grechyn, V., & Huynh, T. B. (2025). Protecting public figures online: How do platforms and regulators define public figures? *Media International Australia*, *196*(1), 156–170.

Cross, C. (2022). Using artificial intelligence (AI) and deepfakes to deceive victims: The need to rethink current romance fraud prevention messaging. *Crime Prevention and Community Safety*, *24*(1), 30–41. psyh. https://doi.org/10.1057/s41300-021-00134-w

Dagar, D., & Vishwakarma, D. K. (2022). A literature review and perspectives in deepfakes: Generation, detection, and applications. *International Journal of Multimedia*

Information Retrieval, *11*(3), 219–289. Scopus. https://doi.org/10.1007/s13735-022-00241-w

de Rancourt-Raymond, A., & Smaili, N. (2022). The unethical use of deepfakes. *Journal of Financial Crime*. Scopus. https://doi.org/10.1108/JFC-04-2022-0090

de Ruiter, A. (2021). The Distinct Wrong of Deepfakes. *Philosophy and Technology*, *34*(4), 1311–1332. Scopus. https://doi.org/10.1007/s13347-021-00459-2

DeGue, S., Valle, L. A., Holt, M. K., Massetti, G. M., Matjasko, J. L., & Tharp, A. T. (2014). A systematic review of primary prevention strategies for sexual violence perpetration. *Aggression and Violent Behavior*, *19*(4), 346–362.

Diakopoulos, N., & Johnson, D. (2021). Anticipating and addressing the ethical implications of deepfakes in the context of elections. *NEW MEDIA & SOCIETY*, *23*(7), 2072–2098. https://doi.org/10.1177/1461444820925811

DiStefano, C., & Morgan, G. B. (2014). A Comparison of Diagonal Weighted Least Squares Robust Estimation Techniques for Ordinal Data. *Structural Equation Modeling: A Multidisciplinary Journal*, *21*(3), 425–438. https://doi.org/10.1080/10705511.2014.915373

Dixon Jr, J. H. B. (2019). Deepfakes: More Frightening Than Photoshop on Steroids. *Judges' Journal*, *58*, 35–37.

Domenteanu, A., Tătaru, G.-C., Crăciun, L., Molănescu, A.-G., Cotfas, L.-A., & Delcea, C. (2024). Living in the age of deepfakes: A bibliometric exploration of trends, challenges, and detection approaches. *Information*, *15*(9), 525.

Dugdale, S., Ward, J., Hernen, J., Elison, S., Davies, G., & Donkor, D. (2016). Using the Behavior Change Technique Taxonomy v1 to conceptualize the clinical content of Breaking Free Online: A computer-assisted therapy program for substance use disorders. *Substance Abuse Treatment, Prevention, and Policy*, *11*(1), 26.

Duquette, H. (2020). Digital Fame: Amending the Right of Publicity to Combat Advances in Face-Swapping Technology. *Journal of High Technology Law*, *20*, 82–118.

Dyer, K. D., & Hall, R. E. (2019). Effect of critical thinking education on epistemically unwarranted beliefs in college students. *Research in Higher Education*, *60*(3), 293–314.

Ebim, M. A. (2022). Putinian Policies of De-Nazification: Discourses of Power in the Deployment of Linguistic Resources on the Russian Invasion of Ukraine. *Putinian Policies of De-Nazification: Discourses of Power in the Deployment of Linguistic Resources on the Russian Invasion of Ukraine (March 13, 2022)*.

Edwards, K. M., Turchik, J. A., Dardis, C. M., Reynolds, N., & Gidycz, C. A. (2011). Rape Myths: History, Individual and Institutional-Level Presence, and Implications for Change. *Sex Roles*, *65*(11–12), 761–773. https://doi.org/10.1007/s11199-011-9943-2

Ekinci, Y., Dam, S., & Buckle, G. (2025). The Dark Side of Social Media Influencers: A Research Agenda for Analysing Deceptive Practices and Regulatory Challenges. *Psychology & Marketing*, *42*(4), 1201–1214. https://doi.org/10.1002/mar.22173

Erin Duffy, B., Ononye, A., & Sawey, M. (2024). The politics of vulnerability in the influencer economy. *European Journal of Cultural Studies*, *27*(3), 352–370. https://doi.org/10.1177/13675494231212346

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

Evans, J. L., Burroughs, M. E., & Knowlden, A. P. (2019). Examining the efficacy of bystander sexual violence interventions for first-year college students: A systematic review. *Aggression and Violent Behavior*, *48*, 72–82.

Fairclough, N. (2013). *Critical discourse analysis: The critical study of language*. Routledge.

Falduti, M., & Tessaris, S. (2023). Mapping the Interdisciplinary Research on Non-consensual Pornography: Technical and Quantitative Perspectives. *Digital Threats: Research and Practice*.

Fallis, D. (2021). The Epistemic Threat of Deepfakes. *Philosophy and Technology*, *34*(4), 623–643. Scopus. https://doi.org/10.1007/s13347-020-00419-2

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*. https://doi.org/10.1016/j.chb.2021.107141

Fiesler, C., & Proferes, N. (2018). "Participant" perceptions of Twitter research ethics. *Social Media+ Society*, *4*(1), 2056305118763366.

Fineman, M. (2012). *Faking it: Manipulated photography before Photoshop*. Metropolitan Museum of Art.

Flood, M. (2014). Current practices to preventing sexual and intimate partner violence. In *Critical issues on violence against women* (pp. 209–220). Routledge.

Flores, S. A., & Hartlaub, M. G. (1998). Reducing rape-myth acceptance in male college students: A meta-analysis of intervention studies. *Journal of College Student Development*, *39*, 438–448.

Flynn, A., Cama, E., Powell, A., & Scott, A. J. (2023). Victim-blaming and image-based sexual abuse. *Journal of Criminology*, *56*(1), 7–25.

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.

Flynn, A., Powell, A., Scott, A. J., & Cama, E. (2022). Deepfakes and Digitally Altered Imagery Abuse: A Cross-Country Exploration of an Emerging form of Image-Based Sexual Abuse. *British Journal of Criminology*, *62*(6), 1341–1358. a9h.

Ford, J. B. (2017). Amazon's Mechanical Turk: A comment. *Journal of Advertising*, *46*(1), 156–158.

Gagliardi, J. (2017). *The Culture of Conspiracy: Evolving Media Technology and Its Relationship to Conspiracy Theories*. State University of New York at Stony Brook.

Gallagher, R. J., Stowell, E., Parker, A. G., & Foucault Welles, B. (2019). Reclaiming stigmatized narratives: The networked disclosure landscape of# MeToo. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–30.

Gamage, D., Ghasiya, P., Bonagiri, V., Whiting, M. E., & Sasahara, K. (2022). Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. *CHI Conference on Human Factors in Computing Systems*, 1–19. https://doi.org/10.1145/3491102.3517446

Gamage, D., Ravinthran, H., & Sasahara, K. (2023). Moral intuitions behind deepfake-related discussions in Reddit communities. *arXiv Preprint arXiv:2303.06216*.

Gamage, D., Sasahara, K., & Chen, J. (2021). The Emergence of Deepfakes and its Societal Implications: A Systematic Review. *TTO*, *1*(1), 28–39.

Gannon, R. (2022). Image-based sexual abuse in Ireland: A thematic and critical discourse analysis of Twitter data surrounding a highly-publicised case in 2013. *DEARCADH: Graduate Journal of Gender, Globalisation and Rights*, *3*, 8–26.

Geddes, K. (2020). Ocularcentrism and Deepfakes: Should Seeing Be Believing? *Fordham Intellectual Property, Media & Entertainment Law Journal*, *31*, 1042–1083.

Geldenhuis, D. L. (2023). *Rape myth acceptance: Investigating the dimensionality of the Illinois rape myth acceptance scale and the male rape myth scale.* [PhD Thesis, Stellenbosch: Stellenbosch University]. https://scholar.sun.ac.za/bitstreams/d3c46d1e-46e9-46b3-b68b-791e2bf6155f/download

Gerger, H., Kley, H., Bohner, G., & Siebler, F. (2007). The acceptance of modern myths about sexual aggression scale: Development and validation in German and English. *Aggressive Behavior: Official Journal of the International Society for Research on Aggression*, *33*(5), 422–440.

Gerstner, C. R., & Farid, H. (2022). Detecting Real-Time Deep-Fake Videos Using Active Illumination. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 53–60.

Gerstner, E. (2020). Face/off: "DeepFake" Face Swaps and Privacy Laws. *Defense Counsel Journal*, *87*, 1–14.

Gilbert, B. J., Heesacker, M., & Gannon, L. J. (1991). Changing the sexual aggression-supportive attitudes of men: A psychoeducational intervention. *Journal of Counseling Psychology*, *38*(2), 197.

Gill, R. (2019). Surveillance is a feminist issue. In *The Routledge handbook of contemporary feminism* (pp. 148–161). Routledge.

GitHub. (n.d.). *Twitter-API-v2-sample-code/Full-Archive-Search at main · twitterdev/Twitter-API-v2-sample-code*. GitHub. Retrieved September 8, 2022, from https://github.com/twitterdev/Twitter-API-v2-sample-code

Giugni, L. (2021). From Individual Perpetrators to Global Mobilisation Strategies: The Micro-Foundations of Digital Violence Against Women. *The Palgrave Handbook of Gendered Violence and Technology*, 75–91.

Godulla, A., Hoffmann, C. P., & Seibert, D. M. A. (2021). Dealing with deepfakes—An interdisciplinary examination of the state of research and implications for communication studies. *Studies in Communication and Media*, *10*(1), 73–96. Scopus. https://doi.org/10.5771/2192-4007-2021-1-72

González-Ibánez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in twitter: A closer look. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 581–586.

Goodman, S. (2017). How to conduct a psychological discourse analysis. *Critical Approaches to Discourse Analysis Across Disciplines*, *9*(2), 142–153.

Google. (2023). *Remove involuntary fake pornography from Google*. Google Search Help. https://support.google.com/websearch/answer/9116649?sjid=12702487589217084664-EU

Gosling, S. D., & Mason, W. (2015). Internet research in psychology. *Annual Review of Psychology*, *66*(1), 877–902.

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.

Gozzi, L. (2024, March 20). Giorgia Meloni: Italian PM seeks damages over deepfake porn videos. *BBC News*. https://www.bbc.com/news/world-europe-68615474

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.

Gregory, S. (2020). Deepfakes and synthetic media: What should we fear? What can we do? *Witness. Https://Blog. Witness. Org/2018/07/Deepfakes*.

Griffin, V. W., Wentz, E., & Meinert, E. (2022). Explaining the why in# WhyIDidntReport: An examination of common barriers to formal disclosure of sexual assault in college students. *Journal of Interpersonal Violence*, *37*(15–16), NP14716–NP14745.

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 OF THE UNITED STATES OF AMERICA*, *119*(1). https://doi.org/10.1073/pnas.2110013119

Guest, G., MacQueen, K. M., & Namey, E. E. (2011). *Applied thematic analysis*. sage publications.

Gutierrez, A. (2023). *Fabricating Misogyny: How Deepfakes Are Weaponized to Maintain Political, Social, and Economic Gender Inequalities*.

Habgood-Coote, J. (2023). Deepfakes and the epistemic apocalypse. *Synthese*, *201*(3), 103. https://doi.org/10.1007/s11229-023-04097-3

Hall, M., Hearn, J., & Lewis, R. (2023). Image-Based Sexual Abuse: Online Gender-Sexual Violations. *Encyclopedia*, *3*(1), 327–339.

Hameleers, M., & Marquart, F. (2023). It's Nothing but a Deepfake! The Effects of Misinformation and Deepfake Labels Delegitimizing an Authentic Political Speech. *International Journal of Communication*, *17*, 21.

Hameleers, M., van der Meer, T., & 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). https://doi.org/10.1177/20563051221116346

Han, C., Li, A., Kumar, D., & Durumeric, Z. (2025). *Characterizing the {MrDeepFakes} Sexual Deepfake Marketplace*. 5169–5188.

Hardy, A., Young, K., & Holmes, E. A. (2009). Does trauma memory play a role in the experience of reporting sexual assault during police interviews? An exploratory study. *Memory*, *17*(8), 783–788. https://doi.org/10.1080/09658210903081835

Harper, C. A., Smith, L., Leach, J., Daruwala, N. A., & Fido, D. (2023). Development and validation of the Beliefs about Revenge Pornography Questionnaire. *Sexual Abuse*, *35*(6), 748–783.

Harris. (2021). Video on demand: What deepfakes do and how they harm. *Synthese*, *199*(5), 13373–13391.

Harris, D. (2018). Deepfakes: False Pornography Is Here and the Law Cannot Protect You. *Duke Law & Technology Review*, *17*, 99–128.

Hawkes, S., Weinert, C., Almeida, T., & Mehrnezhad, M. (2024). Perceptual Hash Inversion Attacks on Image-Based Sexual Abuse Removal Tools. *IEEE Security & Privacy*.

Hawkins, W., Mittelstadt, B., & Russell, C. (2025). Deepfakes on Demand: The rise of accessible non-consensual deepfake image generators. *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 1602–1614. https://doi.org/10.1145/3715275.3732107

Hedling, E. (2024). Gendered Disinformation. In *Feminist Foreign Policy Analysis* (pp. 137–153). Bristol University Press.

Henkhaus, L. (2022). The role of the internet in Ukraine's information war. *Texas A&M Today*.

Henry, N., Flynn, A., & Powell, A. (2020). Technology-facilitated domestic and sexual violence: A review. *Violence against Women*, *26*(15–16), 1828–1854.

Henry, N., & Powell, A. (2015). Embodied Harms: Gender, Shame, and Technology-Facilitated Sexual Violence. *Violence Against Women*, *21*(6), 758–779. https://doi.org/10.1177/1077801215576581

Henry, N., & Powell, A. (2018). Technology-facilitated sexual violence: A literature review of empirical research. *Trauma, Violence, & Abuse*, *19*(2), 195–208.

Her, J., & Ringland, K. E. (2025). *Mitigating Deepfake Harm in Online Communities: Insights from the BTS ARMY Fandom*. 1–6.

Hines, D. A., & Reed, K. M. P. (2017). Bystander Prevention of Sexual and Dating Violence: An Experimental Evaluation of Online and In-Person Bystander Intervention Programs. *Partner Abuse*, *8*(4).

Holliday, C. (2021). Rewriting the stars: Surface tensions and gender troubles in the online media production of digital deepfakes. *Convergence*, *27*(4), 899–918.

Hoppenstedt, M. (2022, August 11). Franziska Giffey: Russische Komiker zeigen Ausschnitt von Gespräch mit Fake-Klitschko. *Der Spiegel*. https://www.spiegel.de/netzwelt/web/franziska-giffey-russische-komiker-zeigen-ausschnitt-von-gespraech-mit-fake-klitschko-a-527ab090-2979-4e70-a81c-08c661c0ef62

Horwitz, J. (2025, August 29). Exclusive: Meta created flirty chatbots of Taylor Swift, other celebrities without permission. *Reuters*. https://www.reuters.com/business/meta-created-flirty-chatbots-taylor-swift-other-celebrities-without-permission-2025-08-29/

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hudspith, L. F., Wager, N., Willmott, D., & Gallagher, B. (2023). Forty years of rape myth acceptance interventions: A systematic review of what works in naturalistic institutional settings and how this can be applied to educational guidance for jurors. *Trauma, Violence, & Abuse*, *24*(2), 981–1000.

Hudspith, L. F., Wager, N., Willmott, D., & Gallagher, B. (2024). The impact of rape myth education on jury decision-making: A systematic review. *Trauma, Violence, & Abuse*, *25*(5), 4062–4077.

Hughes, R., & Huby, M. (2004). The construction and interpretation of vignettes in social research. *Social Work and Social Sciences Review*, *11*(1), 36–51.

Husák, M., Laštovička, M., & Plesník, T. (2022). Handling Internet Activism during the Russian Invasion of Ukraine: A Campus Network Perspective. *Digital Threats: Research and Practice*.

Hwang, Y., Ryu, J., & Jeong, S. (2021). Effects of Disinformation Using Deepfake: The Protective Effect of Media Literacy Education. *CYBERPSYCHOLOGY BEHAVIOR AND SOCIAL NETWORKING*, *24*(3), 188–193. https://doi.org/10.1089/cyber.2020.0174

Internet Matters. (2024). *The new face of digital abuse: Children's experiences of nude deepfakes*. https://web.archive.org/web/20250514234319/https://www.internetmatters.org/wp-content/uploads/2024/11/Childrens-experiences-of-nude-deepfakes-research-1.pdf. https://www.internetmatters.org/wp-content/uploads/2024/11/Childrens-experiences-of-nude-deepfakes-research-1.pdf

Jane, E. A. (2016). Online misogyny and feminist digilantism. *Continuum*, *30*(3), 284–297.

Jegham, N., Abdelatti, M., Koh, C. Y., Elmoubarki, L., & Hendawi, A. (2025). How hungry is ai? Benchmarking energy, water, and carbon footprint of llm inference. *arXiv Preprint arXiv:2505.09598*.

Jenkins, K. (2017). Rape myths and domestic abuse myths as hermeneutical injustices. *Journal of Applied Philosophy*, *34*(2), 191–205.

Jones, R. B., Thapar, A., Stone, Z., Thapar, A., Jones, I., Smith, D., & Simpson, S. (2018). Psychoeducational interventions in adolescent depression: A systematic review. *Patient Education and Counseling*, *101*(5), 804–816.

Karnouskos, S. (2020). Artificial Intelligence in Digital Media: The Era of Deepfakes. *IEEE Transactions on Technology and Society*, *1*(3), 138–147. https://doi.org/10.1109/tts.2020.3001312

Katchanovski, I. (2016). The separatist war in Donbas: A violent break-up of Ukraine? *European Politics and Society*, *17*(4), 473–489.

Katz, J. (2014). Engaging men in prevention of violence against women. In *Critical Issues on Violence Against Women* (pp. 233–243). Routledge.

Kelly, L., & Westmarland, N. (2014). New approaches to assessing effectiveness and outcomes of domestic violence perpetrator programs. In *Critical issues on violence against women* (pp. 183–194). Routledge.

Kelly, N. (2024). *South Korea: The deepfake crisis engulfing hundreds of schools*. *BBC NEWS*. https://www.bbc.com/news/articles/cpdlpj9zn9go

Kerner, C., & Risse, M. (2021). Beyond Porn and Discreditation: Epistemic Promises and Perils of Deepfake Technology in Digital Lifeworlds. *MORAL PHILOSOPHY AND POLITICS*, *8*(1), 81–108. https://doi.org/10.1515/mopp-2020-0024

KhosraviNik, M. (2014). Critical discourse analysis, power, and new media (digital) discourse. *Why Discourse Matters: Negotiating Identity in the Mediatized World*, 283–301.

Kim, E., & Muralidharan, S. (2020). The role of empathy and efficacy in public service announcements: Using narratives to induce bystander intervention in domestic violence. *Journal of Advertising Research*, *60*(4), 452–466.

Kim, H., Hwang, Y., Lee, J., Kwon, Y., Park, Y., & Lee, J. (2022). *Personalization trade-offs in designing a dialogue-based information system for support-seeking of sexual violence survivors*. 1–15.

Kleinsasser, A., Jouriles, E. N., McDonald, R., & Rosenfield, D. (2015). An online bystander intervention program for the prevention of sexual violence. *Psychology of Violence*, *5*(3), 227.

Korshunov, P., & Marcel, S. (2021). *Subjective and objective evaluation of deepfake videos*. *2021-June*, 2510–2514. https://doi.org/10.1109/ICASSP39728.2021.9414258

Krahé, B. (2025). Beware the truth effect: Why efforts to debunk rape myths may backfire. *Journal of Criminal Justice*, *98*(C).

Krishna, D. (2020). Deepfakes, Online Platforms, and a Novel Proposal for Transparency, Collaboration, and Education. *Richmond Journal of Law & Technology*, *27*, 1–74.

Kugler, & Pace, C. (2021). Deepfake Privacy: Attitudes and Regulation. *Northwestern University Law Review*, *116*, 611–680.

Kuzio, T. (2018). Euromaidan revolution, Crimea and Russia–Ukraine war: Why it is time for a review of Ukrainian–Russian studies. *Eurasian Geography and Economics*, *59*(3–4), 529–553.

Kuzio, T. (2023). Imperial nationalism as the driver behind Russia's invasion of Ukraine. *Nations and Nationalism*, *29*(1), 30–38.

Kwok, A. O. J., & Koh, S. G. M. (2021). Deepfake: A social construction of technology perspective. *Current Issues in Tourism*, *24*(13), 1798–1802. https://doi.org/10.1080/13683500.2020.1738357

Labhardt, D., Holdsworth, E., Brown, S., & Howat, D. (2017). You see but you do not observe: A review of bystander intervention and sexual assault on university campuses. *Aggression and Violent Behavior*, *35*, 13–25.

Lamb, H. (2020). Can you believe your eyes?[Issues of Deepfakes]. *Engineering & Technology*, *15*(6), 18–21.

Langa, J. (2021). Deepfakes, real consequences: Crafting legislation to combat threats posed by deepfakes. *BUL Rev.*, *101*, 761.

Lange-Ionatamishvili, E., Svetoka, S., & Geers, K. (2015). Strategic communications and social media in the Russia Ukraine conflict. *Cyber War in Perspective: Russian Aggression against Ukraine*, 103–111.

Laufer, D. (n.d.). *Falscher Klitschko: Was gegen ein Deepfake spricht | tagesschau.de*. Retrieved September 8, 2022, from https://web.archive.org/web/20220630144757/https://www.tagesschau.de/investigativ/rbb/deep-fake-klitschko-101.html

Lawless, B., & Chen, Y.-W. (2019). Developing a method of critical thematic analysis for qualitative communication inquiry. *Howard Journal of Communications*, *30*(1), 92–106.

Lee, C. P., Hong, R., Jiang, H. H., Plotnik, A., Agnew, W., & Morgenstern, J. H. (2025). *How do data owners say no? A case study of data consent mechanisms in web-scraped vision-language AI training datasets*. NeurIPS 2025 Workshop on Regulatable ML.

Lee, Y., Huang, K.-T. (Tim), Blom, R., Schriner, R., & Ciccarelli, C. A. (2021). To believe or not to believe: Framing analysis of content and audience response of top 10 deepfake videos on youtube. *Cyberpsychology, Behavior, and Social Networking*. psyh. https://doi.org/10.1089/cyber.2020.0176

Liu, Y., Nakatsuka, Y., Sani, A. A., Agarwal, S., & Tsudik, G. (2022). Vronicle: Verifiable provenance for videos from mobile devices. *Proceedings of the 20th Annual International Conference on Mobile Systems, Applications and Services*, 565–566.

Liv, N., & Greenbaum, D. (2020). Deep Fakes and Memory Malleability: False Memories in the Service of Fake News. *AJOB Neuroscience*, *11*(2), 96–104. a9h.

Locke, A., & Budds, K. (2020). Applying critical discursive psychology to health psychology research: A practical guide. *Health Psychology and Behavioral Medicine*, *8*(1), 234–247. https://doi.org/10.1080/21642850.2020.1792307

Lonsway, K. A., & Fitzgerald, L. F. (1994). Rape myths: In review. *Psychology of Women Quarterly*, *18*(2), 133–164.

Ma'arif, A., Maghfiroh, H., Suwarno, I., Prayogi, D., Lonang, S., & Sharkawy, A.-N. (2025). Social, legal, and ethical implications of AI-Generated deepfake pornography on digital platforms: A systematic literature review. *Social Sciences & Humanities Open*, *12*, 101882.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84.

Maddocks, S. (2020). 'A Deepfake Porn Plot Intended to Silence Me': Exploring continuities between pornographic and 'political' deep fakes. *Porn Studies*, *7*(4), 415–423. Scopus. https://doi.org/10.1080/23268743.2020.1757499

Maertens, R., Roozenbeek, J., Basol, M., & van der Linden, S. (2021). Long-term effectiveness of inoculation against misinformation: Three longitudinal experiments. *Journal of Experimental Psychology: Applied*, *27*(1), 1.

Mahmud, B. U., & Sharmin, A. (2021). Deep insights of deepfake technology: A review. *arXiv Preprint arXiv:2105.00192*.

Mania, K. (2024). Legal protection of revenge and deepfake porn victims in the European Union: Findings from a comparative legal study. *Trauma, Violence, & Abuse*, *25*(1), 117–129.

Marples, D. R. (2021). *The War in Ukraine's Donbas: Origins, Contexts, and the Future*. Central European University Press.

Martin, N. (2021). Image-based sexual abuse and deepfakes: A survivor turned activist's perspective. *The Palgrave Handbook of Gendered Violence and Technology*, 55–72.

Marwick, A. E. (2017). Scandal or sex crime? Gendered privacy and the celebrity nude photo leaks. *Ethics and Information Technology*, *19*(3), 177–191. https://doi.org/10.1007/s10676-017-9431-7

May, A. (2017). *Pseudoscience and science fiction*. Springer.

Mayring, P. (2004). Qualitative content analysis. *A Companion to Qualitative Research*, *1*(2), 159–176.

McCosker, A. (2022). Making sense of deepfakes: Socializing AI and building data literacy on GitHub and YouTube. *New Media and Society*. Scopus. https://doi.org/10.1177/14614448221093943

McCurry, J. (2024, August 28). South Korea battles surge of deepfake pornography after thousands found to be spreading images. *The Guardian*. https://www.theguardian.com/world/article/2024/aug/28/south-korea-deepfake-porn-law-crackdown

McGlynn, C., Johnson, K., Rackley, E., Henry, N., Gavey, N., Flynn, A., & Powell, A. (2021). 'It's torture for the soul': The harms of image-based sexual abuse. *Social & Legal Studies*, *30*(4), 541–562.

McGlynn, C., Rackley, E., & Houghton, R. (2017). Beyond 'revenge porn': The continuum of image-based sexual abuse. *Feminist Legal Studies*, *25*, 25–46.

McKay, D., & Miller, C. (2021). *Standing in the way of control: A call to action to prevent abuse through better design of smart technologies*. 1–14.

McMahon, S., Burnham, J., & Banyard, V. L. (2020). Bystander intervention as a prevention strategy for campus sexual violence: Perceptions of historically minoritized college students. *Prevention Science*, *21*(6), 795–806.

McPeak, A. (2020). The Threat of Deepfakes in Litigation: Raising the Authentication Bar to Combat Falsehood. *Vanderbilt Journal of Entertainment & Technology Law*, *23*, [i]-450.

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., Eccles, M. P., Cane, J., & Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: Building an international consensus for the reporting of behavior change interventions. *Annals of Behavioral Medicine*, *46*(1), 81–95.

Microsoft. (2023). *Microsoft-Non-consensual Intimate Imagery Reporting Form*. https://www.microsoft.com/en-ie/concern/nonconsensualintimateimagery

Millière, R. (2022). Deep learning and synthetic media. *Synthese*, *200*(3). Scopus. https://doi.org/10.1007/s11229-022-03739-2

Minnaar, A. (2023). An examination of early international and national efforts to combat online child pornography and child sexual exploitation and abuse material on the Internet. *Child Abuse Research in South Africa*, *24*(2), 1–26.

Moitra, A., Ahmed, S. I., & Chandra, P. (2021). Parsing the'Me'in# MeToo: Sexual Harassment, Social Media, and Justice Infrastructures. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW1), 1–34.

Molina, S. E. (2021). Lying Beneath the Surface: The Impacts of Deepfake Technology on the Privacy and Safety of the LGBTQ+ Community Articles and Surveys. *Nova Law Review*, *46*, 251–276.

Montano, D. E., & Kasprzyk, D. (2015). Theory of reasoned action, theory of planned behavior, and the integrated behavioral model. *Health Behavior: Theory, Research and Practice*, *70*(4), 231.

Morse, J. M. (2000). Determining sample size. *Qualitative Health Research*, *10*(1), 3–5.

Mou, N. L., Dufour, G. K., & Jackson, D. L. (2025). Structural equation modelling in male sexual violence against women research: Best practices in statistical reporting and interpretation. *Journal of Sexual Aggression*, 1–22. https://doi.org/10.1080/13552600.2025.2471770

Mujal, G. N., Taylor, M. E., Fry, J. L., Gochez-Kerr, T. H., & Weaver, N. L. (2021). A systematic review of bystander interventions for the prevention of sexual violence. *Trauma, Violence, & Abuse*, *22*(2), 381–396.

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). https://doi.org/10.1002/acp.70061

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., & Flynn, E. (2022). Deepfake false memories. *Memory*, *30*(4), 480–492. Scopus. https://doi.org/10.1080/09658211.2021.1919715

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

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, *154*. Scopus. https://doi.org/10.1016/j.jbusres.2022.113368

Nema, P. (2021). Understanding copyright issues entailing deepfakes in India. *INTERNATIONAL JOURNAL OF LAW AND INFORMATION TECHNOLOGY*, *29*(3), 241–254. https://doi.org/10.1093/ijlit/eaab007

Nissen, A., Conrad, C., & Newman, A. (2023). *Are you human? Investigating the perceptions and evaluations of virtual versus human Instagram influencers*. 1–14.

Nour, N., & Gelfand, J. (2022). Deepfakes: A Digital Transformation Leads to Misinformation. *Conference Papers: International Conference on Grey Literature*, *23*, 55–65. a9h.

Odunze, D. (2018). Cyber victimization by hackers: A criminological analysis. *Public Policy and Administration Research*, *8*(01), 08–15.

Öhman, C. (2020). Introducing the pervert's dilemma: A contribution to the critique of Deepfake Pornography. *Ethics and Information Technology*, *22*(2), 133–140.

Oltermann, P. (2022, June 25). European politicians duped into deepfake video calls with mayor of Kyiv. *The Guardian*. https://www.theguardian.com/world/2022/jun/25/european-leaders-deepfake-video-calls-mayor-of-kyiv-vitali-klitschko

O'Mahony, C., Brassil, M., Murphy, G., & Linehan, C. (2023). The efficacy of interventions in reducing belief in conspiracy theories: A systematic review. *Plos One*, *18*(4), e0280902.

Oudshoorn, N., & Pinch, T. (2005). *How users matter: The co-construction of users and technology*. MIT press.

Paris, B., & Donovan, J. (2019). *Deepfakes and cheap fakes*.

Park, H., Forlizzi, J., & Lee, J. (2022). *Voices of sexual assault survivors: Understanding survivors' experiences of interactional breakdowns and design ideas for solutions*. 485–503.

Pascale, E. (2023). Deeply Dehumanizing, Degrading, and Violating: Deepfake Pornography and the Path to Legal Recourse. *Syracuse L. Rev.*, *73*, 335.

Patrick, S. (2019). *Leaked Sex and Damaged Goods: News Media Framing of Illicit and Stolen Celebrity Images*.

Patrick, S. (2022). More than a sex crime: A feminist political economy of the 2014 iCloud hack. *Media, Culture & Society*, *44*(1), 39–55.

Payne, D. L., Lonsway, K. A., & Fitzgerald, L. F. (1999). Rape myth acceptance: Exploration of its structure and its measurement using theIllinois rape myth acceptance scale. *Journal of Research in Personality*, *33*(1), 27–68.

Pearson, J., & Zinets, N. (2022, March 17). Deepfake footage purports to show Ukrainian president capitulating. *Reuters*. https://www.reuters.com/world/europe/deepfake-footage-purports-show-ukrainian-president-capitulating-2022-03-16/

Peiser, J. (n.d.). J.K. Rowling thought she was talking to Zelensky. It was Russian pranksters. *Washington Post*. Retrieved September 8, 2022, from https://www.washingtonpost.com/nation/2022/06/24/jk-rowling-russian-prank-zelensky/

Perinelli, E., & Gremigni, P. (2016). Use of social desirability scales in clinical psychology: A systematic review. *Journal of Clinical Psychology*, *72*(6), 534–551.

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C. S., RP, L., & Jiang, J. (2020). DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv Preprint arXiv:2005.05535*.

Persson, S., & Dhingra, K. (2022). *Rape myths: Understanding, assessing, and preventing*. Emerald Publishing Limited.

Pfefferkorn, R. (2019a). "Deepfakes" in the Courtroom. *Boston University Public Interest Law Journal*, *29*, 245–276.

Pfefferkorn, R. (2019b). Too Good to Be True. *NWLawyer*, *73*, 22–25.

Pinch, T. J., & Bijker, W. E. (1984). The social construction of facts and artefacts: Or how the sociology of science and the sociology of technology might benefit each other. *Social Studies of Science*, *14*(3), 399–441.

Potter, J. (1996). *Discourse analysis and constructionist approaches: Theoretical background*. https://www.academia.edu/download/50361076/Sessao11_Richardson_Handbook_Chapter.pdf

Potter, J., & Wetherell, M. (2004). Discourse analysis. *Handbook of Data Analysis*, 607–624.

Powell, A., Henry, N., Flynn, A., & Scott, A. J. (2019). Image-based sexual abuse: The extent, nature, and predictors of perpetration in a community sample of Australian residents. *Computers in Human Behavior*, *92*, 393–402.

Powell, A., Scott, A. J., Flynn, A., & Henry, N. (2020). *Image-based sexual abuse: An international study of victims and perpetrators–A Summary Report.*

Pumphrey, J. A. (2024). *Deepfake Media, Cognitive Biases, and US Presidential Elections*.

Qiwei, L., Zhang, S., Kasper, A. T., Ashkinaze, J., Eaton, A. A., Schoenebeck, S., & Gilbert, E. (2024). Reporting Non-Consensual Intimate Media: An Audit Study of Deepfakes. *arXiv Preprint arXiv:2409.12138*.

Rackley, E., McGlynn, C., Johnson, K., Henry, N., Gavey, N., Flynn, A., & Powell, A. (2021). Seeking justice and redress for victim-survivors of image-based sexual abuse. *Feminist Legal Studies*, *29*(3), 293–322.

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access*, *10*, 25494–25513.

Razi, A., Seberger, J. S., Alsoubai, A., Naher, N., De Choudhury, M., & Wisniewski, P. J. (2024). Toward Trauma-Informed Research Practices with Youth in HCI: Caring for Participants and Research Assistants When Studying Sensitive Topics. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–31. https://doi.org/10.1145/3637411

Reddit (u/HauntingFly). (2022, March 14). *The ending we all want to see:* [Reddit Post]. R/Gifs. www.reddit.com/r/gifs/comments/tdm7it/the_ending_we_all_want_to_see/

Reddit (u/reizuki). (2022, March 4). *Be wary of deep fakes! An example of what the Ukraine government warning is about: Putin announces result of peace talks (eng sub)* [Reddit Post]. R/Ukraine. www.reddit.com/r/ukraine/comments/t67ho9/be_wary_of_deep_fakes_an_example_of_what_the/

Reeves, S. (2022). Navigating Incommensurability Between Ethnomethodology, Conversation Analysis, and Artificial Intelligence. *arXiv Preprint arXiv:2206.11899*.

Reface. (2022, March 17). *🇺🇦 Reface Stands with Ukraine. How We Work in a Wartime*. Reface Blog. https://blog.reface.ai/reface-stands-with-ukraine/

Reich, C. M., Pegel, G. A., & Johnson, A. B. (2022). Are survivors of sexual assault blamed more than victims of other crimes? *Journal of Interpersonal Violence*, *37*(19–20), NP18394–NP18416.

Reis, F., de Matos, M. G., & Ferreira, P. (2017). *The Impact of DNS Blocks on Digital Piracy Activity.* ICIS.

Reuters. (2022, March 17). Fact Check-Doctored video appears to show Putin announcing peace. *Reuters*. https://www.reuters.com/article/factcheck-putin-address-idUSL2N2VK1CC

Reynolds, R. M. (2020). When Seeing Is No Longer Believing: Deep Fakes, the Right of Publicity, and CDA 230 Reform. *American University Intellectual Property Brief*, *12*, 1–16.

Riedl, M. J., & Newell, A. (2024). Reporting Image-Based Sexual Violence: Deepfakes,# ProtectTaylorSwift, and Platform Responsibility. *Available at SSRN 4919928*.

Rini, R. (2020). Deepfakes and the epistemic backstop. *Philosophers Imprint*, *20*(24), 1–16. Scopus.

Rizzica, A. (2021). *Sexually explicit deepfakes: To what extent do legal responses protect the depicted persons*.

Robinson, S., Buckley, J., Ciolfi, L., Linehan, C., McInerney, C., Nuseibeh, B., Twomey, J., Rauf, I., & McCarthy, J. (2024). Infrastructural Justice for Responsible Software Engineering. *Journal of Responsible Technology*, 100087.

Rogers, P. (2024). Best practices for your confirmatory factor analysis: A JASP and lavaan tutorial. *Behavior Research Methods*, *56*(7), 6634–6654. https://doi.org/10.3758/s13428-024-02375-7

Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on the psychological theory of "inoculation" can reduce susceptibility to misinformation across cultures. *Harv Kennedy Sch Misinformation Rev 2020b*, *1*.

Roozenbeek, J., Young, D., & Madsen, J. K. (2025). The wilful rejection of psychological and behavioural interventions. *Current Opinion in Psychology*, 102138.

Rottweiler, B., & Gill, P. (2021). *Measuring individuals' misogynistic attitudes: Development and validation of the misogyny scale*. https://osf.io/preprints/psyarxiv/6f829/

Rousay, V. (2023). *Sexual Deepfakes and Image-Based Sexual Abuse: Victim-Survivor Experiences and Embodied Harms*.

Ryan, S. (2018). Epistemic environmentalism. *Journal of Philosophical Research*, *43*, 97–112.

Schewe, P. A., & O'Donohue, W. (1996). Rape prevention with high-risk males: Short-term outcome of two interventions. *Archives of Sexual Behavior*, *25*(5), 455–471.

Schwartz, M. D. (2000). Methodological issues in the use of survey data for measuring and characterizing violence against women. *Violence against Women*, *6*(8), 815–838.

Scott, C. F., Marcu, G., Anderson, R. E., Newman, M. W., & Schoenebeck, S. (2023). *Trauma-informed social media: Towards solutions for reducing and healing online harm*. 1–20.

Shaffer, V. A., Bohanek, J., Focella, E. S., Horstman, H., & Saffran, L. (2019). Encouraging perspective taking: Using narrative writing to induce empathy for others engaging in negative health behaviors. *PloS One*, *14*(10), e0224046.

Shahid, F., Kamath, S., Sidotam, A., Jiang, V., Batino, A., & Vashistha, A. (2022). *" It Matches My Worldview": Examining Perceptions and Attitudes Around Fake Videos*. 1–15.

Sharma, M., & Kaur, M. (2022). A review of Deepfake technology: An emerging AI threat. *Soft Computing for Security Applications*, 605–619.

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. Scopus. https://doi.org/10.1080/21670811.2022.2026797

Shoaib, M. R., Wang, Z., Ahvanooey, M. T., & Zhao, J. (2023). Deepfakes, misinformation, and disinformation in the era of frontier AI, generative AI, and large AI models. *2023 International Conference on Computer and Applications (ICCA)*, 1–7. https://ieeexplore.ieee.org/abstract/document/10401723/

Sims-Schouten, W., Riley, S. C. E., & Willig, C. (2007). Critical Realism in Discourse Analysis: A Presentation of a Systematic Method of Analysis Using Women's Talk of

Motherhood, Childcare and Female Employment as an Example. *Theory & Psychology*, *17*(1), 101–124. https://doi.org/10.1177/0959354307073153

Singer, N. (2024). Teen girls confront an epidemic of deepfake nudes in schools. *The New York Times*, *8*.

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* (No. arXiv:2407.05529). arXiv. https://doi.org/10.48550/arXiv.2407.05529

Skilling, K., & Stylianides, G. J. (2020). Using vignettes in educational research: A framework for vignette construction. *International Journal of Research & Method in Education*, *43*(5), 541–556.

Søe, S. O. (2021). *A unified account of information, misinformation, and disinformation*. 22.

Sriwimon, L., & Zilli, P. J. (2017). Applying Critical Discourse Analysis as a conceptual framework for investigating gender stereotypes in political media discourse. *Kasetsart Journal of Social Sciences*, *38*(2), 136–142.

Stokes, G., Sutcliffe, K., & Thomas, J. (2023). Is a one-size-fits-all '12-month rule'appropriate when it comes to the last search date in systematic reviews? *BMJ Evidence-Based Medicine*, *28*(6), 359–363.

Streiner, D. L. (2003). Starting at the beginning: An introduction to coefficient alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99–103.

Suslavich, B. T. (2023). Nonconsensual Deepfakes: A" Deep Problem" for Victims. *Alb. LJ Sci. & Tech.*, *33*, 160.

Suvarna, A., Bhalla, G., Kumar, S., & Bhardwaj, A. (2020). *Identifying victim blaming language in discussions about sexual assaults on twitter*. 156–163.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: Learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, *36*(4), 1–13.

Sylvester, S. (2020). 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*, *72*, 369–392.

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

Tang, T. L.-P., & Tang, T. L.-N. (2010). Finding the Lost Sheep: A Panel Study of Business Students' Intrinsic Religiosity, Machiavellianism, and Unethical Behavior Intentions. *Ethics & Behavior*, *20*(5), 352–379. https://doi.org/10.1080/10508422.2010.491763

Taylor, B. (2021). Defending the state from digital Deceit: The reflexive securitization of deepfake. *CRITICAL STUDIES IN MEDIA COMMUNICATION*, *38*(1), 1–17. https://doi.org/10.1080/15295036.2020.1833058

Ternovski, J., & Orr, L. (2022). A note on increases in inattentive online survey-takers since 2020. *Journal of Quantitative Description: Digital Media*, *2*.

Thomas, S. (2024). AI and actors: Ethical challenges, cultural narratives and industry pathways in synthetic media performance. *Emerging Media*, *2*(3), 523–546.

Thombre, M. (2021). Deconstructing Deepfake: Tracking Legal Implications and Challenges. *Issue 4 Int'l JL Mgmt. & Human.*, *4*, 2267.

Timmerman, B., Mehta, P., Deb, P., Gallagher, K., Dolan-Gavitt, B., Garg, S., & Greenstadt, R. (2023). Studying the Online Deepfake Community. *Journal of Online Trust and Safety*, *2*(1).

Tong, A., Sainsbury, P., & Craig, J. (2007). Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care*, *19*(6), 349–357.

Trottier, D., Benbouriche, M., & Bonneville, V. (2021). A meta-analysis on the association between rape myth acceptance and sexual coercion perpetration. *The Journal of Sex Research*, *58*(3), 375–382.

Tseng, E., Sabet, M., Bellini, R., Sodhi, H. K., Ristenpart, T., & Dell, N. (2022). *Care infrastructures for digital security in intimate partner violence*. 1–20.

Turkle, S. (1995). *Life on the Screen: Identity in the Age of the Internet*.

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. https://doi.org/10.1371/journal.pone.0291668

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*. https://ieeexplore.ieee.org/abstract/document/10756226/

Twomey, J., Ching, D., O'Mahony, C., Linehan, C., & Murphy, G. (2025). *I Was Never in the AI. RA: an Experimental Study on the Use of Manipulated Audio for the Censorship of Political Dissidents*. 1–8.

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* (No. arXiv:2507.13065). arXiv. https://doi.org/10.48550/arXiv.2507.13065

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

Twomey, J., O'Mahony, C., Linehan, C., & Murphy, G. (2025b). *The Development and Validation of the Deepfake Myth Acceptance Scale (DMAS)*. PsyArXiv. https://doi.org/10.31234/osf.io/6qn7y_v1

Tyagi, K. (2023). *Deepfakes, copyright and personality rights an inter-disciplinary perspective*. 191–210.

Ukraine / Україна [@Ukraine]. (2022, April 21). *Putin telling truth, huh? Well, we tried to imagine what he'd say if he did. Created by ADC*UA, Gvardiya Production House and AI startup Reface. @reface_app @Gvardiya_ph* 👏 *#ArmUkraineNow https://t.co/6P0Fz3h4uR* [Tweet]. Twitter. https://twitter.com/Ukraine/status/1517119052904374272

Ullrich, Q. J. (2022). Is This Video Real? The Principal Mischief of Deepfakes and How the Lanham Act Can Address It. *Columbia Journal of Law and Social Problems*, *55*(1), 1–56. Scopus.

Umbach, R., Henry, N., Beard, G. F., & Berryessa, C. M. (2024). *Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries*. 1–20.

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), Article 1. https://doi.org/10.1177/2056305120903408

van der Sloot, B., & Wagensveld, Y. (2022). Deepfakes: Regulatory challenges for the synthetic society. *Computer Law and Security Review*, *46*. Scopus. https://doi.org/10.1016/j.clsr.2022.105716

Varnhagen, C. K., McFall, G. P., Pugh, N., Routledge, L., Sumida-MacDonald, H., & Kwong, T. E. (2010). Lol: New language and spelling in instant messaging. *Reading and Writing*, *23*, 719–733.

Vasist, P. N., & Krishnan, S. (2022a). Deepfakes: An Integrative Review of the Literature and an Agenda for Future Research. *Communications of the Association for Information Systems*, *51*, 556–602. bth.

Vasist, P. N., & Krishnan, S. (2022b). Engaging with deepfakes: A meta-synthesis from the perspective of social shaping of technology theory. *Internet Research*. Scopus. https://doi.org/10.1108/INTR-06-2022-0465

Vincent, J. (2021, April 30). *'Deepfake' that supposedly fooled European politicians was just a look-alike, say pranksters*. The Verge. https://www.theverge.com/2021/4/30/22407264/deepfake-european-polticians-leonid-volkov-vovan-lexus

Viola, M., & Voto, C. (2023). Designed to abuse? Deepfakes and the non-consensual diffusion of intimate images. *Synthese*, *201*(1), 30.

Vonnegut, K. (1961). *Mother night*. Gold Medal Books. https://go.exlibris.link/1jhdtTth

Wahl-Jorgensen, K., & Carlson, M. (2021). Conjecturing fearful futures: Journalistic discourses on deepfakes. *Journalism Practice*, *15*(6), 803–820.

Wang, S., & Kim, S. (2022). Users' emotional and behavioral responses to deepfake videos of K-pop idols. *Computers in Human Behavior*, *134*. Scopus. https://doi.org/10.1016/j.chb.2022.107305

Wasi, A. T., Priti, R. N., Khan, M. A., Rahman, A., & Islam, M. R. (2025). Seeing Isn't Believing: Addressing the Societal Impact of Deepfakes in Low-Tech Environments. *arXiv Preprint arXiv:2508.16618*.

Weikmann, T., & Lecheler, S. (2022). Visual disinformation in a digital age: A literature synthesis and research agenda. *New Media & Society*, 1. a9h.

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

Whittaker, L., Kietzmann, T. C., Kietzmann, J., & Dabirian, A. (2020). "All around me are synthetic faces": The mad world of ai-generated media. *IT Professional*, *22*(5), 90–99. Scopus. https://doi.org/10.1109/MITP.2020.2985492

Whittaker, L., Letheren, K., & Mulcahy, R. (2021). The Rise of Deepfakes: A Conceptual Framework and Research Agenda for Marketing. *Australasian Marketing Journal*, *29*(3), 204–214. https://doi.org/10.1177/1839334921999479

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.

Widder, D. G., Nafus, D., Dabbish, L., & Herbsleb, J. (2022). Limits and Possibilities for "Ethical AI" in Open Source: A Study of Deepfakes. *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2035–2046. https://doi.org/10.1145/3531146.3533779

Wieberneit, M., Thal, S., Clare, J., Notebaert, L., & Tubex, H. (2024). Silenced Survivors: A Systematic Review of the Barriers to Reporting, Investigating, Prosecuting, and

Sentencing of Adult Female Rape and Sexual Assault. *Trauma, Violence, & Abuse*, *25*(5), 3742–3757. https://doi.org/10.1177/15248380241261404

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*(1), 407–432.

Williams, M. C., Williams, S. E., & Newby, D. E. (2023). Artificial Intelligence–based Text-to-Image Generation of Cardiac CT. *Radiology: Cardiothoracic Imaging*, *5*(2), e220297.

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. *Sociology*, *51*(6), 1149–1168.

Willig, C. (2013). *Introducing Qualitative Research In Psychology*. McGraw-Hill Education (UK). https://books.google.com/books?hl=en&lr=&id=E-lhuM-pNV8C&oi=fnd&pg=PR7&ots=JUj89tHPE-&sig=KNlA8zHC7qmcMI7JdPeIJfJP-H4

Willmott, D., & Hudspith, L. (2024). Jury trials and rape myth bias: Exploring the research evidence, stakeholder perspectives and effective solutions. In *Contemporary Challenges in the Jury System* (pp. 167–188). Routledge.

Winter, R., & Salter, A. (2020). DeepFakes: Uncovering hardcore open source on GitHub. *Porn Studies*, *7*(4), 382–397. https://doi.org/10.1080/23268743.2019.1642794

Wise, A. (2025, May 6). Major deepfake porn site shuts down. *NPR*. https://www.npr.org/2025/05/06/nx-s1-5388422/mr-deepfakes-porn-site-ai-shut-down

Witschas, A. (2025). *Prefabricated futures AI imaginaries between elitist visions and social justice claims*. 335–356.

Wöhler, L., Zembaty, M., Castillo, S., & Magnor, M. (2021). *Towards understanding perceptual differences between genuine and face-swapped videos*. 1–13.

Xue, J., & Lin, K. (2022). Chinese university students' attitudes toward rape myth acceptance: The role of gender, sexual stereotypes, and adversarial sexual beliefs. *Journal of Interpersonal Violence*, *37*(5–6), 2467–2486.

Yadlin-Segal, A., & Oppenheim, Y. (2021). Whose dystopia is it anyway? Deepfakes and social media regulation. *CONVERGENCE-THE INTERNATIONAL JOURNAL OF RESEARCH INTO NEW MEDIA TECHNOLOGIES*, *27*(1), 36–51. https://doi.org/10.1177/1354856520923963

Yang, Y., & Liang, X. (2013). Confirmatory factor analysis under violations of distributional and structural assumptions. *International Journal of Quantitative Research in Education*, *1*(1), 61. https://doi.org/10.1504/ijqre.2013.055642

Yeater, E. A., & O'Donohue, W. (1999). Sexual assault prevention programs: Current issues, future directions, and the potential efficacy of interventions with women. *Clinical Psychology Review*, *19*(7), 739–771.

Appendix I: I Was Never in the AI.R.A: an Experimental Study on the Use of Manipulated Audio for the Censorship of Political Dissidents

Link to published article: https://dl.acm.org/doi/10.1145/3706599.3716238



I Was Never in the AI.R.A: an Experimental Study on the Use of Manipulated Audio for the Censorship of Political Dissidents

John Twomey
School of Applied Psychology
University College Cork
Cork, Ireland
LERO, the Science Foundation Ireland
Research Centre for Software
Limerick, Ireland
118447782@umail.ucc.ie

Didier Ching
School of Applied Psychology
University College Cork
Cork, Ireland
LERO, the Science Foundation Ireland
Research Centre for Software
Limerick, Ireland
116378223@umail.ucc.ie

Cian O'Mahony
School of Applied Psychology
University College Cork
Cork, Munster, Ireland
LERO, the Science Foundation Ireland
Research Centre for Software
Limerick, Ireland
115526303@umail.ucc.ie

Conor Linehan
School of Applied Psychology
University College Cork
Cork, Ireland
LERO, the Science Foundation Ireland
Research Centre for Software
Limerick, Ireland
conor.linehan@ucc.ie

Gillian Murphy
School of Applied Psychology
University College Cork
Cork, Ireland
LERO, the Science Foundation Ireland
Research Centre for Software
Limerick, Ireland
gillian.murphy@ucc.ie

Abstract

During the conflict in Northern Ireland known as "the Troubles", audio recordings of 10 controversial political groups were banned by the British government from being broadcast. Interestingly, video recordings of these people were instead overdubbed by paid actors, implying that the political figures' original voices were somehow more convincing, legitimate or sexy than the actors. This study uses AI-generated audio to examine whether this censorship method affects opinions of political dissidents. We carried out a between-subjects experimental study ($n = 636$), where we compared three different types of AI-generated vocal manipulations with unedited footage of two well-known Northern Irish politicians. We found a significant negative effect of voice manipulation on perceived charisma, legitimization and credibility. Participants also voiced strong disapproval and concern at the idea of using AI to revoice historical footage. We discuss implications of AI assisted audio and video translation of politically sensitive media.

CCS Concepts

• **Empirical studies in HCI**; • **Social aspects of security and privacy**; • **Psychology**;

Keywords

Censorship, GenAI Adams, LondonPsychology,

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
CHI EA '25, Yokohama, Japan
© 2025 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1395-8/2025/04
https://doi.org/10.1145/3706599.3716238

ACM Reference Format:
John Twomey, Didier Ching, Cian O'Mahony, Conor Linehan, and Gillian Murphy. 2025. I Was Never in the AI.R.A: an Experimental Study on the Use of Manipulated Audio for the Censorship of Political Dissidents. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26–May 01, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3706599.3716238

1 Introduction

Recent advances in the quality and accessibility of audio and video manipulation technologies have made the regular creation and dissemination of edited "fake" videos by many actors a reality. Academics and media commentators have pointed out the democratizing potential of AI powered media editing tools. High quality media editing was previously only available to skilled professionals with expensive hardware and tools, whereas now can be carried out using free to download smartphone applications [25]. The majority of academic discourse has focused on identifying the potential dangers of such democratization. For example, concerns have been raised that such videos could be used for spreading effective misinformation regarding elections and referenda and generally carrying out smear campaigns of politicians [26]. However, history tells us that new developments in information technology, in the long term, typically serve to centralize power, and support the interests of corporations and capital more than those of dissidents, separatists, freedom fighters or political agitators. In this paper we focus on an application of generative AI technology that has been previously unexplored in the HCI literature; state censorship of political dissidents.

Media editing has always been a part of the process of political censorship within domestic and international conflict [19]. In this paper, we draw on a bizarre historical case of media censorship in our own country, Ireland, as a model of how traditional and AI fueled media editing technologies can be used for novel and creative

state censorship practices. We carry out and report the findings of a psychological experiment that asks whether subtly manipulating the voices of political speakers can delegitimize their message and undermine their natural magnetism. In this paper we also explore how the same technology could be used in a democratizing manner; in restoring archival footage that has previously been edited and censored. The democratizing of dubbing technologies may facilitate activists who seek to challenge censorship and erasure of the voices of victims of active ethnic cleansing and genocides.

2 Related work

The Troubles" in Northern Ireland describe a conflict that took place from 1969 until 1998, between one group of people (the Unionists) who believe that Northern Ireland should remain in the United Kingdom, and another group (the Republicans) who believe that the island of Ireland should consist of a single nation. The Republican cause is defined by struggles for the voting rights and freedom to demonstrate of the marginalized Catholic communities, as well as the suppression of the native Irish language and culture. On the other side, the Unionist community is concerned with the potential erasure of their freedoms under a Catholic Ireland and were further fueled by the Republic of Ireland's strict and repressive Catholic government at the time of the Troubles.

The conflict has roots in Ireland's history of colonization by the British Empire, in various stages, over a period of 800 years, and the subsequent War of Independence that allowed for the creation of an independent Irish State. However, the Troubles itself emerged from the shutting down of a Civil Rights movement in Northern Ireland by UK authorities and a subsequent descent into inter-group violence [5]. The violence in the North of Ireland was characterized by numerous high-profile acts of political violence by Unionist and Republican paramilitary groups, as well as by the British police force and military. In particular, the bombings carried out by the Irish Republican Army in British cities often characterize popular perception of the conflict. Much academic research has considered the morality, history and ethics of this conflict and the lives it destroyed. [5] [15]. However, little work has considered its implications for human-computer interaction.

Our work draws inspiration from a peculiar law enacted by the UK government during the Troubles that banned audio of members of 10 political groups from TV and radio (mostly focusing on Sinn Féin, an Irish political party concerned with achieving Irish independence from the United Kingdom), notably targeting the president of Sinn Féin, Gerry Adams. The ban was put in place to deprive violent paramilitary groups and political dissidents of airtime. The motivation of media censorship as stated by British prime minister Margaret Thatcher was to deprive political dissidents of the "oxygen of publicity" [18]. While the ban was intended to censor footage entirely, the wording allowed for dubbed versions of original footage to be broadcast [13]. There was an equivalent but much stricter ban in the Republic of Ireland which ran for almost 30 years and included the state broadcaster RTE claiming it would not allow a Sinn Fein member on TV to advertise a bar of soap [12] [21].

For reasons that we do not fully understand, the prevailing method through which British broadcasters chose to comply with



Figure 1: The Day Today - comedy sketch about the IRA revoicing

this censorship law consisted of over-dubbing Irish republican figures with different voices, while the text of the message was kept the same. It has been suggested by some commentators that this method of censorship (like similar "security" measures [20]) counterintuitively increased support for the speakers. Instead of hearing a nervous and defensive representative being pushed on challenging questions, audiences were treated to the calm authoritative voice of a professional actor smoothly and charmingly explaining the political motivations behind dissident and loyalist violence [3]. The practice was famously lampooned in "The Day Today" (fig. 1)., a British parody news show, which suggested members of Sinn Fein be forced to inhale helium before speaking to reduce the credibility of their voices [2]. In other popular culture around the conflict, it has been suggested that Gerry Adam's voice was dubbed over because his voice was "too sexy" and thus was unfairly persuasive [11]. Surprisingly, there has been a lack of empirical work testing this hypothesis and demonstrating the effectiveness of traditional and AI overdubbing for political censorship of dissidents.

Our work tests the effectiveness of AI powered voice overdubbing by; 1) dubbing original videos of northern Irish political figures Gerry Adams and Ian Paisely, 2) testing how the original is perceived versus the dubbed voice, 3) testing the affect other vocal manipulations/qualities have on its persuasiveness, and 4) accounting for whether prior familiarity with the politicians' effects this. We have carried out a small demonstrative study where we show users interview footage with either an original voice, a deepfake dub, or a mocking/humorous voice (e.g. squeaky like a mouse or demon). We measure the agreement of participants towards the statement made by the speaker, and how persuasive they find the voice. We also control for participants' pre-existing opinions on the politicians. Our Null hypothesis is that there is no difference in persuasiveness between the three conditions.

Of course, there is some previous research that explores the effect of different voice qualities on perception of the speaker. For example, prosody has been linked with credibility in political speech [6]. Specific aspects of speech such as (prosody and volume for example) are associated with perceptions of charisma [17] [22]. Little work has been done approaching how these findings apply to censorship practices which use media editing technology to

manipulate voices. Lots of work focuses on specific characteristics of voices such as pitch, prosody and tone – our work situates this in a real-world example drawing from the Troubles as a real-life application of this work.

This paper proudly follows a rich tradition of HCI research that inserts generative AI into places it does not belong. We speculatively ask whether it is "good" to use AI in restoring historical footage of controversial historical conflicts. Specifically, existing reviews of generative AI technologies such as deepfakes have highlighted that many of the supposedly positive uses of these technologies in historical education materials have not been demonstrated empirically [25]. We explore the potential implications of restoring the voices of separatist leaders in previously censored archival footage.

Our research questions are as follows:

Do manipulated vocal characteristics affect the perceived credibility, legitimacy and charisma of historical politicians and their statements?

How do viewers feel about the potential of using generative AI to revoice existing censored footage?

2.1 A special note to readers

With regard to the broader application of this work; we must acknowledge that there is none. Our research is focused specifically on Irish and British conflicts and should be judged solely within that sphere. We make no claims around the effects of political censorship and counterterrorism. The Irish experience has little relevance to broader global political affairs, as our history is quite specific and unlike any other violence in the world today- a former colony of the British Empire divided along religious grounds between the settler-colonizer and indigenous populations - the establishment of the resulting two state system contributing to a vicious cycle of terrible political violence and unrest on both sides- defined by many powerful figures of the British Empire including William Churchill, David Lloyd George, Lord Mountbatten and Arthur Balfour – a conflict defined by unnecessary death and violence, organized acts of revenge, ethnic cleansing, land seizures and the suppression and destruction of cultural identities.

3 Methods

3.1 Ethics and preregistration

This project received ethical clearance from our institution's ethical board as part of a series of projects on the potentials of deepfake media in political disinformation. We obtained consent for participants prior to the study and gave them the option of withdrawing this consent before submitting their data. This study was also preregistered (available online here: https://aspredicted.org/m6gt-g4d4.pdf).

3.2 Participants

We recruited 636 valid participants from Britain and Ireland using the website Prolific. We did not include participants who did not fully complete the survey or who failed the attention checks. For the attention checks we asked participants a simple question about the content of the video and removed those who answered incorrectly for either of the two politicians. We gathered participants from the United Kingdom and Ireland. While we didn't specifically collect location as demographic data – the majority of participants on Prolific were British. Three participants withdrew their consent at the end of the study and their data was not analyzed. We regathered participants to make up for those who had failed one of the attention checks. In total 25 people failed the Ian Paisely attention checks and 36 failed the Gerry Adams attention checks. We also regathered participants who reported being unable to hear one or more of the conditions. We had 702 distinct participants in total in the dataset. The majority of the participants described themselves as female (n = 384), followed by male (n = 312). Four participants selected non-binary/third gender. Another self-described as bi-gender and one participant chose not to say. Participants had an average age of 43 (SD = 13.91).

3.3 Materials

3.3.1 Videos. For both of our target politicians, we created four versions of the same brief interview clip – original video, AI generated dubbed voice with similar Northern Irish accent, AI generated dubbed voice with an accent approximating received pronunciation (RP – an accent associated with upper-class England notable for its historical use in television news broadcasts) and a high-pitched voice similar to an individual who had inhaled helium. The motivation for the dubbed voice with a similar accent was to emulate the methods used during the historical censorship. The received pronunciation (RP) condition was meant to simulate newsreader voices, the type which translated political speech is often translated for viewers using. The high-pitch conditions were selected as ways that broadcasters might seek to delegitimize speakers – by casting them as unserious (high-pitch). The two target politicians were Gerry Adams and Ian Paisley. Gerry Adams was chosen as he is the most well-known target of the broadcasting ban and one of the figureheads of the Republican side of the Troubles. Ian Paisley was chosen as he could be considered a similarly high profile and controversial Unionist figure. The specific clips we chose both speak to the differing interpretations of the two politicians on the idea of peace in the North and how they saw themselves as advocating for self-defense and non-violence (fig. 2).

The AI dubbed voice was generated using play.ht, an online voice cloning software. For Gerry Adams, the voice was generated based on a recording of the voice actor who was used to overdub Gerry Adams during the ban. However, Ian Paisley was never overdubbed – for his AI generated dubbed replacement we picked Jim Alister, a unionist politician from the same part of Northern Ireland as Ian Paisley. For the received pronunciation voice, we used narakeet, an online AI voice generator and we chose a British accent which most closely resembled received pronunciation. For the high-pitched condition, we raised the pitches by 35 per cent in both politicians as we found this value was a good balance between being noticeable but not unintelligible. All materials are available on our OSF repository (https://osf.io/jvgx2/?view_only=1bcf196ad3094e498a3dc98a6e76fc1c).

3.3.2 Outcome Measures. While we set out to measure the role of audio dubbing in the perceived sexiness of political dissidents, there are in fact few robust measures for "sexiness" and it is hard to account for the impact of sexuality and sexual preference. As a result, we chose to measure charisma as a way of measuring the



Figure 2: Transcripts of the two authentic videos used in this study, Participants saw a version of either video, with either unaltered audio, or video that was doctored in one of three ways.

magnetism of these politicians. The Generalised Charisma Inventory [24] is a 6-item psychometric scale with 3 items measuring influence (e.g. knows how to lead a group) and 3 items measuring affability (e.g. "smiles at people often"). The scale is measured on a five-point Likert scale. We also used measures of political legitimacy and credibility drawing from existing deepfake literature [4]. We measured political legitimacy by asking "Based on the video you just watched, please indicate how you feel about this politician by indicating to what extent you agree with the following statements." (1 = completely disagree, 7 = completely agree)", with items such as "This politician sells out their principles". Participants were asked to rate the credibility of the video by their agreement with certain descriptors on the same scale. (E.g. this video is: Informative (1 = completely disagree, 7 = completely agree)).

3.4 Procedure.

After providing informed consent and demographic information, participants were shown three videos in a random order: one of the four versions of the Gerry Adams video, one of the four versions of the Ian Paisley video, and a control video of the former British Prime Minister Margaret Thatcher. After each video, participants completed an attention check and then rated the charisma, legitimacy and credibility of the video. After they were debriefed, participants were then told about the purpose of the study and shown an original dubbed video of Gerry Adams that was broadcast during the ban. We also showed them a redubbed version we made using deepfake technology, and asked participants to reflect on how they felt about this use of AI technology.

For the revoiced Gerry Adams we used the same voice cloning software on the original audio that we used for the control condition. This generated our Gerry Adams audio clone. We manually matched this to the original video with the intent to match the lips of the original video as we felt using lip-syncing technology would defeat the purpose of "restoring" the original video. This process was carried out iteratively with multiple versions of each sentence trialed to capture the cadence and also his accent. We were limited in our choice for the dubbed audio by the few recordings of dubbed Gerry Adams that exist online (this was the most significant effect of the broadcasting ban: it reduced coverage of Sinn Fein altogether due to extra hassle of revoicing footage) [13]. This task was perhaps the most labor intensive of the materials to produce, and inevitably also the poorest quality.

3.5 Positionality.

As an Irish research team, we acknowledge the sensitive issue of political violence in Ireland and Britain. The authors do not make any claims on being pro-terrorism or censorship but simply encourage HCI to realize that political agitators are a unique stakeholder group who often get ignored by traditional user-focused HCI literature. Though we as HCI researchers may ignore how people will use technology for political violence across the globe [1], it won't stop this from happening. Similarly, while HCI has made some

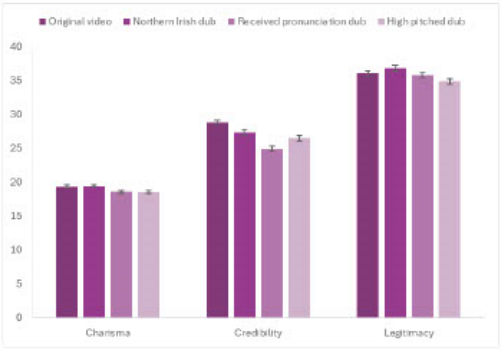


Figure 3: Perceived charisma, legitimacy and credibility of the target politicians, as a function of audio condition.

significant and laudable contributions to the field of censorship by media giants and authoritarian states, we urge readers to further consider how generative AI may revolutionize political censorship, especially for social media companies who are already party to genocides worldwide [14]. As part of this positionality statement, we also must disclose that our university was built by the British government during the state-sponsored famine of the 1840's which killed a quarter of the Irish population and forced another quarter to emigrate. One university building is also built on the bones of 13 IRA prisoners of war who were interned and executed during the Irish war of Independence. While we plan to visit the mass grave to pay respects, we cannot make any claims as to what they would have thought about the use of deepfake technology in political conflicts.

4 Results

We first checked whether there was any evidence that Gerry Adams was particularly magnetic, relative to our other politician, Ian Paisely, as has been claimed [11]. We found that in fact, in every condition of our experiment, Gerry Adams was rated as marginally but significantly less charismatic (M = 18.44, SD = 4.33) than Ian Paisley (M = 19.50, SD = 3.88), $t(1347) = 4.74$, $p < .001$, at least in the eyes of our (mostly British) sample.

4.1 Do manipulated vocal characteristics affect the credibility, legitimacy, and perceived charisma of politicians and their statements?

As per our preregistration, we conducted 3 one-way between groups ANOVAs to find out whether the perceived charisma, legitimization and credibility of political statements differed between our four conditions. The perceived charisma differed significantly across all four conditions $F(3, 1345) = 4.579$, $p = .003$. The Northern Irish dub (M = 19.42, SD = 4.08) and the original video (M = 19.35, SD =3.93) were viewed as more charismatic than the received pronunciation (M =18.57, SD = 4.07) and high pitched (M = 18.97, SD = 4.41) conditions (see fig. 3). Post-hoc Tukey tests indicated that the latter two conditions were significantly different from the first two.

We also found significant differences in perceived legitimacy between the conditions, $F(3, 1345) = 4.08$, $p = .009$, The AI dubbed video with the Northern Ireland voice was rated the highest (M = 36.77, SD = 7.69), followed by the original condition (M = 35.99, SD = 7.35) and received pronunciation dub (M = 35.77, SD = 6.76). The high-pitched voice was seen as the least legitimate (M = 34.81, SD = 7.57). Post-hoc Tukey tests indicated that there were significant differences between the Northern Irish dub and high-pitched conditions.

There were also significant results for the credibility ratings $F(3,1345) = 16.275$, $p < .001$. Unsurprisingly the unedited footage was seen as the most credible (M = 28.80, SD = 6.32), followed by the dubbed footage with a similar accent (M = 27.33, SD = 7.92) and the high pitched voice (M = 26.46, SD = 7.58) For this condition the received pronunciation dub scored lowest on average (M = 24.88, SD = 7.78) perhaps reflecting how it is a more obvious manipulation. Post-hoc Tukey tests indicated that the unedited condition was significantly more credible than the high pitched and received pronunciation dubs.

We carried out a robustness check by using an ANCOVA which showed that these results were stronger when controlling for participants pre-existing support for the target politicians (further details on this are available on our OSF repository).

4.2 How do viewers feel about the potential of using generative AI to revoice existing censored footage?

For our second research question, we asked the participants for their opinion of our completely novel1 and groundbreaking AI revoiced Gerry Adams and compared their responses and found an overwhelmingly negative response to the potentials of generative AI (fig. 4). We discuss these findings in further detail and

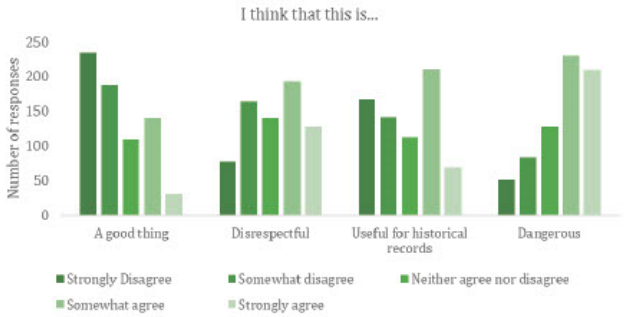


Figure 4: Opinions on the use of AI revoiced audio of Gerry Adams

also present the qualitative data which often contextualizes their responses.

When asked as to whether they considered the use of AI revoicing a good thing on a scale of one to 5, participants had mixed feelings. Even participants who saw the revoicing manipulation as a good thing (M = 2.36, SD = 1.25) usually added caveats around the potential misuse of the technology and the controversial nature of the footage/conflict: *"I think it's good to help recreate the past for educational purposes but only if it's made clear it is AI. Also, it should not be used to manipulate, misinform or misrepresent"*. Some participants noted that the restored footage matched the original lip movements better. *"The AI generated voice gives more credence as it fits with speech pattern better. We should be able to look back at history - whether we agree with the politics of the time, it is still important to learn from it, and to give a voice to those involved"*.

> *"It was very hard to spot the difference I wouldn't approve in these circumstances due to the person being involved and his reputation. There are times when AI dubbed interviews is a good thing to bring back memories of good times and there are other times when it shouldn't happen like in the above."*

On the whole, participants tended to view the revoiced footage quite negatively. When asked to rate from 1-5 whether they agree this use of AI was disrespectful, participants tended to agree (M = 3.19, SD = 1.28).

> *"It is scary, it's messing with history. I remember those times clearly, lived through the troubles & this manipulation of history for future generations will distort the truth & could be used to insight trouble again. It's totally pointless."*

There were many different reasons given for this including the potential offensive connotations of modified accents and their values *"Dubbing is inaccurate and highlights negative connotations associated with certain accents"*. There were many who viewed this use of AI as being particularly controversial due to the nature of the conflict and the various controversies surrounding Gerry Adams as a politician. *"I think this was a diffrent time and to return back to it is negitive hopefully Ireland has moved on and realise the pen is mightier than the sword. While I doubt very much Irelands troubles will be gone in my lifetime we cannot return to a time when people lived in fear blaming each other."* Many focused specifically on Gerry Adams as a particularly controversial politician whose inclusion cast shade on the legitimacy of the entire idea: *"It is irrelevant as Gerry Adams is trying to legitimise himself through politics when all he is, is a blatant TERRORIST."*

On average, most participants disagreed that AI could be useful for Historical records (M = 2.82, SD = 1.35).

> *"I just don't like this idea. It makes me feel very uncomfortable. Further strengthens my doubts and worries surrounding AI."*

Some people agreed that the AI would be useful for recreating historical records *"I think that AI has the ability to add context to these old pieces of dialogue helping to bring greater clarity to what was attempting to be conveyed at the time, thus helping to provide a more*

accurate historical record" / *"I think, although it is a simulation of the original recording, it is closer to the person's voice and intonation than the actor - so in a way, you are getting a more accurate snapshot of the interview as it actually was"*. Many more expressed the opinion that it was "rewriting history.

> *"Whatever AI is used requires some human input in training which particularly in a hostile political sphere may introduce biases"*

Participants felt that a better approach would involve releasing the original recordings. *"The AI restoration is clumsy, and actually uninspiring because of that. The original dubbed footage is more convincing! (AI can, surely, do better!) But regardless of that shoddy AI the concept is just odd - the 'ban' was how things were, it's better to have the footage as it was then viewed - not least as the politicians involved knew they would be dubbed, so we cannot judge how they spoke in reality, and they never meant us to know. By the way, if you want to 'restore' these interviews someone - the BBC etc - must have the actual real recordings, so why faff about with shoddy AI?"*. They often complained about the quality of the AI revoiced video – which the primary researcher took very personally. *"It didn't sound like a real person far less Gerry Adams-unless he's suddenly developed a psuedo American accent. It actually felt as if I was watching some bizarre Planet of the Apes re-enactment"*.

Many agreed with the idea that historical revoicing of AI is dangerous (M = 3.66, SD = 1,223) Most criticism more broadly criticized AI as being inherently biased and hostile towards truth. *"I am hesitant to support any form of AI at all, let alone when it comes to representing historical events."* Perhaps the most interesting takes on the historical context of this redubbing were from those who remembered and experienced the Troubles, the broadcasting ban and its consequences. *"I'm old enough to have heard Gerry Adams speak, so I know what he sounds like. I think the content of what he says is more important than the way it sounds. I think that from an historical point of view hearing his voice is important."*. Participants described the importance of learning about censorship and political conflict without the need for gaudy "restored" content.

> *"I was born and raised in Belfast and lived there until 1999. I remember Sinn Féin not being allowed to have their voice on any broadcast. I think it would better to leave these videos the way that they were originally shown, but also explain exactly why they were shown like this and how things come to be like this."*

Overall, like most use cases of AI, the use of AI in historical records perhaps brings more challenges than benefits, whether it be the inherent falsity of the content or the questions it raises about how easy it may be to revoice historical footage.

5 Implications

We believe that our work is the first to experimentally test the use of AI supported censorship of political voices. We find that the voice used to dub a politician affects how they are perceived, highlighting the potentially significant power that lies in the hands of those who manipulate the voices of political speakers. This finding has implications beyond the absurdly specific context of the ban enforced during the Troubles. Most significantly, it underscores how international media can sway public opinion with their choice of translator. Historically too, the current study provokes important questions about how this ban was conceived of and strictly enforced for so many years, without any evidence to support the idea that simply revoicing figures like Gerry Adams could reign in their apparently dangerous magnetism. The condition which most closely approximated the manipulation of voices used in the Troubles was rated as significantly more credible, legitimate and charismatic than either the received pronunciation or high-pitched dubs. Perhaps the trick to censoring dissident Irish republican voices this whole time was, ironically, to make them sound English. This work recontextualizes the censorship employed during the Troubles by asking if and how such censorship works.

5.1 Implications for HCI non-work in historical records

This work provides strong critiques of the use of generative AI in editing historical records. We hope that this work is off-putting for researchers who seek to design technologies which use AI to "restore" existing footage, particularly where such footage is already controversial or edited. We have shown empirically that the use of AI in historical education and record keeping is tainted by overwhelming fears of the technology being used to rewrite history – adding to calls for more work interrogating the positive uses of media editing technology [25]. Our work highlights how footage edited by AI is not seen by viewers as historical anymore and is perhaps tantamount to automating the job of Winston Smith in 1984. Other forms of automated historical recreation such as the restoring or recoloring of footage should also be critically considered. Our work further highlights once again how every well-intentioned use of generative AI is tainted by the well of negative public opinion on the technology. This is consistent with work done on the supposedly positive, but mostly creepy use of deepfakes in recasting movies [16]. Faith in this technology was hard to find in the responses. The sentiment of responses indicates that the marketing of hype of AI is dangerous and disruptive may be turning people away from the technology itself.

Outside of the implications for generative AI, there is a broader implication here for how translated political speech is dubbed over using technology. As international conflict is often subject to translation – there should be an increased focus on how the voices we use for translated political speech affect people's attitudes. As previous literature has explored the potential of "resurrecting Margeret Thatcher" using deepfakes [7], republican activists may attempt to resurrect dissident republican leaders to challenge her. There are also broader implications for the ethics of overdubbing translated footage that need to be encouraged. If a hypothetical French TV station ran a clip of, for example, a buffoonish former/current U.S.

president, would the French language dub affect the persuasiveness of the content? The attractiveness of the politician? How do the voices we apply affect the impact of political statements? There are numerous design implications for the applications of AI voices.

5.2 Future HCI work in political censorship

There are two main takeaways of this paper for HCI researchers: different voices have different values, and neutrality does not exist. While technology is often used by those intending political violence, there has been a generalized handwaving regarding the use of technology by state and corporate actors, and the role we as HCI researchers have to play in this area. We strongly believe that it is time to take a stance on whether or not we approve of the wholesale use of technology in murdering and censoring political radicals. Certainly, a large number of industry researchers at CHI may already think this is an admirable goal based on their companies' track records. Instead of assuming our work won't fuel acts of violence and acting surprised when it does [23] - we urge researchers to take a more active position on how media editing technology begets censorship.

An alternate reading of this paper is that censorship works and governments looking to manipulate audio should particularly consider making enemy politicians sound like chipmunks. We situate this potential in an approach we think of as unprincipled, dishonourable, uncaring, sellout HCI inspired by existing alt.chi practices [9] [10]. Our paper is submitted at a time when alt.chi is discontinued, as CHI embraces the move towards a conference where it becomes impossible to tell which papers are intentionally disgusting alt.chi papers and which papers are just chi papers with poor taste. Building on established ideas such as anarchi and punk HCI [8], we acknowledge the "you know I was punk once" HCI or "conservatives are the new punks" HCI approaches. This is a radical move away from utopian futures and instead imagines a dystopian past/present. Is it time to embrace CHI's move away from overly provocative and speculative work. We should instead move toward "non-speculative fact" - everything is fine as it is, and our work aims to keep it that way.

Acknowledgments

This publication is the result of research conducted with the financial support of Taighde Éireann – Research Ireland under Grant number 13/RC/2094_2.

References

[1] Aniruddha Kumar Baro. 2019. Information and Communication Technology and Ethnic Riots: Riots in Goalpara District of Assam, India. Int. J. Innov. Technol. Explor. Eng, 8(12), 2072-2075.

[2] James Brassett. 2021. Irony and the Liminality of Resistance. In The Ironic State (pp. 77-96). Bristol University Press.

[3] Tony Hall. 1994. A gag that chokes freedom: 3 edition. Independent (London, England : 1986)

[4] Michael Hameleers, Toni GLA. van der Meer, and Tom Dobber. 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.

[5] Caroline Kennedy-Pipe. 2014. The origins of the present troubles in Northern Ireland. Routledge.

[6] Gabrijela Kišiček. 2018. Persuasive power of prosodic features. Argumentation and advocacy, 54(4), 345-350.

[7] Dominic Lees, Tom Bashford-Rogers, and Marcus Keppel-Palmer (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.

[8] Conor Linehan and Ben Kirman. 2014. Never mind the bollocks, i wanna be anarCHI: a manifesto for punk HCI. In CHI'14 Extended Abstracts on Human Factors in Computing Systems (pp. 741-748).

[9] Conor Linehan, and Ben Kirman. "Mc hammer presents: the hammer of transformative nostalgification-designing for engagement at scale." In Proceedings of the 2017 CHI conference extended abstracts on human factors in computing systems, pp. 735-746. 2017.

[10] Conor Linehan, Sabine Harrer, Ben Kirman, Shaun Lawson, and Marcus Carter. 2015. "Games against health: a player-centered design philosophy." In Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, pp. 589-600.

[11] Lisa McGee. 2018. Derry Girls (TV series). Hat Trick Productions.

[12] Niall Meehan. 2003. How RTE censored its censorship. Sunday Business Post.

[13] Eoin Moloney. 1991. Closing down the airwaves: The story of the broadcasting ban. In The media and Northern Ireland: covering the troubles (pp. 8-50). London: Palgrave Macmillan UK.

[14] Paul Mozur. 2018. A genocide incited on Facebook, with posts from Myanmar's military. The New York Times, 15(10), 2018.

[15] Orla T. Muldoon. 2004. Children of the Troubles: The impact of political violence in Northern Ireland. Journal of Social Issues, 60(3), 453-468.

[16] Gillian Murphy, Didier Ching, John Twomey and Conor Linehan. 2023. Face/Off: Changing the face of movies with deepfakes. Plos one, 18(7), e0287503.

[17] Oliver Niebuhr and Radek Skarnitzl. 2021. Those who shout the loudest–Do they sound more charismatic?. In Proc. 1st International Conference on Tone and Intonation (TAI) (pp. 1-5).

[18] Max Pettigrew. 2016. The 'oxygen of publicity' and the suffocation of censorship: National newspaper representations of the British broadcasting ban (1988–94). In The Northern Ireland Troubles in Britain (pp. 227-242). Manchester University Press.

[19] Julie Posetti, and Alice Matthews.2018. "A short guide to the history of 'fake news' and disinformation." International Center for Journalists 7.

[20] Nathan Pluquet. 2023. The Dangers of Emergency Powers and the Downfall of Democracy: British Security Response in Northern Ireland during the "Troubles".

[21] Betty Purcell. 1991. The silence in Irish broadcasting. In The Media and Northern Ireland: covering the troubles (pp. 51-68). London: Palgrave Macmillan UK.

[22] Andrew Rosenberg and Julia Hirschberg. 2009. Charisma perception from text and speech. Speech communication, 51(7), 640-655.

[23] Miriam Sturdee, Joseph Lindley, Conor Linehan, Chris Elsden, Neha Kumar, Tawanna Dillahunt, Regan Mandryk, and John Vines. 2021. "Consequences, Schmonsequences! Considering the Future as Part of Publication and Peer Review in Computing Research." In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1-4. 2021.

[24] Konstantin O. Tskhay., Rebecca Zhu, Christopher Zou, and Nicholas O. Rule. 2018. Charisma in everyday life: Conceptualization and validation of the General Charisma Inventory. Journal of personality and social psychology, 114(1), 131.

[25] John Twomey, Didier Ching, Matthew Peter Aylett, Michael Quayle, Conor Linehan, and Gillian Murphy. 2024. What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology. IEEE Transactions on Technology and Society.

[26] John Twomey, Didier Ching, Matthew Peter Aylett, Michael Quayle, Conor Linehan, and Gillian Murphy. 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, no. 10: e0291668.

Appendix II: Face/Off: Changing the face of movies with deepfakes

Link to published article: https://dx.plos.org/10.1371/journal.pone.0287503

PLOS ONE

RESEARCH ARTICLE

Face/Off: Changing the face of movies with deepfakes

Gillian Murphy[1,2]*, **Didier Ching**[1,2], **John Twomey**[1,2], **Conor Linehan**[1,2]

1 School of Applied Psychology, University College Cork, Cork, Ireland, 2 Lero, The Science Foundation Ireland Centre for Software Research, Limerick, Ireland

* gillian.murphy@ucc.ie

Abstract

There are growing concerns about the potential for deepfake technology to spread misinformation and distort memories, though many also highlight creative applications such as recasting movies using other actors, or younger versions of the same actor. In the current mixed-methods study, we presented participants (N = 436) with deepfake videos of fictitious movie remakes (such as Will Smith staring as Neo in The Matrix). We observed an average false memory rate of 49%, with many participants remembering the fake remake as better than the original film. However, deepfakes were no more effective than simple text descriptions at distorting memory. Though our findings suggest that deepfake technology is not uniquely placed to distort movie memories, our qualitative data suggested most participants were uncomfortable with deepfake recasting. Common concerns were disrespecting artistic integrity, disrupting the shared social experience of films, and a discomfort at the control and options this technology would afford.

Check for updates

OPEN ACCESS

Citation: 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

Editor: Michal Ptaszynski, Kitami Institute of Technology, JAPAN

Received: December 1, 2022

Accepted: June 5, 2023

Published: July 6, 2023

Copyright: © 2023 Murphy et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: The materials and data are available at https://osf.io/czy9w/.

Funding: This work was supported by funding from LERO awarded to G.M (https://lero.ie/, Award PP5004). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. There was no additional external funding received for this study

Competing interests: The authors have declared that no competing interests exist.

1 Introduction

Deepfakes are a type of manipulated media created using artificial intelligence technologies [1]. Deepfakes most often consist of videos where an artificial face has been superimposed onto another person's face, resulting in a highly convincing recording of someone doing or saying something that they never did [2]. While computer-generated imagery is not a new technical development, the rise of deepfakes and the development of deepfake creation apps promises an especially realistic product without significant training or expensive equipment [3], and research has shown that most individuals cannot reliably distinguish AI generated faces from real faces [4]. Almost as soon as deepfake technology emerged on Reddit in 2017, discussions around the grave harm they could cause was widespread [5]. Indeed, the first deepfakes posted online were fabricated pornographic videos featuring celebrities, and non-consensual deepfake pornography remains a pressing issue [6] As well as this, there are concerns that widespread use of deepfake technology could give rise to a society where individuals lose faith in the veracity of media [7] and where video evidence cannot be used in court cases [8,9]. It has been argued that deepfakes may be especially potent at distorting memories, giving rise to false memories for events that never occurred [10,11].

More recently, the positive applications of deepfakes have been discussed. Deepfakes allow for quick and relatively cheap translation of films, marketing material and educational resources into other languages, as well as potentially restoring speech to individuals who have lost their voice to disease or injury [7,8]. For example, deepfake technology was used to depict the footballer David Beckham "speaking" nine languages in a campaign against malaria [12] and to protect the identity of LGBT individuals speaking out about persecution in Russia [13]. In 2018, the speech synthesis company CereProc used deepfake technology to produce audio of John F. Kennedy delivering the speech he was due to give on the day he was assassinated [14]. Some have suggested that deepfake technology may soon allow us to hyper-personalise our shopping and entertainment–using virtual deepfakes of ourselves to try on a pair of sunglasses before purchasing, or inserting our own faces onto video game characters [3].

One field which could be revolutionised by deepfake technology is the creative arts, especially television and cinema. It has been suggested that deepfakes could be used to translate films, edit misspoken words in a recorded scene, or to put the leading actor's face on the body of a stunt-double [3]. Current estimates are that deepfake videos can be produced at less than ten percent of the typical cost of other synthetic media [12]. We are seeing the adoption of this technology already, with deepfakes being used creatively to improve storytelling. For example, in the Star Wars TV show "The Book of Boba Fett", the character of Luke Skywalker is played by an actor with a younger version of the original Luke Skywalker (actor Mark Hamill) superimposed onto their face [15]. This technology has been embraced by art museums and galleries too. Painter Salvador Dalí was recreated using deepfake technology in the Dali Museum in Florida [16]. Projector screens showed Dalí move, speak, and interact with museum visitors, even inviting audience members to take a selfie with him. An exhibition in the International Centre of Photography in New York City used deepfake technology to recast the film *The Warriors*, inserting visitors into key scenes in the film [17]. In less official capacities, in the homes of movie-lovers around the globe, individuals are recasting films and posting clips to Youtube. The site is awash with highly realistic face-swaps and as noted by Meskys et al. [12] these videos are made for pure fun and as a form of creative expression.

In the imagined not-too-distant future then, deepfakes may become ubiquitous and may allow us to personalise our media and entertainment. In the current work, we explore the potential issues that may arise from such a reality. Firstly, we ask, can deepfakes can distort our memories and beliefs Moreover, are deepfakes so realistic and convincing that they will distort our memories more drastically than any other current misinformation medium? This has been suggested by some authors [10] and indeed, deepfakes may harbor a situation where "seeing is believing", distorting our shared understanding of culture and history [18]. Secondly, we ask how audiences feel about the promised possibilities of deepfakes. In what ways would the ability to personalise our entertainment be appealing or unappealing? Understanding these questions is key in designing and regulating for a future in which deepfake technology becomes more widespread.

The current paper reports a mixed-methods study concerning deepfakes in cinema. We presented participants with a series of clips from remakes of well-known films. Some of these clips were genuine remakes that were released commercially in cinemas (such as the remakes of *Aladdin* and *Carrie*), while others were deepfakes made by Youtube and Reddit users. We assessed whether participants would form false memories for these fictional remakes, and we compared the false memory rate to participants who merely read a written description of the film, in order to assess whether deepfakes are especially likely to deceive. After completing the experiment, we showed participants some deepfake clips and explained how the technology works. Participants were asked to reflect on whether they would use this recasting technology in future if it became available. We analysed their responses using discursive thematic analysis

to gain greater understanding of the benefits and risks of this technology for cinemagoers. The current study is one of the first to empirically assess the potential for deepfakes in revolutionising cinema and the implications for developers, moviemakers and society are discussed.

2 Literature review

While the definition of deepfakes varies slightly between papers, it is generally accepted that deepfakes are audio, images, or videos which have been manipulated using AI neural networks to swap the voice or face of one person with another [19]. Generative Adversarial Networks (GAN) are usually required to create these manipulated audio, images and videos. GAN's consist of two neural networks. Network A (generator) is fed information from a source material and produces its own "fake" version of it, and Network B (discriminator) discriminates and identifies the fakes from the generator. The generator produces the "fake" media and this and the original source material is fed into the discriminator. If the discriminator finds discrepancies between the original and the fake, then the generator learns from this discrepancy and produces better fakes. This process continues to improve the fake media until it is indistinguishable to the discriminator. The resulting synthetic media is known as a deepfake [20]. While the early deepfake materials that emerged online in 2017 showed clear signs of manipulation, they soon became more realistic and convincing due to technological advancements [21].

There are widespread concerns that deepfakes may mislead and spread misinformation, by presenting a vivid depiction of a person doing something they did not really do. These concerns continue to grow as deepfake technology becomes ever more widely-available, in the form of mobile face-swap apps, which have topped the most dowloaded charts when released [12]. Yet, others have argued that the unrelenting focus on the negative aspects of deepfakes reflects a techno-paranoia that is devoid of nuance. For example, Broinowski [22, p. 122] argued that "dystopian predictions about deepfakes side-step their socially enriching capacity to inspire, transform, and engage, and instead position the technology as harmful by design—an instrument as unequivocal in its mendacious intent as guns are unequivocal instruments of death, irrespective of the context in which it is used, or the intentions of its user". In the current study, we assess both the potential harms and benefits of deepfakes, as they become ever more accessible to members of the public.

2.1 Misinformation & false memories

Decades of psychological research has demonstrated that memory is reconstructive and does not work like a video camera [23]. Our memories for what we originally experienced can be distorted by post-event information ("the misinformation effect"). Early research in this field found that after watching a recording of a car accident, participants' recollection of the speed the cars were going was significantly affected by the wording of the question. Those asked what speed a red car was going when it 'hit' the other car gave lower estimates that those asked about when the red car 'smashed' into the other car [24]. Those who heard the word 'smashed' in the investigator's question were also more likely to report seeing broken glass at the scene of the accident, though there was none. Since this ground-breaking research, many studies have demonstrated that it is possible to plant memories of entirely false events–for spilling punch all over a bride at a wedding [25], taking a hot air balloon ride as a child [26], or witnessing fabricated political events during the Irish abortion referendum [27], Brexit [28] or the US Capitol Riots [29].

The source-monitoring framework (SMF) is one way of understanding and predicting the occurrence of false memories [30]. According to the SMF, remembering is an inferential

process where we attribute our mental experiences (including images and scenes) to a particular origin (something we truly witnessed, a dream, an imagined situation, etc.). Typically, we are correct in these judgements, and we use reliable indicators of veracity to establish the source of a mental event (e.g. how vivid and detailed is the scene, how plausible is the event, etc.). False memories can occur when we inadvertently and unconsciously attribute an internally generated mental experience to an incorrect source. For example, a wealth of research has demonstrated that simply imagining a fictional event can lead to significant levels of false memories for that event ("imagination inflation";[31]). With regard to deepfakes, there are concerns that this form of hyper-realistic doctored media may bias these judgements, making viewers more susceptible to false memories. Deepfakes are incredibly convincing and can be hard to detect, and we know that even more basic forms of doctored videos have effects on memory. For example, in a 2010 study, Wade and colleagues found that participants who viewed a doctored video that showed a fellow participant cheating at a gambling task were three times more likely to sign a statement confirming they had witnessed the cheating than those who were merely told the video existed [32]. A false confession study using doctored videos also found that those who viewed a doctored video that showed themselves cheating at a task were more likely to come to believe they had in fact cheated [33]. From a source-monitoring perspective, viewing doctored videos may increase familiarity with a fake event and increase the ease with which the event can be pictured. This can make the event seem more credible and the ease with which such a 'memory' can come to mind may lead participants to falsely believe it truly occurred [33]. This has led some to conclude that deepfakes are likely to be especially potent in distorting our memories of the past [10].

There is limited research on doctored video effects on memory, perhaps in part because the technical skills and equipment required to create convincing materials for such a study may not often be found in psychology research labs. This appears to be changing with the rise of simple deepfake creation apps, however we do not, as yet, have firm evidence for the effects of deepfakes on memory. A recent study compared false memories for fabricated news stories depicted as deepfakes against the same fake stories presented as either text or text alongside a photograph [34]. Though participants had negative perceptions of deepfakes after taking part in the study (rating them as dangerous, unethical and highly convincing) and many participants reported false memories for fabricated events after viewing a related deepfake, deepfakes did not reliably increase the incidence of false memories relative to less technologically advanced mediums. Thus, while the literature around deepfakes frequently cites the spread of misinformation and memory distortion as a likely harm [10], the empirical evidence to support such a claim is not yet apparent.

2.2 Democratising deepfakes

Deepfake technology has spread rapidly from a dedicated community on Reddit with significant technical skills, to become accesible to an average internet user with no specific training or equipment [35]. This democratisation of deepfake technology has had measurable harms for many individuals. In a classifying exercise, cyber-security company Deeptrace found in 2019 that 96% of all deepfake videos on the internet were of a pornographic nature [5]. Almost all of these were created nonconsenually and the majority, but not all, featured celebrities. With the advances in deepfake technology, a wealth of material is no longer required to train the AI and many internet users are using deepfake apps to create fake pornographic footage of those they know personally [6,13]. This has been classed as a form of technology-facilitated image-based abuse [36].

Beyond pornography, accessible deepfake technology may soon allow members of the public to regularly create and consume deepfake media. Many of these applications are creative, educational and playful, yet they do not receive the same attention in the literature as the more negative elements [22]). It has been speculated that deepfakes can and will be used to customise and personalise movies, video games, and other media by superimposing one's face onto the characters [37]. Deepfakes are also beginning to be used by creative artists. The rapper Kendrick Lamar released a music video in 2022 ('*The Heart Part 5*') depicting himself rapping but superimposing different influential black men onto his face using deepfake technology. A trailer for the 2019 film *Gemini Man* starring Will Smith used deepfake technology to depict an older clip of Smith, from the show *The Fresh Prince of Bel Air*, speaking about the film. These are examples of deepfakes being used with the full awareness and consent of the stars, but popular media has featured much speculation as to how this technology will transform the film industry. A 2019 article in Collider predicted "Within five years, deepfakes will have advanced to the point that it will not only be possible but probable we'll be watching the likes of Marilyn Monroe co-starring opposite Heath Ledger in entirely original feature films" [38]. Others have speculated that the option to recast oneself in popular media will soon be extended to members of the public; "it is inevitable that, quite possibly by the time you read this article, the technology will be available for anyone to 'star' not only in high quality deepfakes of scenes from TV shows, but also in any commercial they want" [3, p. 476]. Despite these technological advances opening up new possibilities for public participation in films, little is known about how we should design these digital tools or how we would limit any harms they may cause.

2.3 Research gap

Existing research has highlighted the potential for misinformation to distort memory [23] and the emergence of deepfakes has heightened these concerns [10]. However, the positive and creative applications of deepfakes are also an exciting area of speculation, one that promises to offer new possibilities and a world of choice to members of the public [3]. In the current study, we sought to answer two research questions to better understand how deepfakes may affect cinema in the near future: Q1. Do deepfake clips increase false memories for fabricated films? Due to a lack of empirical deepfake research, we do not yet know if deepfakes are uniquely poised to distort memories and the current study will assess this in regards to memories of films. Q2. Why would moviegoers choose to use this technology, or not? Though it is clear that the possibilities of this technology are expanding, we do not yet understand what would make cinematic applications attractive to members of the public. What features ought to be considered when designing this technology and what harms should we plan to mitigate? The current study will thus advance our understanding of deepfakes in the film industry.

3 Method

3.1 Method

This study was preregistered (available here https://aspredicted.org/blind.php?x=RHM_5BB). The materials and data are available at https://osf.io/czy9v/.

3.2 Participants

Participants were recruited by undergraduate students as part of a class project. 570 participants completed the survey in full, however, in line with our preregistration we removed those who reported searching the internet or asking for help to answer the questions (n = 20) or

Participants were then asked "Some people suggest that in future we could use deepfake technology to choose the actors and actresses in films to suit our personal tastes (e.g. if you didn't like Jack Nicholson in The Shining, you could simply recast the character as Brad Pitt). Would you like to see this type of technology become available?" and could answer from Definitely Yes, Probably Yes, Might or Might Not, Probably Not, Definitely Not. Participants were then prompted to tell us why or why not, responding using an open text box. Lastly, participants were asked "If you could recast actors in films or TV series, what would motivate you to do so? (select all that apply)" and could select from "If I found the original actor annoying", "If I disagreed with the original actor's politics or actions and didn't want to watch a film with them in it", "I would use this for the novelty", "If I preferred the replacement actor for any reason", "If the replacement actor better fitted my perception of the character", "If I wanted to increase the diversity of the cast", "Other (please describe)". Participants were then further debriefed and thanked for their participation.

3.4 Analysis

The quantitative data was analysed and reported as per our preregistered plans. The qualitative data was analysed following Braun and Clarke's [39,40] guidelines for carrying out reflexive thematic analysis. Specifically, the data was first read closely. Where a response included a number of distinct ideas, it was divided into separate chunks for the purposes of open coding. Thus, 292 individual points, taken from the responses of 263 participants, were open coded. Analysis then progressed towards defining and iteratively refining a set of themes that meaningfully summarise the thoughts expressed by participants, making sure that themes always reflected the data.

3.5 Ethics

The study was approved by the School of Applied Psychology Ethics Committee at University College Cork. Participants provided informed consent at the start of the online survey, by clicking to indicate their consent. As misinformation researchers must ensure they are not causing harm with their experimental methods [41]), participants were debriefed using techniques that have shown to be effective in prior false memory research [42,43].

4 Results

4.1 Quantitative analysis

Participants rated the deepfakes as moderately convincing on average, on a scale of 0–100, where 100 is extremely convincing—The Shining remake (M = 61.78, SD = 31.02), The Matrix remake (M = 50.46, SD = 33.11), Captain Marvel remake (M = 70.39, SD = 31.08), Indiana Jones remake (M = 71.96, SD = 27.47). Participants gave similar answers when asked how likely others would be to fall for this video, with 100 being extremely likely—The Shining remake (M = 61.70, SD = 30.12), The Matrix remake (M = 55.66, SD = 32.76), Captain Marvel remake (M = 68.60, SD = 30.28), Indiana Jones remake (M = 67.65, SD = 28.81).

Participants readily formed false memories for these fictitious movies. Captain Marvel was most frequently falsely recalled (73%), followed by Indiana Jones (43%), The Matrix (42%), and The Shining (40%). Of those who falsely remembered each film, 41% remembered the Captain Marvel remake was better than the original, 13% remembered the Indiana Jones remake was better than the original, 12% remembered The Matrix remake was better than the original, and 9% remembered The Shining remake was better than the original.

tic the video was and how likely most people would be to fall for it, on a scale of 0 (not at all) to 100 (extremely).

Participants were then asked "Some people suggest that in future we could use deepfake technology to choose the actors and actresses in films to suit our personal tastes (e.g. if you didn't like Jack Nicholson in The Shining, you could simply recast the character as Brad Pitt). Would you like to see this type of technology become available?" and could answer from Definitely Yes, Probably Yes, Might or Might Not, Probably Not, Definitely Not. Participants were then prompted to tell us why or why not, responding using an open text box. Lastly, participants were asked "If you could recast actors in films or TV series, what would motivate you to do so? (select all that apply)" and could select from "If I found the original actor annoying", "If I disagreed with the original actor's politics or actions and didn't want to watch a film with them in it", "I would use this for the novelty", "If I preferred the replacement actor for any reason", "If the replacement actor better fitted my perception of the character", "If I wanted to increase the diversity of the cast", "Other (please describe)". Participants were then further debriefed and thanked for their participation.

3.4 Analysis

The quantitative data was analysed and reported as per our preregistered plans. The qualitative data was analysed following Braun and Clarke's [39,40] guidelines for carrying out reflexive thematic analysis. Specifically, the data was first read closely. Where a response included a number of distinct ideas, it was divided into separate chunks for the purposes of open coding. Thus, 292 individual points, taken from the responses of 263 participants, were open coded. Analysis then progressed towards defining and iteratively refining a set of themes that meaningfully summarise the thoughts expressed by participants, making sure that themes always reflected the data.

3.5 Ethics

The study was approved by the School of Applied Psychology Ethics Committee at University College Cork. Participants provided informed consent at the start of the online survey, by clicking to indicate their consent. As misinformation researchers must ensure they are not causing harm with their experimental methods [41]), participants were debriefed using techniques that have shown to be effective in prior false memory research [42,43].

4 Results

4.1 Quantitative analysis

Participants rated the deepfakes as moderately convincing on average, on a scale of 0–100, where 100 is extremely convincing—The Shining remake (M = 61.78, SD = 31.02), The Matrix remake (M = 50.46, SD = 33.11), Captain Marvel remake (M = 70.39, SD = 31.08), Indiana Jones remake (M = 71.96, SD = 27.47). Participants gave similar answers when asked how likely others would be to fall for this video, with 100 being extremely likely—The Shining remake (M = 61.70, SD = 30.12), The Matrix remake (M = 55.66, SD = 32.76), Captain Marvel remake (M = 68.60, SD = 30.28), Indiana Jones remake (M = 67.65, SD = 28.81).

Participants readily formed false memories for these fictitious movies. Captain Marvel was most frequently falsely recalled (73%), followed by Indiana Jones (43%), The Matrix (42%), and The Shining (40%). Of those who falsely remembered each film, 41% remembered the Captain Marvel remake was better than the original, 13% remembered the Indiana Jones remake was better than the original, 12% remembered The Matrix remake was better than the original, and 9% remembered The Shining remake was better than the original.

Our primary question was whether presenting the movie in the format of a deepfake would increase the false memory rate. As shown in Fig 1, this was not the case. The Shining was remembered at a rate of 40% when presented as a deepfake and 39% when presented as text. Captain Marvel was remembered at a rate of 75% when presented as a deepfake and 70% when presented as text. The Matrix was remembered at a rate of 39% when presented as a deepfake and 45% when presented as text. Indiana Jones was remembered at a rate of 42% when shown as a deepfake and 44% when presented as text. False memory rates did not differ for any of the movies according to presentation format (all $p > .05$).

Interestingly, presenting participants with videos did not have a major impact on memories for the real films either. The Tomb Raider remake was remembered by 74% as a video, 64% as text, the Charlie & The Chocolate Factory remake was remembered by 97% as a video, and 98% as text, Total Recall was remembered by 49% as video and 43% as text, and Carrie was remembered by 61% as video and 55% as text.

After viewing each film and reporting their memories, participants were asked to consider whether all the films they saw were real, and to rate the truthfulness of each film on a scale of 0 (definitely not a real film) to 100 (definitely a real film). To assess whether truthfulness varied by presentation format and whether the movie was real or fake, we conducted a 2x2 repeated measures ANOVA with the factors of presentation format (text, video) and film type (real, fake). We found a significant main effect of presentation format $F(1, 344) = 7.05$, $p = .008$, $\eta p2 = 0.02$. On average, films presented as videos were rated as more truthful ($M = 49.88$, $SD = 21.10$) than those presented as text ($M = 45.89$, $SD = 21.47$). We also found a large significant main effect of film type, $F(1, 344) = 335.16$, $p < .001$, $\eta p2 = 0.49$, with real films rated as more truthful ($M = 66.24$, $SD = 19.69$) than fake films ($M = 29.53$, $SD = 28.60$). We also found an interaction effect between presentation format and film type ($F(1, 344) = 14.15$, $p < .001$, $\eta p2 = 0.04$) such that the difference between truthfulness ratings for real movies presented as video ($M = 70.79$, $SD = 25.24$) and text ($M = 61.70$, $SD = 26.78$) was greater than the difference between fake movies presented as video ($M = 28.97$, $SD = 35.59$) and text ($M = 30.08$, $SD = 34.83$).

When asked whether they would like to see wider availability of this kind of technology (using deepfakes to swap actors in and out of films), participants had widely diverging responses. About a third were keen to see this technology become more widespread (12% Definitely Yes, 19% Probably Yes), 15% were unsure, and over half did not want to see this technology become more widespread (31% Definitely Not, 24% Probably Not). When asked why they would use this technology, participants selected many of our suggested answers as well as providing their own reasons. The most popular selected reason was "If the replacement actor better fitted my perception of the character" (n = 174), followed by "If I found the original actor annoying" (n = 139) and "I would use this for the novelty" (n = 121). Some participants also selected reasons related to diversity or politics: "If I disagreed with the original actor's politics or actions and didn't want to watch a film with them in it" (n = 96), "If I wanted to increase the diversity of the cast" (n = 80). Others simply selected "If I preferred the replacement actor for any reason" (n = 74).

4.2 Qualitative analysis

Our intention was to understand, describe and discuss our participants' thoughts about the use of deepfakes in films, given all they had learned about deepfakes through participating in the study. Data describes our participants' excitement and imagination, as well as their fears and concerns. 263 participants responded to the prompt; "Some people suggest that in future we could use deepfake technology to choose the actors and actresses in films to suit our

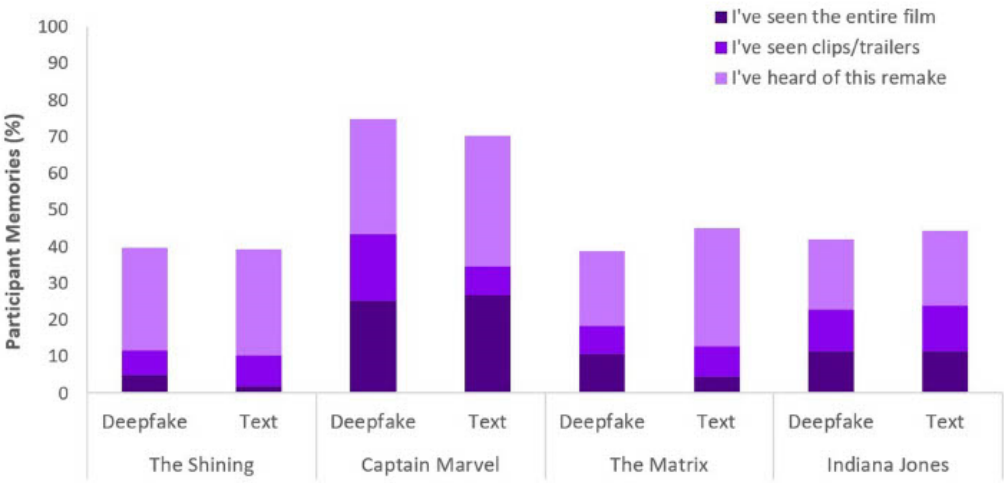


Fig 1. Participant memories for each of the four fictitious movie remakes.

https://doi.org/10.1371/journal.pone.0287503.g001

personal tastes (e.g. if you didn't like Jack Nicholson in The Shining, you could simply recast the character as Brad Pitt). Would you like to see this type of technology become available? Why, Why not?." Responses varied in terms of length and depth, from short, vaguely positive responses, to lengthy discussions of art, employment and copyright law, and ethics.

Some brief comments expressed positive sentiments about deepfake technology being used to replace actors, but without any justification or argumentation to support that position. For example, many participants made statements such as, "It would be cool and interesting," or, "could improve the movie." While it is important to acknowledge that these positions exist, it is difficult to analyse such short comments in a meaningful manner. Similarly, there were many vaguely negative comments entered without any justification. Statements such as, "it's dumb," "unnecessary," "little more than a novelty", "creepy" and "it is weird and unnerving" were entered. However, the vast majority of comments were at least a full sentence in length, and provided some justification for the author's position. Fig 2 illustrates the overall thematic structure of our findings. There are six main themes, with five focused on negative aspects and one focused on positive aspects. Four of these main themes had multiple subthemes. We will discuss each theme in turn, providing examples of participant comments described by each theme.

4.2.1 Seeing positive use cases. Participants mostly demonstrated negative views when asked about the possibility of using deepfake technology to replace actors in a film. Indeed, much of the research and scholarship on deepfakes expresses similarly negative viewpoints [13,44]. However, there were some positive responses from participants also. These responses can mostly be classified into two sub-themes; separating actor from role, and supporting creativity.

Many participants expressed that the possibility of separating actors from roles could be a positive development, and two contrasting reasons were provided. Firstly, a number of participants mentioned situations where their dislike of a specific actor (e.g., "If the actor was a bad

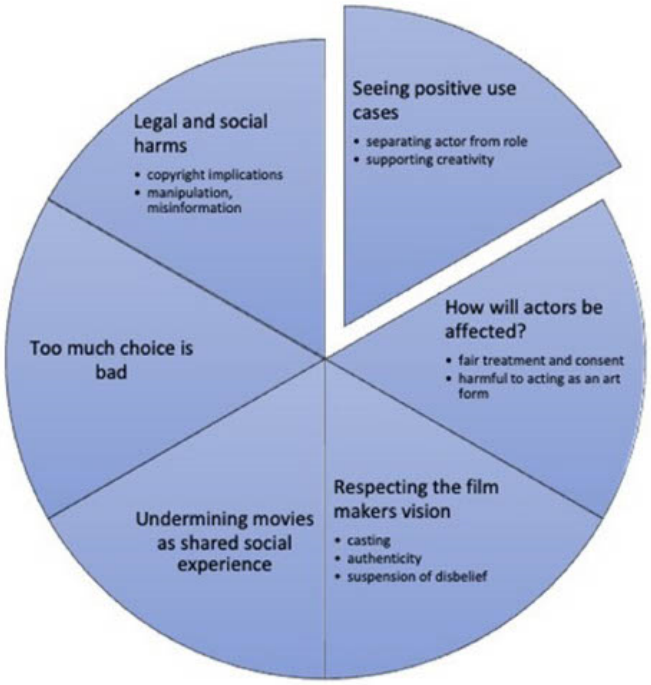


Fig 2. The thematic structure of participant responses.

https://doi.org/10.1371/journal.pone.0287503.g002

person") impairs their enjoyment of a film. For example, "beloved movies that cast a controversial or hated actor/actress have the opportunity to recast and make the film more enjoyable." Indeed, many participants mentioned that they do not like specific actors, and consequently do not watch their films. Interestingly, the idea of re-casting controversial actors may provide a way for people to feel they are "separating the artist from the art", a challenge that has received renewed attention since the onset of the #MeToo movement [45]. Secondly, many participants mentioned that they have "favourite actors" who they would enjoy watching in new roles, or in some of their favourite films; "you can always have your favourite actors in whatever movie you want" and "if you like watching certain actors for certain reasons (their looks!!) I would cast them in a lot more movies." Participants also caution that the ability to separate actor and role could be abused and could, "in some situations lead to racism (replacing POC actors with white ones)." Indeed, anywhere that power is given to audiences to participate in the generation of content, there should be some consideration given to the abuse of that power, and to mitigating actions [46].

A number of participants suggested positive uses of deepfakes that were not simply utilitarian, but could be considered creative. Indeed, participants were excited that deepfake technology could, "allow for audience participation in the creative process." It would "be interesting and comical to change different actors into different movie parts" and "could be nice to have Nicholas Cage in every movie." This is in line with the enormous volume of recast film clips currently available on Youtube and Reddit, contributed by average internet users, as well as the explosion of accessible and low-cost apps that allow the creation of deepfakes or less sophisticated 'cheap fakes' [7]). Participants also mentioned that movie makers may find uses for the technology themselves, "especially the potential for creative approaches to archive animation and the replacement of lips instead of dubbing." There are relatively few positive examples of deepfakes, or even positive discussions of deepfakes, in the existing academic literature, with most publications focusing on the harms and risks [13,37,47]. Though the majority of our participants' responses focus on describing harms, it is valuable to see our participants thinking about how deepfake technology can also be used in playful, creative and empowering ways.

4.2.2 How will actors be affected?. Many participants chose to answer the question by speculating how deepfake technology could have harmful effects on both the profession and the art of acting. Some discussed how using deepfake technology to replace actors in films is unfair to actors; either to those who originally starred in the films, whose "hard work" and performance could now go uncredited or unacknowledged, or to an actor who had not agreed to star in a particular type of film through choice, and whose choice has now been removed, "people deserve to have their reputation protected, and their face attached to only films they consent to" "like when Galaxy [a brand of chocolate] cast a deepfake Audrey Hepburn in their commerical a few years ago" Indeed, participants frequently discussed the harms caused by famous faces being applied to pornographic movies, a common concern in the academic literature [6,36]. Participants expressed concern about how, and whether, actors would get paid in a situation where they are replaced in a film, "It depends on whether the actor whose likeness is used will be properly compensated" and, "people should be paid for their labour." They also worried that the technology could lead to an ever decreasing pool of actors starring in films, as the same famous faces become applied to an increasingly broad range of films.

One of the most frequent points made across all participants, was a concern over how deepfaked performances could be lower in technical quality than genuine performances. Participants made the point that acting is not just a way of selling films; it is an art form; deepfakes, "will take away from the beauty of authentic, real acting." Participants complained that the digital swapping of actors faces is likely to diminish the quality and nuance of acting in the film, "Faking these things takes the power out of the original performance. It would be nothing but a detriment to the craft", "I think the performance of the actor or actress is more important than their face" and "some actors aren't interchangeable, their talent is unique." Nobody suggested that deepfaked performances could be more convincing or engaging than live recorded performances, indeed a number of participants made the point that, "there's no way to deepfake a bad performance away."

4.2.3 Respecting the film makers' vision. Many participants felt that the prospect of replacing actors in films using deepfake technology is ultimately unfair on the people who make movies. Participants wrote about the importance of respecting the original vision of the director and producers, and the choices they make in bringing a story to film. Again, participants pointed out that filmmaking is an art form, and that "the artistic direction of the director and production should be valued and respected" where "the director knows what they are doing better than the viewer most of the time." Participants mentioned that the technology may be commercially successful, but would be harmful to movies as an artistic expression, "the technology would only serve to profit media conglomerates, and would disempower the

artistic choices of the filmmaker", "the loss of humanity behind art scares me" and "art should not be customisable to the viewer." This theme has three sub-themes, all of which received significant discussion from our participants.

Many participants commented that casting decisions are intentional, and that finding the right actor for the role is a core part of what makes a film unique and memorable. "I think that casting directors know what they're doing when they're choosing actors to play characters" and "As much as the world would currently love a button that turns every action hero into Keanu Reeves, it's artistically bankrupt and dismisses the work that goes into casting appropriately for movies."

Frequently, participants suggested that the "authenticity" of a film would be harmed in a situation where actors were replaced using deepfake technology. Typically, this critique was quite vague, (e.g. "Authenticity cannot be deep faked") and, indeed, similar points were made when referring to casting choices and the art of acting. However, given the frequency with which the word "authenticity" was used in posts, we felt this was worth identifying as a separate concern. It appears that participants are concerned with an overall trend of decreasing "realism" in films, "Too much tech. Need something to keep movies somewhat real", "Removes the validity & originality for future generations to come. Brings in the question of 'who/when was the original'."

A number of participants mentioned how deepfaked versions of films may undermine the viewers' experience of immersion in a story, or harm their experience of the magic of cinema, due to noticing "uncanny" errors with the technology, or wondering whether the version they are watching is the one intended by producers. It "sort of interrupts the fantasy aspect of the film." One participant suggested "I like movies to escape. Not to select variables."

4.2.4 Undermining movies as a shared social experience. A number of participants focused not on how deepfakes could impact the quality of the media itself, but on how deepfakes could impact how people use and discuss that media socially. For example, one participant noted that "if everyone is experiencing a different version of the film they watch, it wouldn't make for very good conversation or debate amongst people who watch it." Another commented "I think part of the enjoyment of a film is seeing the same one as others and being able to talk about it." The concern participants expressed regarding the social experience of film viewing is perhaps not the first that comes to mind when thinking of deepfakes, but it certainly reflects a branch of academic scholarship that sees media as shared cultural capital, important for building cultural ties [48]. In this regard, the erosion of film as a shared experience is certainly a harm worth considering.

4.2.5 Legal and social harms. Given that much of both the academic literature and media reporting on deepfakes focuses on their potential to cause serious harm through manipulation, misinformation and personal attacks [8,13], surprisingly few of our participants focused on these issues when considering the replacement of actors in movies. However, some participants did raise concerns over deepfakes that went beyond considering their impact on movies and movie watching. These points form two subthemes, manipulation and misinformation, and copyright implications.

A number of participants made statements about how deepfakes could be used in future to misrepresent, manipulate and misinform. Participants discussed how deepfakes could be used for propaganda purposes, "imagine a place with heavy censorship using deep fake videos in their news cycle to convince the populace of things that aren't true. Things like that make it dangerous." They expressed concerns over how doctored videos could be used in legal proceedings to undermine the justice system, "Could lead to doctored evidence", "everyone is at risk of having their face placed on any video/film." Finally many argued that the technology could be used to place unknowing people into sexualised or pornographic videos without their

consent, "Has dangerous repercussions re IBSA [image based sexual abuse]". These findings suggest that many of the concerns that have dominated the literature concerning deepfake technology are also prominent in the minds of the public [6,36].

A surprisingly small number of participants cited concerns over copyright law, licensing and plagiarism. Participants questioned who would get paid for films that included re-casted actors, and how that payment process would work. In practice, actors negotiate different types of contracts for different types of films; some based on flat fees, some including profit sharing, etc. Making those calculations for each actor that may be re-cast in a role involves significant legal work, especially as copyright exemptions may apply for educational or satirical use of images [49]. For example, "you would need to license the likeness of an actor and the tech for paying those royalties does not exist yet." Participants seemed skeptical over whether studios would endorse the deepfake re-casting of their films.

4.2.6 Too much choice is bad. Interestingly, a number of participants took the opportunity to reflect more broadly on recent trends in media, where consumer choice is prioritised over artistic merit. Participants reported frequently feeling pandered to. For example, one participant complained, "why do we need everything to please us?. . . .. it's a pity to see human emotion and capacity and skill being replaced by technology, mainly to 'please'". Another argued, "we don't and shouldn't always get what we want. I feel like this is another step towards a Wall-e type future." Participants here appear to be expressing a general concern for how technological development is being directed. Deepfake movie replacement is seen by these participants as part of throw-away culture (e.g.,[50]), but more broadly as a symptom of a system designed to please rather than excite, challenge, or inspire [51].

Overall, while there certainly were a large number of comments from participants who were interested or excited about the prospect of being able to swap a favoured actors face into an existing film, the majority of comments explored various negative implications of doing so, which ranged from concerns about the quality of acting, empathy for the movie makers and their artistic vision, to thoughts about how the shared social experience could be undermined.

5 Discussion

As expected, participants readily formed false memories of fictitious remakes of popular films. This is in line with much false memory research [23] and demonstrates the malleability of memory when exposed to misinformation. Our primary question, though, was whether the presentation format (deepfake video vs. text) would influence false memory rates. Our data suggests that there is nothing especially potent about deepfake media when it comes to manipulating memory for cinema. This replicates prior work on deepfake fake news stories [34] and other studies on misinformation media. Many studies have demonstrated that misinformation in non-technical forms like simple narratives are extremely effective at distorting memory [25,28,52,53]. One study of autobiographical memory exposed individuals to either a doctored photograph or misleading text that suggested they had taken a hot air balloon ride as a child. They found the misleading text was actually *more* effective at distorting memory, with 80% of those in the text condition forming a false memory, compared to 50% of those who saw a doctored image of themselves in the hot air balloon [54]. This somewhat unintuitive finding may be due to the nature of memory–by providing doctored images or deepfake videos to participants, we supply such concrete information that the participant doesn't need to do any imagining or mental construction of the event, thus they potentially have less information which they might later mistake for a real memory in a source-monitoring error [55]. Our findings suggest that this fluency-based account, which was developed in relation to autobiographical memory,

may also apply to memory for cultural artifacts such as films. This should be further assessed in future research across other kinds of memory.

The implications of our findings for memory research are perhaps not as simple as suggesting we may be overestimating the memory distortion effect of deepfakes. We would note that the false memory rate for the deepfake prompts were still quite high (49% on average). Instead, we suggest that we might be underestimating how readily our memories can be distorted without any technological input. A great deal of research has demonstrated that relatively simple methods can be extremely effective at planting false memories, such as encouraging participants to imagine an event [31] or reading a headline describing an event [27]. Despite this wealth of evidence, we as researchers might still assume that doctored media would be a more effective means of distorting technology [10].

In light of this evidence, it is interesting to reflect on the alarmist discourse around deepfakes and misinformation. Some have predicted that deepfakes will bring about an 'infocalypse' [56]and may be particularly likely to distort our memories of the past (Liv & Greenbaum, 2020). Why have these narratives proliferated without much in the way of empirical evidence? Broinowski [22] argued that moral panics around emerging technology are nothing new–from fears that the written word would destroy human memory and that lightbulbs may cause blindness. We often project our current fears and negative experiences when we attempt to predict how we will use technology in the future [57]. It is perhaps not surprising then, in this 'post-truth' age of 'infodemics' and 'fake news' [58–60] that we have established a shared narrative that casts deepfakes as a threat to democracy and society. Without a doubt, deepfakes have the potential to misinform and to cause real harm, but more empirical evidence is required before we can quantify these harms, weigh the benefits, and intervene where necessary.

Our qualitative data revealed a wide variety of perspectives on the use of deepfakes in cinema. We were impressed with the depth and consideration of participant responses, with many considering the potential harms; to actors, film makers, and the art itself. There is already some scholarship predicting the not-too-distant future where deepfakes become widespread and considering the many ethical and legal implications of 'resurrection technology' where deceased actors are cast in films [61] though our participants raised wider issues worthy of consideration. Our participants also reflected on the nature of cinema and the potential disruptive effects of deepfake recasting, voicing concerns about losing the shared social experience of cinema and their concerns that so much choice would not be enjoyable. Positive remarks were far less frequent than negative remarks, reflecting the negative discourse in academic literature around deepfakes [22]. However, the focus on the negatives may in part have been due to the study design. By priming participants to consider the misinformation risks inherent in deepfake technology and attempting to distort their own memories for films, they may have more easily conjured negative applications rather than positives.

5.1 Implications

The primary implication of the current study for research is the need for more empirical investigations of deepfake technology. Though there are a great many published papers discussing potential regulation and mitigation methods for deepfakes [62], we cannot hope to mitigate against harms that we have not yet quantified. While some have suggested deepfakes may have a uniquely powerful effect on memory [10] further research is needed to explore this across different domains, assessing deepfakes covering different topics. In particular, this research should draw on existing theories and frameworks of memory distortion, such as the source-monitoring framework [30] to establish how and when deepfakes may distort memory. We

also support growing calls for a greater focus on the potential positive applications of deepfakes, encouraging researchers to consider the potentially beneficial uses alongside the predicted harms [63]. As outlined by Broinowski [22], utopian deepfake predictions can be ameliorative (improving everyday life), transformative (reshaping reality) and natural (they are a logical extension of existing social systems). There is existing evidence of helpful and playful applications of deepfakes for learning (creating videos of historical figures to teach students), wellbeing (creating AI-generated fitness coaches that resemble someone you look up to —or a fitter version of yourself), and medicine (allowing dementia patients to interact with younger versions of their friends and family that they may remember) [7,64], though much of this research is speculative and descriptive rather than empirical.

Our findings also have implications for design, particularly in relation to participant concerns that the ability to recast films would offer too much choice. Indeed, even without the abundance of options offered by deepfake recasting, streaming viewers already report being paralysed by choice. The 'paradox of choice' reflects human tendencies to get overwhelmed by too many choices and either giving up or making poor choices [65]. Research suggests an average Netflix viewer loses interest after 60–90 seconds of searching, having reviewed 10–20 titles, after which the risk of abandoning the service increases substantially [66]. In 2021, Netflix launched a 'Play Something' feature, a shuffle option where Netflix will select something for you to watch. This feature allows the viewer to bypass all decision making and its very existence speaks to the pressure and paralysis that can accompany too much choice [67]. Deepfake recasting then is perhaps a risky addition to the viewing experience, as viewers would be required not only to select the film or series they would like to watch, but to cast the show as well. This has implications for the design of such features, where perhaps limiting the options would somewhat paradoxically improve the viewer experience.

5.2 Limitations & future research

A limitation of our false memory experiment is that the experience of viewing a deepfake clip in an online study is likely different to how participants may be exposed to deepfakes in the wild. Our participants saw the deepfake only once, while research shows that repeated exposure to misinformation typically increases belief–the illusory truth effect [68,69]. Source credibility can also affect misinformation susceptibility [70] and participants may make judgements about deepfakes encountered on social media or received from friends that may differ from those within the confines of a controlled psychology experiment. Many other factors may also play a role in susceptibility to deepfake movie remakes (such as having seen the original, how recently it was released, etc.), as previous research has suggested factors such as interest and knowledge can affect false memory susceptibility [71,72]. However, the current study represents an important step in establishing the baseline risks of memory distortion as a consequence of deepfake exposure. As noted above, while the generally negative responses in the qualitative responses to our study is similar in tone to the typical pessimistic academic publications and media coverage of deepfakes [22] the design of our study may have encouraged participants to consider the negative applications of deepfakes such as misinformation. The use of an online survey also meant that we couldn't further interrogate participants about their responses or tease out their reasons further, though we would note that many participants gave thoughtful and considered answers reflecting complex feelings about the subject.

The current study considers deepfakes in just one domain (cinema) and we make no claims that our findings would generalise to other applications. Further research is required to understand user perspectives on the inclusion of deepfakes in education, marketing, gaming and the many other fields where the potential benefits of deepfakes have been discussed [3,7]. In

addition, the effects of deepfakes on viewers' beliefs, memories and behaviours needs to be considered across a wide variety of settings and subjects. As deepfake technology become ever more accessible, with apps that require less expertise and are cheaper than before, it is vital that we gather a broader evidence base to understand this technology which is likely to become a part of everyday life.

6 Conclusion

The findings of the current study suggest that with regard to cinema, deepfakes may not have the powerful effect on misleading memory that some have suggested [10]. Though deepfakes are currently being used in the creative arts and there is speculation as to how ubiquitous and democratised the technology may become [3,38], our participants expressed some reservations about such a future, while also considering the potential benefits. Deepfakes are sometimes positioned as an 'alien technology' and are the subject of growing social concern [63]. We support growing calls to understand deepfakes as a cultural technology, where social concerns and fears should be engaged with critically and any interventions or regulations should be evidence-based.

Acknowledgments

We would like to acknowledge the students of UCC Applied Psychology who assisted with sourcing and selecting the materials and with data collection. This project was supported by a grant from Lero, the Science Foundation Ireland Centre for Software Research.

Author Contributions

Conceptualization: Gillian Murphy.

Data curation: Gillian Murphy.

Formal analysis: Didier Ching, John Twomey, Conor Linehan.

Funding acquisition: Gillian Murphy.

Investigation: Gillian Murphy.

Methodology: Gillian Murphy, Conor Linehan.

Project administration: Gillian Murphy.

Supervision: Gillian Murphy, Conor Linehan.

Writing – original draft: Gillian Murphy.

Writing – review & editing: Gillian Murphy, Didier Ching, John Twomey, Conor Linehan.

References

1. Celli F. Deepfakes are coming: Does Australia come prepared?. Canberra L. Rev., 2020; 17, 193.
2. Sylvester S. 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. 2021; 73(3), 369–392.
3. Kietzmann J, Lee LW, McCarthy IP, Kietzmann TC. Deepfakes: Trick or treat?. Business Horizons. 2020; 63(2), 135–146.
4. Nightingale SJ, Farid H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. Proceedings of the National Academy of Sciences. 2022; 119(8), e2120481119. https://doi.org/10.1073/pnas.2120481119 PMID: 35165187

5. Ajder H, Patrini G, Cavalli F, Cullen L. The State of Deepfakes: Landscape, Threats, and Impact. Deeptrace. 2019.
6. Kikerpill K. Choose your stars and studs: the rise of deepfake designer porn. Porn Studies. 2020; 7(4), 352–356.
7. Westerlund M. The Emergence of Deepfake Technology: A Review. Technology Innovation Management Review. 2019.
8. Chesney B, Citron D. Deep fakes: A looming challenge for privacy, democracy, and national security. Calif. L. Rev. 2019; 107, 1753.
9. Maras MH, Alexandrou A. Determining authenticity of video evidence in the age of artificial intelligence and in the wake of Deepfake videos. The International Journal of Evidence & Proof. 2019; 23(3), 255–262.
10. Liv N, Greenbaum D. Deep fakes and memory malleability: False memories in the service of fake news. AJOB neuroscience. 2020; 11(2), 96–104. https://doi.org/10.1080/21507740.2020.1740351 PMID: 32228386
11. Vaccari C, Chadwick A. Deepfakes and disinformation: exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media+ Society. 2020; 6(1).
12. Meskys E, Kalpokiene J, Jurcys P, Liaudanskas A. Regulating deep fakes: legal and ethical considerations. Journal of Intellectual Property Law & Practice. 2020; 15(1), 24–31.
13. De Ruiter A. The distinct wrong of deepfakes. Philosophy & Technology. 2021; 34(4), 1311–1332.
14. Rogan A. JFK Video: Hear Kennedy's 'lost' Dallas speech in his own voice. The Times. 2018 March 15th. Available from https://www.thetimes.co.uk/article/jfk-video-hear-kennedys-lost-dallas-speech-in-his-own-voice-xtkvhm255.
15. King J. Deepfake Luke Skywalker is another step down a ghoulish CGI path. GQ. 2022 February 10. Available from: https://www.gq-magazine.co.uk/culture/article/boba-fett-luke-skywalker.
16. Lee D. Deepfake Salvador Dali takes selfies with museum visitors. The Verge. 2019 May 10. Available from: https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum.
17. Mihailova M. To dally with Dalí: Deepfake (Inter) faces in the art museum. Convergence. 2021; 27(4), 882–898.
18. Beridze I, Butcher J. When seeing is no longer believing. Nature Machine Intelligence. 2019; 1(8), 332–334.
19. Guarnera L, Giudice O, Battiato S. Fighting deepfake by exposing the convolutional traces on images. IEEE Access. 2020; 8, 165085–165098.
20. Kwok AO, Koh SG. Deepfake: a social construction of technology perspective. Current Issues in Tourism. 2021; 24(13), 1798–1802.
21. Ascot T. Microfake: How small-scale deepfakes can undermine society. Journal of Digital Media & Policy. 2020; 11(2), 215–222.
22. Broinowski A. 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. 2022; 7(1), 109–139.
23. Loftus EF. Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. Learning & memory. 2005; 12(4), 361–366. https://doi.org/10.1101/lm.94705 PMID: 16027179
24. Loftus E, Palmer JC. Reconstruction of automobile destruction: An example of the interaction between language and memory. Journal of verbal learning and verbal behavior. 1974; 585–589.
25. Hyman IE, Husband TH, Billings FJ. False memories of childhood experiences. Applied cognitive psychology. 1995; 9(3), 181–197.
26. Wade KA, Garry M, Don Read J, Lindsay DS. A picture is worth a thousand lies: Using false photographs to create false childhood memories. Psychonomic bulletin & review. 2002; 9(3), 597–603.
27. Murphy G, Loftus EF, Hofstein Grady R, Levine LJ, Greene CM. False Memories for Fake News During Ireland's Abortion Referendum. Psychological Science. 2019; 1449–1459. https://doi.org/10.1177/0956797619864887 PMID: 31432746
28. Greene CM, Nash RA, Murphy G. Misremembering Brexit: Partisan bias and individual predictors of false memories for fake news stories among Brexit voters. Memory. 2021; 587–604. https://doi.org/10.1080/09658211.2021.1923754 PMID: 33971789
29. Calvillo DP, Harris JD, Hawkins WC. Partisan bias in false memories for misinformation about the 2021 US Capitol riot. Memory. 2023; 31(1), 137–146.

30. Johnson MK, Hashtroudi S, Lindsay DS. Source monitoring. Psychological bulletin. 1993; 114(1), 3. https://doi.org/10.1037/0033-2909.114.1.3 PMID: 8346328

31. Garry M, Polaschek DL. Imagination and memory. Current Directions in Psychological Science. 2000; 9(1), 6–10.

32. Wade KA, Green SL, Nash RA. Can fabricated evidence induce false eyewitness testimony?. Applied Cognitive Psychology. 2010; 24(7), 899–908.

33. Nash RA, Wade KA. Innocent but proven guilty: Eliciting internalized false confessions using doctored-video evidence. Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition. 2009; 23(5), 624–637.

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

35. Paris B, Donovan J. Deepfakes and cheap fakes. United States of America: Data & Society. 2019; 1.

36. Lucas KT. Deepfakes and domestic violence: perpetrating intimate partner abuse using video technology. Victims & Offenders. 2022; 17(5), 647–659.

37. Albahar M, Almalki J. Deepfakes: Threats and Countermeasures Systematic Review. Journal of Theoretical and Applied Information Technology. 2019; 3242–3250.

38. Castoro R. Deepfakes Could Mean Better Movies but a Worse Society. Collider. 2019 November 7th. Available from https://collider.com/deepfakes-explained-better-movies-worse-society/.

39. Braun V, Clarke V. Using thematic analysis in psychology. Qualitative research in psychology. 2006; 3 (2), 77–101.

40. Clarke V, Braun V, Hayfield N. Thematic analysis. Qualitative psychology: A practical guide to research methods. 2015; 222(2015), 248.

41. Greene C, Murphy G, de Saint Laurent C, Prike T, Hegarty K, Ecker U. Best Practices for Ethical Conduct of Misinformation Research: A Scoping Review and Critical Commentary. European Psychologist. 2022.

42. Murphy G, Loftus E, Grady RH, Levine LJ, Greene CM. Fool me twice: How effective is debriefing in false memory studies?. Memory. 2020; 28(7), 938–949. https://doi.org/10.1080/09658211.2020.1803917 PMID: 32762497

43. Greene CM, Murphy G. Debriefing works: Successful retraction of misinformation following a fake news study. PloS one. 2023; 18(1):e0280295. https://doi.org/10.1371/journal.pone.0280295 PMID: 36662686

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

45. Ralston R. Separating the art from its artist: Film reviews in the era of# MeToo. Poetics. 2022; 101653.

46. Kirman B, Linehan C, Lawson S. Exploring mischief and mayhem in social computing or: how we learned to stop worrying and love the trolls. CHI'12 Extended Abstracts on Human Factors in Computing Systems. 2021, 121–130.

47. Harris D. Deepfakes: False pornography is here and the law cannot protect you. Duke L. & Tech. Rev. 2019; 17, 99.

48. Van de Vijver L. The cinema is dead, long live the cinema! Understanding the social experience of cinema-going today. Participations. 2017; 14(1), 129–144.

49. Gieseke AP. "The New Weapon of Choice": Law's Current Inability to Properly Address Deepfake Pornography. Vand. L. Rev. 2020; 73, 1479.

50. Whiteley N. Toward a throw-away culture. Consumerism,'style obsolescence'and cultural theory in the 1950s and 1960s. Oxford Art Journal. 1987; 10(2), 3–27.

51. Horkheimer M, Adorno TW. The culture industry: Enlightenment as mass deception. In Karl Marx. 2017; pp. 405–424. Routledge.

52. Loftus EF, Pickrell JE. The formation of false memories. Psychiatric annals. 1995; 25(12), 720–725.

53. Murphy G, Dawson CA, Huston C, Ballantyne L, Barrett E, Cowman CS, Fitzsimons C, Maher J, Ryan KM, Greene CM. Lost in the mall again: a preregistered replication and extension of Loftus & Pickrell (1995). Memory. 2023; 5:1–3.

54. Garry M, Wade KA. Actually, a picture is worth less than 45 words: Narratives produce more false memories than photographs do. Psychonomic bulletin & review. 2005; 12(2), 359–366.

55. Garry M, Gerrie MP. When photographs create false memories. Current Directions in Psychological Science. 2005; 14(6), 321–325.

56. Lenters M. "Deepfakes: Danger to Democracy or Creativity for All?" Innovating Origins. 2021 March 30. https://innovationorigins.com/deepfakes-danger-to-democracy-or-creativity-for-all/.

57. Natale S, Balbi G. Media and the imaginary in history: The role of the fantastic in different stages of media change. Media History. 2014; 20(2), 203–218.

58. Lazer DM, Baum MA, Benkler Y, Berinsky AJ, Greenhill KM, Menczer F, Metzger MJ, Nyhan B, Pennycook G, Rothschild D, Schudson M. The science of fake news. Science. 2018; 359(6380):1094–6. https://doi.org/10.1126/science.aao2998 PMID: 29590025

59. Greene CM, Murphy G. Quantifying the effects of fake news on behavior: Evidence from a study of COVID-19 misinformation. Journal of Experimental Psychology: Applied. 2021; 27(4), 773. https://doi.org/10.1037/xap0000371 PMID: 34110860

60. Zarocostas J.. How to fight an infodemic. The lancet. 2020; 395(10225), 676. https://doi.org/10.1016/S0140-6736(20)30461-X PMID: 32113495

61. Lees D, Bashford-Rogers T, Keppel-Palmer M. The digital resurrection of Margaret Thatcher: Creative, technological and legal dilemmas in the use of deepfakes in screen drama. Convergence. 2021; 27(4):954–73.

62. Caldera E. Reject the evidence of your eyes and ears: deepfakes and the law of virtual replicants. Seton Hall L. Rev. 2020; 50, 177.

63. Cover R. Deepfake culture: the emergence of audio-video deception as an object of social anxiety and regulation. Continuum. 2022; 1–13.

64. Pataranutaporn P, Danry V, Leong J, Punpongsanon P, Novy D, Maes P, Sra M. AI-generated characters for supporting personalized learning and well-being. Nature Machine Intelligence. 2021; 3(12):1013–22.

65. Schwartz B, Ward A. Doing better but feeling worse: The paradox of choice. Positive psychology in practice. 2004; 86–104.

66. Gomez-Uribe CA, Hunt N. The netflix recommender system: Algorithms, business value, and innovation. ACM Transactions on Management Information Systems (TMIS). 2015; 6(4), 1–19.

67. Rodríguez Ortega V. 'We Pay to Buy Ourselves': Netflix, Spectators & Streaming. Journal of Communication Inquiry. 2022; 01968599211072446.

68. Fazio LK. Repetition increases perceived truth even for known falsehoods. Collabra: Psychology. 2020; 6(1).

69. Pennycook G, Cannon TD, Rand, DG. Prior exposure increases perceived accuracy of fake news. Journal of experimental psychology: general. 2018; 147(12), 1865.

70. Traberg CS, van der Linden S. Birds of a feather are persuaded together: Perceived source credibility mediates the effect of political bias on misinformation susceptibility. Personality and Individual Differences. 2022; 185, 111269.

71. O'Connell A, & Greene CM. Not strange but not true: Self-reported interest in a topic increases false memory. Memory. 2017, 25(8), 969–977. https://doi.org/10.1080/09658211.2016.1237655 PMID: 27710198

72. Greene CM, Murphy G. Individual differences in susceptibility to false memories for COVID-19 fake news. Cognitive research: principles and implications. 2020; 5(1), 1–8.

Appendix III: Encountering Deepfakes: A Thematic Analysis of Comments on a Music Video featuring Deepfake Content

Link to published article: https://dl.acm.org/doi/abs/10.1145/3757406



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

DIDIER CHING, School of Applied Psychology, University College Cork, Cork, Ireland. Lero, Science Foundation Ireland research Centre for Software, Limerick, Ireland.
JOHN TWOMEY, School of Applied Psychology, University College Cork, Cork, Ireland. Lero, Science Foundation Ireland research Centre for Software, Limerick, Ireland.
CONOR LINEHAN, School of Applied Psychology, University College Cork, Cork, Ireland. Lero, Science Foundation Ireland research Centre for Software, Limerick, Ireland.
GILLIAN MURPHY, School of Applied Psychology, University College Cork, Cork, Ireland. Lero, Science Foundation Ireland research Centre for Software, Limerick, Ireland.

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 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.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; *Empirical studies in HCI*

Additional Key Words and Phrases: Deepfake, Entertainment

ACM Reference format:
Didier Ching, John Twomey, Conor Linehan, and Gillian Murphy. 2025. Encountering Deepfakes: A Thematic Analysis of Comments on a Music Video Featuring Deepfake Content.. *Proc. ACM Hum.-Comput. Interact.*, 9, 7, Article CSCW225 (November 2025), 23 pages, https://doi.org/10.1145/3757406

1 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 [18]. 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 [22].

Author's Contact Information: Didier Ching, University College Cork, Cork, Ireland, 116378223@umail.ucc.ie; John Twomey, University College Cork, Cork, Ireland, 118447782@umail.ucc.ie; Conor Linehan, University College Cork, Cork, Ireland, conor.linehan@ucc.ie; Gillian Murphy, University College Cork, Cork, Ireland, gillian.murphy@ucc.ie.



This work is licensed under Creative Commons Attribution International 4.0.
© 2025 Copyright is held by the owner/author(s).
ACM 2573-0142/2025/11 – CSCW225.
https://doi.org/10.1145/3757406

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 [2, 87], manipulating election results through the defamation of political candidates [19], and undermining epistemic trust in society due to citizens being unable to differentiate between real and false information [31]. 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 [5, 12] and non-consensual deepfake porn of mostly female celebrities which can cause psychological and reputational harm [56].

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 itis 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 [48] and deepfakes incorporated into awareness and educational videos on YouTube have almost 10 million views [17]. 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 [23, 34], music videos [40, 48, 79] and museums [62], 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 [48]. 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." [73]. 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.

2 Background

2.1 Research on Deepfakes

Deepfakes" are a form of digital video manipulation which first emerged on Reddit in 2017 [22]. 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 [2, 82]. 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 [2, 19, 31, 87]. Those papers often do not involve empirical analyses and discuss deepfake technology in legal [19], philosophical [31], and societal [87] contexts. Existing research often discusses the potential of deepfake technology to spread misinformation. For example, Chesney and Citron [19] 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 [19] 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 [1, 2, 87].

Empirical work on deepfake technology often focuses on detection and countermeasures against the aforementioned threats of deepfakes [52, 55, 65]. 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 [55]. 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 [27, 50]. For example, Dobber et al. [27] 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 [1]. Current research in this area focuses on identifying the harms of deepfake technology against women [36, 56] and how legislation is currently equipped to deal with this issue [35, 43, 47]. Gosse and Burkell [36] 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 [69]. Some HCI and CSCW literature have adopted this perspective in their research as well [24, 57] 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 [8]. 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 [34], 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.

2.2 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 [34]. Fans reportedly had mixed responses to seeing a de-aged Mark Hamill [23, 66].

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 [62]. 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 AKand "Self" by Steven Wilson utilising deepfake technology to superimpose faces of celebrities onto their own whilst singing [40, 79]. Both videos have over 500,000 views at the time of writing and gained considerable publicity [7, 37]. Responses to these videos were generally positive, but some acknowledged the controversial nature of the technology [3].

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., [63] 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 Ms 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. [63] encourages further research into understanding deepfakes as a cultural phenomenon rather than viewing deepfakes as a uniquely advancing "alien technology".

2.3 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 [53, 54]. 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 [64], or when the late rapper Tupac performed with Snoop Dogg during the Coachella festival in 2012 [64]. Sherlock [77] 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-aging 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., [53] 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. [53] 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 [68] and ultimately disrespects the dignity of the deceased [13]. 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 [25], some people still expressed discomfort to seeing the deceased victim being manipulated [54]. 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 Methods

3.1 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[48], 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 [38, 68, 78].

3.2 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 [32, 46, 58]. 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.

3.3 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 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 6comments 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 ÷ 9000 ≈ 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.

3.5 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 [10]. 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 [76] 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. [21]. 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 [11].

Guidance from Clarke et al. [21] 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 [21].

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 [10]. 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 [9] 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 [11].

4 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.

4.1 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 [2, 87].

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 [20] and "rotoscoping" in Celine Dion's duet with Elvis on "American Idol" in 2007 [8] 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 [81] 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. [41] 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.

4.2 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 [30]. 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 [84]. 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.

4.3 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 [31] and to defame individuals [19].

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. [58] 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 [19, 87].

4.4 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 [62]. 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 [62]. Similarly, deepfakes in the music video serve to invoke empathy in viewers, as well as produce a more vivid emotional experience for the viewer.

4.5 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 [62] 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 [6].

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 [74] 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 [13] 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.

4.6 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 [16]. While the hypothetical harms of deepfakes focus on the total erosion of truth in society [19], 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 [63], 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 [45]. 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 [24], 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 [70]. 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.

5 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.

5.1 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 were unaware of the capabilities of deepfake technology discussed it in a completely different manner. Responses were often characterized by confusion, doubt, and uncertainty. Commenters seemed to be doubting their senses and reality as they can't comprehend what they are seeing, leading to use of strategies such as familiarization [75] to attempt to explain away the confusion. Surprisingly, some commenters did not even notice the deepfakes and simply stated how similar Kendrick Lamar looked to the celebrities that were being deepfaked. Most interestingly, we did not find a single comment from someone who was unaware of deepfake technology who suspected the visuals were fake or edited in some way. 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 [51]. 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 what they were seeing was real because they did not have any reason to think they were being deceived. This theory and subsequent supporting research are typically based off individuals receiving textual and verbal

information/misinformation 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. Those who were less digitally literate and unaware of the existence of deepfake technology were often left confused and doubting themselves, as well as having 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 [39]. 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 [57]) or host educational programmes among their users to improve their literacy skills, which Google and Facebook have done [49]. However, it is also important to raise awareness of deepfake technology without undermining individual's trust in legitimate media [83]. Therefore, a balance must be achieved with informing the public without overexaggerating the impact of deepfakes.

5.2 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 [62] or with Buzzfeed's video "You Won't Believe What Obama Says In This Video!" [17]. 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 [29] and can enhance the overall reception of the music [14, 15]. 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 [66]. 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 [62] or in the Gipsoteca Museum [88], but the entertainment and educational value of these applications have yet to be empirically tested.

5.3 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 research papers [1,2,19,80], 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 [12] or emulate friends or family to solicit funds [5]. 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 [4, 86]. This strike highlights issues surrounding the licensing 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 [35, 56]. 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 [2, 19].

Legal and ethical considerations surrounding deepfake technology in entertainment were also overlooked by commenters who focused on negative aspects of deepfakes. These are imminent

issues that may concern society as a whole, not just actors or individuals in the entertainment industry. The current inability of legal systems to protect individual's likeness online and prevent creators of deepfakes from manipulating their identities in an entertainment setting can be attributed to multiple factors. Firstly, any copyright of one's identity and likeness would be difficult to enforce in a court of law as creators of those deepfakes can cite statutory exemptions such as satire or parody [59]. Secondly, even if a copyright infringement case was taken to trial, the deepfake creator may also be able to invoke a fair use defence depending on several factors of the deepfake [59]. This is assuming that the original creator of the deepfake can be identified in the first place. While there may be legal recourse if the deepfake ultimately caused mental or emotional harm, the law is not equipped to deal with cases surrounding deepfakes of deceased individuals as no mental or emotional harm has come to the deceased [85]. Therefore, issues surrounding deepfakes of deceased individuals are currently ethical ones surrounding the human dignity of the deceased and whether creating deepfakes of deceased individuals is disrespectful and offensive. However, discussion surrounding the ethical and legal issues surrounding deepfake technology were comparatively little compared to the exaggerated dangers of deepfakes, further highlighting a divide between what individuals think is dangerous about deepfake technology, and what harms deepfake technology realistically brings.

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 [16] 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 [43] hypothesis 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.

5.4 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., [76]. 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 [76]. 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.

6 Conclusion

The 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)

References

[1] Henry Ajder, Giorgrio Patrini, Francesco Cavalli, and Laurence Cullen. 2019. The State of Deepfakes: Landscape, Threats, and Impact. Amsterdam: Deeptrace.

[2] Marwan Albahar and Jameel Almalki. 2019. Deepfakes: Threats and Countermeasures Systematic Review. Journal of Theoretical and Applied Information Technology, (November 2019), 3242-3250.

[3] 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

[4] Janelle Ash. 2023. George Clooney supports actor's union strike, says it's an 'inflection point in our industry'. Fox News, Retrieved November 26, 2023 from https://www.foxnews.com/entertainment/george-clooney-supports-actors-union-strike-inflection-point-industry

[5] Divya Bhati. Kerala man loses Rs 40,000 to AI-based Deepfake WhatsApp fraud . India Today. Retrieved December 9, 2023 from 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

[6] James Bingaman. 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, (December 2022) 86(2), 364-381. https://doi.org/10.1177/0030222820971531

[7] Jon Blistein. 2021. Steven Wilson Becomes Bowie, Biden, Trump, and More in New 'Self' Video. Rolling Stone, Retrieved August 28, 2023 from https://www.rollingstone.com/music/music-news/steven-wilson-self-music-video-deep-fake-technology-1121914/

[8] Susanne Bødker. 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).

[9] Alexandrine Boudreault-Fournier. 2016. Microtopia in Counterpoint: Relational Aesthetics and the Echo Project. Cadernos de arte e antropologia, (April 2016), 5(1), 135-154. https://doi.org/10.4000/cadernosaa.1060

[10] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. Qualitative Research in Psychology, (2006), 3(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

[11] Virginia Braun and Victoria Clarke. 2013. Successful qualitative research: A practical guide for beginners.

[12] Thomas Brewster. Fraudsters Cloned Company Director. Forbes. Retrieved December 9, 2023 from https://www.forbes.com/sites/thomasbrewster/2021/10/14/huge-bank-fraud-uses-deep-fake-voice-tech-to-steal-millions/

[13] Lisa Bode. 2007. 'Grave Robbing'or 'Career Comeback'? On the Digital Resurrection of Dead ScreenStars. History of Stardom Reconsidered, (2007), 36-40.

[14] Marilyn Boltz. 2013. Music videos and visual influences on music perception and appreciation: Should you want your MTV. The psychology of music in multimedia, (June 2013), 217-234.

[15] Marilyn Boltz, Brittany Ebendorf, and Benjamin FIeld. 2009. Audiovisual interactions: The impact of visual information on music perception and memory. Music Perception, (September 2009), 27(1), 43-59

[16] Anna Broinowski. 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 (May 2022): 109-139. https://doi.org/10.5325/jasiapacipopcult.7.1.0109

[17] Buzzfeed. 2018. "You Won't Believe What Obama Says In This Video!". Retrieved March 29, 2023 from https://www.youtube.com/watch?v=cQ54GDm1eL0&ab_channel=BuzzFeedVideo

[18] R. Chawla. 2019. Deepfakes: How a pervert shook the world. International Journal of AdvanceResearch and Development, (2019), 4-8.

[19] Bobby Chesney and Danielle Citron. 2019. Deep fakes: A looming challenge for privacy, democracy, and nationalsecurity. Calif. L. Rev., 1753-1819. https://doi.org/10.15779/Z38RV0D15J

[20] Elizabeth Chin. 2011. Michael Jackson's panther dance: double consciousness and the uncanny business of performing while black. Journal of Popular Music Studies, (March 2011), 23(1), 58-74.

[21] Victoria Clarke, Virginia Braun, and Nikki Hayfield. 2015. Thematic analysis. Qualitative psychology: A practical guide to research methods, (2015), 3, 222-248.

[22] Samantha Cole. 2017. AI-Assisted Fake Porn is Here and We're All Fucked.Vice. Retrieved July 22, 2022 from https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

[23] James David. 2022. 'Star Wars' fans divided on whether they want to see deepfake Luke Skywalker again. WGTC. Retrieved 12 September, 2023 from https://wegotthiscovered.com/tv/star-wars-fans-divided-on-whether-they-want-to-see-deepfake-luke-skywalker-again/

[24] Yue Deng, Zheng Chen, Changyang He, Zhicong Lu, Z, and Bo Li. 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).

[25] Adrienne De Ruiter. 2021. The distinct wrong of deepfakes. Philosophy & Technology, (December 2021), 34(4),1311-1332.

[26] Ann-Christine Diaz. 2020. "PARKLAND VICTIM JOAQUIN OLIVER COMES BACK TO LIFE INHEARTBREAKING PLEA TO VOTERS". Adage. Retrieved January 12, 2023 from https://adage.com/article/advertising/parkland-victim-joaquin-oliver-comes-back-life-heartbreaking-plea-voters/2285166

[27] Tom Dobber, Nadia Metoui, Damian Trilling, Natali Helberger, and Claes de Vreese. 2021. Do(microtargeted) deepfakes have real effects on political attitudes? The International Journal of Press/Politics, (January 2021), 69-91. https://doi.org/10.1177/1940161220944364

[28] Denver D'Rozario. 2016. The market for 'Delebs' (dead celebrities): A revenue analysis. Journal of Customer Behaviour, (December 2016), 15(4), 395-414.

[29] Brittany Ebendorf. 2007. The impact of visual stimuli on music perception (Doctoral dissertation).

[30] Francesca Emiliani and Stefano Passini. 2016. Everyday Life in Social Psychology. Journal for the Theory of Social Behaviour 47, 1 (2016), 83-97. DOI: https://doi.org/10.1111/jtsb.12109

[31] Don Fallis. 2020. The epistemic threat of deepfakes. Philosophy & Technology, (August 2020), 1-21. https://doi.org/10.1007/s13347-020-00419-2

[32] Amy Ferrey, Georgia Ashworth, Mark Cabling, Gabriella Rundblad, and Khalida Ismail. 2023. Athematic analysis of YouTube comments on a television documentary titled 'Diabulimia: The World's most dangerous eating disorder'. Diabetic Medicine, (May 2023), 40(5), e15025., DOI:10.1111/dme.15025

[33] Dilrukshi Gamage, Jiayu Chen, and Kazutoshi Sasahara. 2021. The Emergence of Deepfakes and its Societal Implications: A Systematic Review. Proceedings of the 2021 Truth and Trust Online Conference 1, 1 (October 2021), 28-39.

[34] Dilrukshi Gamage, Piyush Ghasiya, Vamshi Bonagiri, Mark Whiting, and Kazutoshi Sasahara, 2022. Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. In CHI Conference on Human Factors in Computing Systems, (April 2022), 1-19. DOI: https://doi.org/10.1145/3491102.3517446

[35] Anne Pechenik Gieseke. 2020. "The New Weapon of Choice" Law's Current Inability to Properly Address Deepfake Pornography. Vand. L. Rev., (2020), 1479-1514.

[36] Chandell Gosse and Jacquelyn Burkell, 2020. Politics and porn: how news media characterizes problems presented by deepfakes. Critical Studies in Media Communication, (October 2020), 37(5), 497-511. DOI:10.1080/15295036.2020.1832697

[37] Jon Hadusek. 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/

[38] Marc Hogan. 2022. "How Kendrick Lamar's "The Heart Part 5" Video Subverts Deepfake Technology". Pitchfork. Retrieved 28 July 2022 from https://pitchfork.com/thepitch/kendrick-lamar-the-heart-part-5-video-deepfake-technology/

[39] Hwang, Yoori, Ji Youn Ryu, and Se-Hoon Jeong. 2021. Effects of disinformation using deepfake: The protective effect of media literacy education. Cyberpsychology, Behavior, and Social Networking 188-193

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

[41] Iikka Kaate, Joni Salminen, Soon-Gyo Jung, Hind Almerekhi, and Bernard J. Jansen. 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).

[42] Rob Jenner. 2022. "How Kendrick Lamar and Nipsey Hussle's Yearslong Friendship Informed "The Heart Part 5"". GQ. Retrieved 29 August, 2022 from https://www.gq.com/story/kendrick-lamar-nipsey-hussle-the-heart-part-5

[43] Deborah G. Johnson and Mario Verdicchio. 2017. AI Anxiety. Journal of the Association for Information Science and Technology 68, 9 (2017), 2267-2270. DOI: https://doi.org/10.1002/asi.23867

[44] Vasileia Karasavva and Aalia Noorbhai. 2021. The Real Threat of Deepfake Pornography: A Review of Canadian Policy. Cyberpsychology, Behavior, and Social Networking, (March 2021), 203-209. https://doi.org/10.1089/cyber.2020.0272

[45] Kristjan Kikerpill. 2020. Choose your stars and studs: the rise of deepfake designer porn. Porn Studies, (October2020), 7(4), 352-356. https://doi.org/10.1080/23268743.2020.1765851

[46] Clare King and Darragh McCashin. 2022. Commenting and connecting: A thematic analysis of responses to YouTube vlogs about borderline personality disorder. Internet interventions, (April 2022), 28, 100540.

[47] Matthew Kugler and Carly Pace. 2021. Deepfake privacy: Attitudes and regulation. Nw. UL Rev. (2021).

[48] Kendrick Lamar. 2022. Kendrick Lamar - The Heart Part 5. Retrieved May 31, 2022 from: https://www.youtube.com/watchv=uAPUkgeiFVY&ab_channel=KendrickLamarVEVO

[49] Dami Lee. "Facebook launches Digital Literacy Library to help young people use the internet responsibly." The Verge, August 2, 2019. https://www.theverge.com/2018/8/2/17643586/facebook-digital-literacy-library-educators-schools

[50] Jiyoung Lee, and Soo Yun Shin. "Something that they never said: Multimodal disinformation and source vividness in understanding the power of ai-enabled deepfake news." Media Psychology (February2022), 1-16. https://doi.org/10.1080/15213269.2021.2007489

[51] Timothy R. Levine. 2014. Truth-Default Theory (TDT). Journal of Language and Social Psychology 33, 4(2014), 378-392. DOI: https://doi.org/10.1177/0261927x14535916

[52] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. 2020. Face x-ray for more general face forgery detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, (2020), 5001–5010

[53] Dominic Lees, Tom Bashford-Rogers, and Marcus Keppel-Palmer. 2021. The digital resurrection of Margaret Thatcher: Creative, technological and legal dilemmas in the use of deepfakes in screen drama. Convergence, (August 2021), 27(4), 954-973.

[54] Hang Lu and Haoran Chu. 2023. Let the dead talk: How deepfake resurrection narratives influence audience response in prosocial contexts. Computers in Human Behavior, (August 2023). https://doi.org/10.1016/j.chb.2023.107761

[55] Falko Matern, Christian Riess, and Marc Stamminger. 2019. Exploiting visual artifacts to expose deepfakes and face manipulations. In 2019 IEEE Winter Applications of Computer Vision Workshops(WACVW), IEEE, (2019) 83–92.

[56] Sophie Maddocks. 2020. 'A Deepfake Porn Plot Intended to Silence Me': exploring continuities between pornographic and 'political' deep fakes. Porn Studies, (October 2020), 7(4), 415-423. https://doi.org/10.1080/23268743.2020.1757499

[57] Rachel Magee, Denise Agosto, Andrea Forte. 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).

[58] Martel, Cameron, and David G. Rand. "Misinformation warning labels are widely effective: A review of warning effects and their moderating features." Current Opinion in Psychology (2023): 101710.

[59] Alisa McLellan, Keana Schmidt-Waselenchuk, Kari Duerksen, and Erica Woodin. 2022. Talking back to mental health stigma: An exploration of YouTube comments on anti-stigma videos. Computers Inhuman Behavior, (June 2022), 131, 107214., https://doi.org/10.1016/j.chb.2022.107214

[60] Meskys, Edvinas, Julija Kalpokiene, Paulius Jurcys, and Aidas Liaudanskas. "Regulating deep fakes: legal and ethical considerations." Journal of Intellectual Property Law & Practice 15, no. 1 (2020): 24-31.

[61] Metaphysic.ai. 2023. Deeptomcruise. TikTok. Retrieved November 26, 2023 from https://www.tiktok.com/@deeptomcruise?lang=en

[62] Mihaela Mihailova. 2021. To dally with Dalí: Deepfake (Inter) faces in the art museum. Convergence, (August 2021), 27(4), 882-898. https://doi.org/10.1177/13548565211029401

[63] Gillian Murphy, Didier Ching, John Twomey, and Conor Linehan. 2023. Face/Off: Changing the face of movies with deepfakes. Plos one, (July 2023), 18(7), e0287503. https://doi.org/10.1371/journal.pone.0287503

[64] Chenda Ngak. 2012. Tupac Coachella hologram: Behind the technology. Retrieved 23 June, 2023 https://www.cbsnews.com/news/tupac-coachella-hologram-behind-the-technology/

[65] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. 2020. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In European Conference on Computer Vision, Springer, (August 2020), 86–103.

[66] Omar Oakes. 2019. 'Deepfake' voice tech used for good in David Beckham malaria campaign. Campaign, retrieved21 September 2024 from https://www.campaignlive.co.uk/article/deepfake-voice-tech-used-good-david-beckham-malaria-campaign/1581378

[67] Ameena Qobrtay. 2022. 'Star Wars' fans love CGI Luke Skywalker, but deepfake implications are ˙ dangerous. Retrieved 24 August, 2023 https://dailytargum.com/article/2022/03/star-wars-fans-love-cgi-luke-skywalker-but-deepfake-implications-are

[68] Konstantin Pantserev. 2020. The Malicious Use of AI-Based Deepfake Technology as the New Threat to Psychological Security and Political Stability. Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity, 37-55. https://doi.org/10.1007/978-3-030-35746-7_3

[69] Pistol Pete. 2022. All of the deepfakes in Kendrick Lamar's "The Heart Part 5" video. Hiphopcanada. Retrieved 19 September 2023 https://hiphopcanada.com/kendrick-lamar-heart-part-5-deepfakes/

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

[71] Shannon Reid. 2021. The deepfake dilemma: Reconciling privacy and first amendment protections. U. Pa. J. Const.L., (2021), 23, 209.

[72] Rebecca Roberts. 2023. You're Only Mostly Dead: Protecting Your Digital Ghost from Unauthorized Resurrection. Federal Communications Law Journal, (2023), 75(2), 273-296

[73] Yvonne Rogers and Paul Marshall. 2017. Research in the Wild. San Rafael: Morgan & Claypool Publishers.

[74] Alexi Sargeant. 2017. The undeath of cinema. The New Atlantis, 53(Summer/Fall), (July 1 2017),17-32.

[75] Ernst Schraube and Charlotte Højholt. 2015. Psychology and the Conduct of Everyday Life. Routledge.

[76] Peter Schultes, Verena Dorner, and Franz Lehner. Leave a comment! An in-depth analysis of user comments on YouTube.

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

[78] Bethy Squires. 2022. Kendrick Lamar Brings Deepfake Deep Truths to 'The Heart Part 5' Music Video. Vulture. Retrieved 28 December 2022, https://www.vulture.com/2022/05/kendrick-lamar-deepfakes-the-heart-part-5-nipsey-hussle.html

[79] Steven Wilson. 2021. Steven Wilson - SELF (Official Video). Retrieved 2 May 2023 from https://www.youtube.com/watch?v=UroVey4fJ_g

[80] Shannon Sylvester. 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

[81] Jan-Philipp Stein and Peter Ohler. 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, (November 2018), 30(6), 480-491.

[82] Erkam Temir. 2020. Deepfake: new era in the age of disinformation & end of reliable journalism. SelçukİletişIm, (January 2020), 13(2), 1009-1024. DOI: 10.18094/JOSC.685338

[83] Twomey, John, Didier Ching, Matthew Peter Aylett, Michael Quayle, Conor Linehan, and Gillian Murphy. "Deep fake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine." Plos one 18, no. 10 (2023): e0291668.

[84] Cristian Vaccari and Andrew Chadwick, "Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news." Social Media+ Society 6, 1(February 2020) https://doi.org/10.1177/2056305120903408

[85] Olivia Wall. "A Privacy Torts Solution to Postmortem Deepfakes." Wash. UL Rev. 100 (2022): 885.

[86] Angela Watercutter and Will Bedingfield. 2023. Hollywood Actors Strike Ends With a Deal That Will Impact AI and Streaming for Decades. Wired, retrieved 26 November 2023 from https://www.wired.com/story/hollywood-actors-strike-ends-aistreaming

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

[88] Matteo Zaramella, Irene Amerini, and Paolo Russo. 2023. Why Don't You Speak?: A Smartphone Application to Engage Museum Visitors Through Deepfakes Creation. In Proceedings of the 5thWorkshop on the analySis, Understanding and proMotion of heritAge Contents (SUMAC '23), November 2, 2023, Ottawa, ON, Canada. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3607542.3617359

Received January 2024; revised October 2024; accepted February 2025.

Appendix IV: Qualitative coding and analysis example in Chapter 5

Partial extract of one of the texts and its analysis – we would have preferred to reproduce the full texts for each interview in the corpus but reproducing entire texts would be outside the bounds of fair use and thus represent a copyright issue.

Text	Genre	Action orientation, situation, construction	Discourse elements
. … Even among my peers, so many of us have had these same exact experience. And it kills me to know that this is an inevitability that future generations will also be victim to. And even worse, because technology advances so quickly. Right now, the laws surrounding deep fake porn and other intimate image abuse are almost non-existent. It has felt hopeless and honestly unsafe to ever speak out. However, I am relieved to finally be talking about this with some support. I was reached out to by 'My image. My choice'. It's a group of very hardworking activists who are openly pushing for change and becoming a resource for the countless people who have been targeted and abused by others online. So I am finally looking into what I can do, especially because I received a very interesting message recently and that's the other part of my story that I wanted to address. A lot of people wrongfully think that deep fake porn is a joke, or that because I post my face publicly, I deserve this. But it can happen to anyone. I was contacted by someone who thinks she has actually discovered the identity of a person who has created deep fakes of her. This person hiding behind a username on one of the biggest deep fake sites was recognised because he made pornographic videos of at least seven people he knows, personally, from his time at school, so not people in the public eye, but classmates. But this same username also made a tonne of videos of me and a dozen of my friends. So she reached out to me and let me know. And she shared his identity with me. I now know his name, where he lives and where he works. And I am so ready to finally be able to pursue whatever action I can. To find some semblance of justice on behalf of all the people he's done this to. Unfortunately, he's obviously just one of many, many offenders who have exploited others in this way. But to me it does feel like a huge step closer in finally fighting back because I don't care if it's just one. I want to let him know that I know. This affects people's lives. It's fully ingrained in mine. It dictates how I act, where I feel safe and is something that I'm forced to come to terms with in order to just continue existing.	(G) Advertising (G) Narrative	(A) Highlights group experience of deepfakes (C) Community of targets which is rapidly expanding (A) Justifying silence as safety (C) Constructs internet people who are shaming (S) responses situated on timeline before and after identifying the perpetrator, changes in self over time (C) Construction of community of targets in solidarity, construction of the deepfake creator, non celebrity targets (C) Anonymity and identity (C) Constructs Deepfake recourse	(P) Positions the experiences as belonging to self and other (My and us), both a personal experience and a communal narrative (ID) Balancing want to speak out with potential for more harm, and lack of actual consequences (P) positioning self within a popular shaming narrative that public celebrities as being uniquely targeted because of lack of image rights, (F) lack of self framed as trade-off of celebrity, power disparity between public figure and anonymous users (F) Anonymity as disguise, username as masks (L) Not just public figures but individual known targets (F) Loss of anonymity of perpetrator - evening out power disparity, action is something which is owed to other victims (F) Deepfakes as having large amount of anonymous perpetrators (F) Deepfake recourses as a stepped process (see[Co-auuthor name] comment on back to square one) (F) Deepfakes affecting multiple different domains of life (GR) (F) "It" is privileged over targets behaviour and safety, framed as inescapable/ a requirement of continuing ones life

Appendix V: Example paragraph of writing up the analysis for Study 4

Community values/attitudes encourage silence/blaming

Public coverage facilitating harms and traumatisation Internet control over narrative -speaking to the internet

In our dataset, victims position themselves and experience but also construct the public discourse on deepfake abuse imagery and frame publicity as drowning out the experiences of victims while privileging the experiences of outside commentators and perpetrators. One distinct part of the public coverage which emerges in relation to online celebrities such as streamers is how the reactions of the perpetrators and other male streamers are privileged over the experiences of the victims. “The debate over our experience as women in this is, not shockingly, amongst men. None of you should care or listen to what any male steamer's "take" is on how we feel”…… “This is not your debate. Stop acting like it is.” This example shows the inconsistency where the coverage of this abuse overly focuses on commentary and reactions over lived experience. The focus on perpetrators also came up in regard to public coverage of deepfakes “I don't know. I think what's frustrating is because I had to do. It's just like so. It's so angry, anger inducing cause like the first thing he did was go live and apologise to his stream. He didn’t apologise to any of the women, he apologised to his fucking stream. And it's like what are you doing?” Public apologies to the general public were framed as especially harmful as it furthers publicity about the incident without any benefit to the target. “I think I could forgive if he would have like called me on a Tuesday and he's like, hey, I looked at this three weeks ago. I'm so sorry. I think I could find a place and like, I could forgive him for that. But the problem is that the platform and the wildfire that it's spread is just like, it feels so unrepairable at this point”. Two distinct harms are framed here, consumption of deepfakes and contributing to the spread of deepfake content. The spreading awareness of the platform and causing the stressful public coverage highlights the role of publicity in the harms of deepfakes. The news publicity around deepfake abuse incidents was also identified as a moment where some targets first

became aware of their own imagery. “And when you see things like that, it's very hard not to sympathise and the whole situation led me to be a bit curious and be like, “wait am I on these websites” so I went and Googled it and unfortunately I had also been deep faked by this person and you're about to learn how this is gonna fuck with my life forever.” The role of technology and search engines as first point of contact with trauma and abuse is important to note here as. Outside of the direct consumers and disseminators of deepfake imagery, the spreading of awareness of the imagery is also framed as contributing to the harm. “I'm. I'm so disgusted. I mean, I guess it's giving you clout. I guess. You got 2.6 million views, so It's a great way to go viral. I mean, I guess. I've never met you. I'm sorry if anything I've ever done has upset you, but you went out of your way to put hashtags about me being a pickme and leaks and stuff when it took two seconds to know that those are fake and really dehumanising to do to another woman.” She expands on this further: “It's really hurtful to see another young woman be so dehumanising because it took two seconds to know that those are fake instead of spreading that everywhere and me getting disgusting umm DM's about it, which I really don't want to see. I already had to take a couple days off of um social media this week because of the way that people um sometimes message me and it's just really disgusting.“ In this example, another influencer is positioned as contributing to deepfake abuse by spreading awareness about “leaked” images, leading to the victim being harassed and having to take time off from the internet. The streamers in the dataset particularly highlight how the public awareness feeds into patterns of online harassment and having their experiences disqualified You get accused of things like every day and you get called names every day and you just get to a point where you say,” OK, I don't. I don't. I can't care”. / I forget that there's no humanity on the Internet. This quotation also highlights how the targets position themselves as part of an uncaring internet, aimed on challenging their experiences. We will further discuss the characterisation of the internet and how it is used to voice shaming/silencing narratives around deepfake abuse.

Appendix VI: One week follow up results for study 4.

	Control group mean (SD)		Experimental group mean		Interaction Effect Time*Groups			Time effects			Group effects	
	1	2	1	2	Wilk's Lambda Time and groups (sig)	F* Time and Groups	η^2	Wilk's Lambda Time (sig)	F* Time	η^2	F* Groups (Sig)	η^2
Deep fake myths	13.204 (5.5233)	12.806 (5.554)	11.361 (5.459)	12.121 (5.933)	**.978 (.004)**	8.359	.022	.998 (.368)	.813 (.368)	.002	**5.41 (.021)**	.014
NSII Watch Past	2.42 (1.58)	2.25 (1.559)	1.9 (1.367)	1.9 (1.395)	.994 (.132)	2.278	.006	.993 (.106)	2.622 (.106)	.007	**9.153 (.003)**	.006
NSII Watch Future	2.33 (1.361)	2.20 (1.379)	1.90 (1.204)	1.86 1.269	.997 (.304)	1.060	.003	.991 (.069)	3.332 (.069)	.009	**9.266 (.002)**	.024
NSII Share Future	1.25 (.717)	1.25 (.747)	1.13 (.494)	1.19 (.660)	.997 (.277)	1.185	.277	.998 (.345)	.896 (.345)	.002	2.071 (.151)	.005
NSII Create Future	1.26 (.741)	1.31 (.878)	1.20 (.689)	1.23 (.752)	1 (.836)	.043	.000	.996 (.200)	1.649 (.200)	.004	.896 (.344)	.002
NSII Future Watch Intentions												
Friend	1.61 (1.176)	1.54 (1.132)	1.51 (1.008)	1.40 (.960)	1 (.707)	.142	.000	.990 (.050)	3.856 (.050)	.010	1.290 (.257)	.003
Ex	1.65 (1.206)	1.58 (1.130)	1.57 (1.124)	1.40 (1.0690	.996 (.213)	1.559	.004	.977 (.003)	8.979 (.003)	.023	1.435 (.232)	.004
Stranger	2 (1.303)	1.86 (1.265)	1.66 (1.164)	1.69 (1.170)	.991 (.071)	3.285	.009	.996 (.237)	1.401 (.237)	.004	**4.894 (.028)**	.013
Celebrity	2.40 (1.466)	2.20 (1.408)	1.91 (1.302)	1.86 (1.242)	.994 (.137)	2.221	.006	.982 (.010)	6.791 (.010)	.018	**9.849 (.025)**	.025
NSII Future Share Intentions												
Friend	1.11 (.455)	1.17 (.717)	1.08 (.403)	1.16 (.645)	1 (.835)	.044	.000	.987 (.027)	4.901 (.027)	.013	.044 (.835)	0.000
Ex	1.14 (.594)	1.16 (.68)	1.08 (.446)	1.15 (.619)	.998 (.428)	.631	.002	.993 (.1)	2.712 (.100)	.007	.355 (.552)	.001
Stranger	1.20 (.634)	1.20 (.723)	1.13 (.541)	1.16 (.642)	.999 (.634)	.043	.001	.999 (.634)	.227 (.634)	.001	.968 (.326)	.003
Celebrity	1.30 (.805)	1.26 (.824)	1.13 (.534)	1.19 (.641)	.995 (.163)	1.950	.005	1 (.739)	.111 (.739)	.000	3.261 (.072)	.009
NSII Future Create Intentions												
Friend	1.16 (.623)	1.25 (.818)	1.14 (.585)	1.16 (.632)	.997 (.253)	1.309	.003	.994 (.127)	2.343 (.127)	.006	.717 (.398)	.002
Ex	1.21 (.740)	1.21 (.794)	1.16 (.616)	1.16 (.632)	1 (.904)	.015	.904	1 (.904)	.015 (.904)	.000	.502 (.479)	.001
Stranger	1.20 (.677)	1.27 (.881)	1.18 (.681)	1.19 (.666)	.997 (.319)	.996	.003	.996 (.241)	1.38 (.241)	.004	.589 (.443)	.002
Celebrity	1.29 (.808)	1.30 (.888)	1.21 (.720)	1.19 (.693)	.999 (.590)	.291	.001	1 (.885)	.021 (.885)	.001	1.674 (.196)	.004