

Department of Psychological and Behavioural Science

**COURSEWORK SUBMISSION FORM AND
PLAGIARISM/ACADEMIC HONESTY DECLARATION**

Please ensure that a completed copy of this form is uploaded as part of your coursework submission.

Candidate Number: 25719

Course code: PB4D3 Behavioural Science in an Age of New Technology - Dissertation

Word-count: 9,990

Date: 20 August 2024

The Department wishes to draw your attention to the School Calendar Regulations on Assessment Offences and Plagiarism:

<https://info.lse.ac.uk/Staff/Divisions/Academic-Registrars-Division/Teaching-Quality-Assurance-and-Review-Office/Assets/Documents/Calendar/RegulationsAssessmentOffences-Plagiarism.pdf>

All work submitted as part of the requirements for any assessment of the School (e.g., examinations, essays, dissertations, and any other work, including computer programs), whether submitted for formative or summative assessment, must be expressed in your own words and incorporate your own ideas and judgments. Plagiarism must be avoided in all such work. Plagiarism can involve the presentation of another person's thoughts or words as if they were your own. However, please note that plagiarism also includes self-plagiarism, which is where you as the author re-use your own previously submitted work or data in a "new" written piece of work without letting the reader know that this material has appeared elsewhere. The definition of "your own work" also includes work produced by collaboration or group-work expressly permitted by the Department.

Please also note that plagiarism as defined by the School above, need not be deliberate or intentional for it to constitute an assessment offence.

Declaration (without signature, to preserve anonymity): Having read and understood LSE's guidelines on plagiarism/academic honesty, I hereby confirm by completing and attaching this form that the work submitted is my own. By submitting this form I hereby confirm I understand the Department's policy on summative assessment, and that I have read the relevant parts of the MSc programme handbook.

MSc Behavioural Science Dissertation 2023 / 2024

Deepfakes: Exploring the effectiveness and unintended
consequences of content warnings on human detection and
perception of visual misinformation

Supervised by Professor Liam Delaney at the Department of Psychological and Behavioural
Science at The London School of Economics and Political Science

20 August 2024

Abstract

‘Deepfakes’ are becoming increasingly realistic and prevalent online, leaving individuals at risk of deception, misinformation and fraud. It is crucial to understand how susceptible people are to the harms of deepfakes, and whether behavioural interventions can reduce susceptibility without unintended consequences. In an online randomised control experiment ($N = 163$, $M_{age} = 29.97$, $SD_{age} \approx 9.77$, females = 52.15%), we asked participants to categorise 16 videos as real or deepfakes. 8 of the videos were real and 8 were deepfakes, with content warnings added to 4 deepfakes in the treatment group. We measured susceptibility to deepfakes by analysing participants’ overall categorisation accuracy as well as overall perceptions of authenticity. We found (1) participants accurately categorised 60.46% of videos on average ($SD = 14.55$); (2) participants were biased towards categorising videos as real; (3) content warnings did not reduce susceptibility to deepfakes through either an increase in accuracy or decrease in overestimations of authentic content; (4) content warnings did not have unintended consequences by way of backfire or implied truth effects. Our findings suggest people are susceptible to visual misinformation through systematic errors in accuracy judgments and biased perceptions of authenticity. Content warnings were ineffective at reducing susceptibility, but they did not increase susceptibility either. With new technologies such as generative AI capable of producing manipulated media at scale, policymakers should commission further research as a priority.

Acknowledgements

First, I would like to thank Professor Liam Delaney who has been generous with his time during office hours this year and supervising me with this dissertation. He made the MSc in Behavioural Science a rewarding experience from both a theoretical and applied perspective and I will remember my time at LSE fondly.

There are many other academics at the LSE I'd like to thank, including my personal tutor Dr Laura Giurge, whose research on the science of time at work and beyond I found deeply fascinating. Thank you also to Dr Matteo Galizzi who taught us everything we need to know about running behavioural science experiments, and Dr Dario Krpan for introducing me to the fascinating intersection between behavioural science and technology. I would also like to thank Dr George Melios for providing us with the fundamental quantitative methods needed to conduct behavioural research, and Dr Christian Krekel for the rigorous seminars and teaching me to approach behavioural questions from a methodological perspective. Thank you also to Will Stubbs for answering all my questions this year and making the program so well organised.

I am also sincerely grateful to the staff at LSE Life for the study skills workshops they ran this year. Returning to education in my thirties was somewhat daunting at first, but they made the transition seamless. I would also like to thank Heather Dawson as the department's librarian for her guidance on literature this year, Ledia Pelivani for additional quantitative support at LSE Life, and Tarsha Vasu and Michael Wiemers from the LSE Digital Skills Lab for their support with statistical software.

Lastly, I'd like to thank my father and late mother for always encouraging curiosity and critical thinking, and always supporting me. I wouldn't be where I am today without them.

Table of Contents

ABSTRACT	3
ACKNOWLEDGEMENTS	4
1. INTRODUCTION	6
2. LITERATURE REVIEW	6
2.1. TRADITIONAL FORMS OF FALSE INFORMATION	6
2.2. THE EMERGENCE OF ‘DEEPPAKES’	8
2.3. DETECTION AND PERCEPTION OF DEEPPAKES	10
2.4. BEHAVIOURAL BIASES.....	10
2.5. REDUCING THE HARM OF DEEPPAKES.....	12
3. PRESENT STUDY	16
3.1. OVERVIEW	16
3.2. HYPOTHESES AND DEPENDENT VARIABLES	16
3.3. METHODOLOGY	18
3.4. RESULTS: TRUTH DISCERNMENT	24
3.5. RESULTS: OVERALL BELIEFS	31
4. DISCUSSION.....	34
4.1. ANSWERS TO RESEARCH QUESTIONS.....	34
4.2. ADDITIONAL FINDINGS	37
4.3. APPLIED IMPLICATIONS.....	38
4.4. LIMITATIONS.....	40
5. CONCLUSION.....	41
REFERENCES	42
APPENDIX A.....	53
APPENDIX B.....	55
APPENDIX C.....	56
APPENDIX D.....	57
APPENDIX E.....	59
APPENDIX F	65
APPENDIX G	68
APPENDIX H	70

1. Introduction

‘Deepfakes’ have emerged as a powerful video, image and audio manipulation technology. While they could benefit society, there is growing concern about their potential for harm. Research suggests humans struggle to spot manipulated visual content and may be biased towards perceiving such content as real. As such, people may be susceptible to harmful deepfake material. In this study, we ran an online experiment to explore people’s susceptibility to deepfake content, and whether content warnings reduce susceptibility without having any unintended consequences. In the sections that follow, we first explore the literature on false information and deepfakes. We then introduce our study and present our findings. We conclude with a discussion of the implications of our study and avenues for further research.

2. Literature review

2.1. Traditional forms of false information

Humans have created, shared and encountered false information throughout history, with ancient civilisations using rock carvings and papyrus to create propaganda (Burkhardt, 2017). During the Roman Empire, Gaius Octavian and Mark Antony used speeches and engravings on physical currency to spread false information about one another (van der Linden, 2023). As literacy rates increased and advances in technology paved the way for a printed press, gossip and speculation became easy to create and share through textual modalities (Berkhardt, 2017; Darnton, 2017). More recently, the internet and social media have increased the scale and dissemination of false information (Vizoso, 2021).

According to Wardle and Derakhshan (2017), there are three types of false information. Misinformation is false information shared with no malicious intent. Disinformation is false information shared deliberately to cause harm. Mal-information is

authentic information shared maliciously, for example by putting non-public information into the public realm. Although not new concepts, concerns have grown in recent years about the spread of misinformation and disinformation online (Pennycook & Rand, 2021a). Examples include false information about the coronavirus vaccination (Kouzy et al., 2020), misleading claims about climate change (Treen et al., 2020) and fake news articles about elections (Allcott & Gentzkow, 2017).

There are differing accounts for why people believe false information. For example, motivated reasoning accounts suggest that people believe political misinformation because it aligns with their political beliefs (Kahan et al., 2017; Kunda, 1990). Analytical thinking has been found to influence belief in false information (Kelley et al., 2023; Pennycook et al., 2012; Swami et al., 2014). Depleted cognition may also be a factor (Gilbert et al., 1993), as well as lack of deliberation (Bago et al., 2020), distraction (Pennycook et al., 2021b) and heightened emotions (Martel et al., 2020).

Belief in false information is conceptualised in two distinct ways. Pennycook and Rand (2021a) distinguish between *truth discernment* and *overall beliefs*. Truth discernment measures belief in misinformation from an overall accuracy perspective (i.e. the extent to which an individual correctly discerns between true and false information). Overall beliefs measure an individual's overall perceptions of authenticity (i.e. the extent to which individuals perceive all information to be true, regardless of accuracy). Both concepts measure susceptibility to false information, but truth discernment captures accuracy of judgments, whereas overall beliefs capture bias in judgments.

While research on textual misinformation is helpful to understand the underlying mechanisms, other modalities of false information and fake content are rapidly emerging, and research is needed on whether existing mechanisms explain susceptibility to such content.

2.2. The emergence of ‘deepfakes’

Deepfakes are “*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*” (Regulation (EU) 2024/1689, 2024, Article 3(60)). They swap the face or voice of one individual onto the face or voice of another, giving a highly realistic depiction of another person (Iacobucci et al., 2021; Mirsky & Lee, 2021). The term ‘deepfake’ is a portmanteau of the words *deep learning* and *fake*, with deep learning being the branch of artificial intelligence that is used to create synthetic media (Westerlund, 2019). The term is believed to have emerged in 2017, with subsequent high-profile deepfakes catching the attention of the mainstream media (Mirsky & Lee, 2021; Ternovski et al., 2022).

On a technical level, the most realistic deepfakes are the product of powerful machine learning networks, including Generative Adversarial Networks (GANs) (Westerlund, 2019). Mirsky and Lee (2021) explain how GANs comprise a generative network that creates synthetic output based on authentic input, and a discriminative network that is trained to differentiate between the two. The networks compete against each other and eventually, the discriminative network can no longer distinguish between the authentic input and fake output. The result is a synthetic video, image or audio that is highly realistic.

Some commentators suggest deepfakes have commercial, educational and social value. Examples in the literature include virtual store experiences, foreign language video dubbing and the recreation of historic events for schoolchildren (Mirsky & Lee, 2021; Nas & Kleijn, 2024). Other use cases include multi-lingual marketing and virtual brand ambassadors (Mustak et al., 2023), grief management (Yang, 2024), and voice reconstruction (“Lawmaker uses AI voice clone”, 2024). While deepfakes may have value, many commentators cite the dangers of deepfakes.

Deepfakes distort perceptions of reality, which erodes trust in individuals, institutions and society (Pinhanez et al., 2022). They can destabilise democratic functions through political impersonation and microtargeting (Chesney & Citron, 2018). A deepfake of Sir Keir Starmer berating staff was disseminated widely on social media, its release timed to coincide with the Labour Party conference in 2023 (Bristow, 2023). Images of US Presidential candidate Donald Trump surrounded by Black voters during the 2024 election campaigns were quickly identified as deepfakes (Spring, 2024). This threatens democracy because voters could make voting decisions based on fake political messages (Dobber et al., 2021). Online environments are vulnerable to polarisation and inequality of political knowledge, making political deepfakes a high risk (van Aelst et al., 2017).

Deepfakes also threaten privacy, reputation and security. Machine learning models are used to create deepfake pornography (Cook, 2019; Hao, 2020; Westerlund, 2019) and to facilitate crime (Damiani, 2019, as cited in Köbis et al., 2021). Cyberfraud has been rated as the most concerning type of digitally enabled crime (Caldwell et al., 2020). In the health domain, cyberhackers could use deepfake technology to manipulate health scan imaging for blackmail and insurance fraud (Mirsky et al., 2021). In the security domain, deepfake fingerprints could unlock personal devices (Bontrager et al., 2018).

It is becoming easier for members of the public to access deepfake technology. In the past, specialist knowledge, computing power and access to large training datasets were required to create deepfakes (Fletcher, 2018). Advances in computer graphics, availability of open-source software, and the release of apps such as FakeApp have made it easy to create deepfakes at scale (Fletcher, 2018; Vizoso, 2021). Deepfakes therefore pose risks in terms of scale of production, which is problematic if people struggle to detect them.

2.3. Detection and perception of deepfakes

When exploring people's susceptibility to deepfakes, research often uses the truth discernment measure explored in Section 2.1. Studies compute accuracy scores based on participants' video categorisation decisions. Human performance is mixed, with accuracy rates of 57.6% (Köbis et al., 2021), 60% (Bray et al., 2023) 60.70% (Somoray & Miller, 2023), 62% (Nightingale et al., 2017), 80% (Nas & Kleijn, 2024) and 88.9% (Groh et al., 2022). Human performance is comparable to, and in some cases better than, leading machine learning models (Dolhansky et al., 2020; Groh et al., 2022). Although computers can detect deepfake artifacts that are not visible to the naked eye, they struggle to detect deepfakes that humans would easily identify (Korshunov & Marcel, 2020). Given the variation in human performance, we formulate our first research question:

RQ1: How do humans perform at accurately categorising videos as real or deepfakes, and what does this imply about susceptibility to deepfake content?

Accuracy rates do not provide the full picture on susceptibility to deepfake material. Given the risks of deepfakes distorting perceptions of reality, it is also important to measure overall perceptions of videos to capture bias. The lack of coverage of overall beliefs in the deepfake literature is a key limitation of current research.

2.4. Behavioural biases

Underlying behavioural biases may influence whether individuals become susceptible to deepfake content. Tversky and Kahneman (1974) conceptualised biases as the product of *heuristics*, which they define as mental shortcuts that simplify complex decision making under uncertainty. The authors acknowledge that heuristics can be useful, but they can lead to systematic errors, including in online contexts (Pennycook & Rand ,2021a; Pennycook et al., 2021b).

There is disagreement about the direction of bias in perceptions of visual content. One view is that as individuals become more aware of manipulated media content, they become more sceptical of all media and may even discredit authentic content (Chesney & Citron, 2018). This implies high sensitivity to manipulated media, leading to an overestimation of its prevalence and bias towards categorising content as fake (Köbis et al., 2021). This problem has been conceptualised as a *liar's dividend*, since an individual could exonerate themselves from a malicious act by claiming authentic evidence of their wrongdoing was in fact a deepfake (Chesney & Citron, 2018). While bias towards perceiving videos as fake may make individuals less susceptible to falling for deepfakes, this could result in them discrediting authentic content (Ternovski et al., 2022) through overestimation of deepfakes.

A view that has more support in the literature is that individuals generally assume content to be authentic unless there is clear evidence of manipulation, under a *seeing is believing* heuristic (Farid, 2019, as cited in Köbis et al., 2021). Evidence of this heuristic is found by Frenda et al. (2013) who found individuals claimed to remember fictitious political events when shown an image depicting the event. The seeing is believing heuristic implies individuals are less sensitive to manipulated media and underestimate its prevalence, resulting in a bias towards categorising content as real (Köbis et al., 2021; Somoray & Miller, 2023). While this heuristic may reduce the likelihood of individuals incorrectly categorising real content as fake, they will be susceptible to falling for deepfakes through underestimation of their prevalence.

In addition, individuals are susceptible to deepfakes because the *realism heuristic* makes multimodal content such as videos appear more credible and trustworthy (Sundar, 2008). Multimodal content is richer than textual information and has a higher degree of fidelity to the real world (Sundar, 2008; Weikmann, 2024). Textual information is abstract and requires cognitive effort to impute meaning and to imagine events (Messaris & Abraham,

2001, as cited in Hameleers, 2020). Research supports this hypothesis, finding deepfakes to be more vivid, credible and persuasive (Hwang et al., 2021). Together, this implies visual information engages the faster, automatic and intuitive *System 1* thinking, whereas textual information engages the slower, deliberative and reflective *System 2* thinking (Kahneman, 2011; Pinhanez et al., 2022), leaving people at risk of believing a deepfake to be authentic. Additionally, the *fluency heuristic* posits that individuals will process information more fluently where the information appears familiar (Shin, 2022). Research from psychology suggests fluency influences accuracy judgments, where high levels of fluency (e.g. from repeated exposure to false claims) are positively associated with truth judgments (Berinsky, 2017; Newman et al., 2015).

The heuristics explored above can be conceptualised as resulting in a broader *confirmation bias* (Wason, 1960). Under confirmation bias, individuals seek out and interpret evidence in a way that confirms existing beliefs (Nickerson, 1998). This leads to lower quality decisions since people choose to rely on evidence that aligns with their beliefs, disregarding evidence to the contrary (Hernandez & Preston, 2013). Of the few human deepfake detection studies that consider bias, Köbis et al., (2021) suggests a confirmation bias towards categorising videos as real, whereas Somoray and Miller (2023) did not find evidence of this. We therefore formulate our second research question:

RQ2. Do individuals show an overall bias towards categorising videos as real or deepfakes, and if so, what does this imply about susceptibility to deepfake content?

2.5. Reducing the harm of deepfakes

Underlying behavioural mechanisms appear to influence people's susceptibility to false information, including visual disinformation such as deepfakes. An exploration of potential interventions to help people overcome such biases is needed.

2.5.1. Technical interventions

Technical interventions use machine learning models to automatically flag when a video is a deepfake. Mirsky and Lee (2021) summarise how machine learning tools can detect pre-defined indicators of manipulation, such as image boundaries and background and foreground contrasts. They can also detect anomalies in the artificially reconstructed footage, comparing these against authentic training footage. Although technical measures are valuable because they detect subtle artifacts that are not visible to humans (Mirsky & Lee, 2021), they are not sufficient to reduce susceptibility to deepfakes because they do not help individuals overcome underlying biases (Chesney & Citron, 2018).

2.5.2. Legal interventions

Governments and institutions are taking gradual interest in regulating deepfakes. In the European Union, the EU AI Act will impose transparency obligations on businesses that create deepfakes, with requirements to clearly label deepfake content (Regulation (EU) 2024/1689; Romero Moreno, 2024). The UK is less interventionist, adopting a principles-based approach (Office for Artificial Intelligence, 2023), although it is now a criminal offence to share deepfake revenge pornography (Online Safety Act, 2023). In the USA, there are no federal laws regulating deepfakes, although some States regulate deepfakes in election campaigns (Graham, 2024). In China, recent laws subject app developers and platform providers to compliance requirements (Herbert Smith Freehills, 2023). While legislative interventions are welcome, it is unrealistic to expect broader legal developments to reduce susceptibility to deepfakes. As Chesney and Citron (2018) explain, blanket bans on deepfakes would fetter innovation and enforcement of rights would be complex. Legal interventions also do not help individuals overcome underlying behavioural biases that make them susceptible to false information.

2.5.3. Behavioural interventions

As an alternative to technical and legal interventions, recent research explores the effectiveness of behavioural interventions to reduce susceptibility to deepfake content, but the results are mixed. Some studies have explored the use of information, incentives and education to reduce susceptibility to deepfakes. Köbis et al. (2021) found that giving people a description of deepfakes or financial incentives had no effect on deepfake truth discernment. Somoray & Miller (2023) found that giving people training on common deepfake artifacts was also ineffective, and Bray et al. (2023) found giving people visual examples of deepfakes to be similarly ineffective. One study found that priming people with the definition of a deepfake improved detection (Iacobucci et al., 2021), whereas another study found a digital literacy intervention to be effective at reducing susceptibility to deepfakes (Hwang et al., 2021). Consequently, a mixed picture emerges on using awareness and training to reduce susceptibility to deepfakes.

Content warnings may be a more promising intervention, but their effectiveness at reducing susceptibility to non-textual misinformation is unclear. Content warnings are labels attached to online content that alert a user to potentially false or misleading content (Martel & Rand, 2023). Conceptually, they can be described as a *nudge* (Thaler & Sunstein, 2008), since they change the choice architecture without removing people's options, guiding people to make better decisions. Content warnings achieve this by adding friction to decision-making, forcing more reflective and deliberative thinking (Cox et al., 2016 as cited in Guo et al., 2024). Content warnings have been found to be effective at reducing belief in fake news (Pennycook et al., 2018; Pennycook et al., 2020), but there are mixed findings in the context of deepfakes. Lewis et al., (2023) found content warnings ineffective, whereas Ahmed (2021) found content warnings reduced the likelihood of participants perceiving a deceptive message contained within a deepfake as real. Research emphasises that the type and design of a

warning can influence effectiveness (Martel & Rand, 2023). Guo et al., (2024) found contextual warnings added at the individual video level were more effective at reducing belief in Covid-19 misinformation than general warnings inviting individuals to learn about the coronavirus vaccination. Contextual warnings that mention fact-checkers have been found to be effective for textual misinformation (Clayton et al., 2020), although there are suggestions from the research that more categorical statements may be needed for visual information (Guo et al., 2024). We therefore formulate our third research question as follows:

RQ3. If individuals are susceptible to deepfake content, do content warnings reduce susceptibility?

As with any behavioural intervention, unintended consequences could undo the effect of an intervention. An unintended consequence of adding content warnings to false information is that they might backfire by increasing, rather than decreasing, belief in a false claim (Nyhan & Reifler, 2010). A suggested mechanism for backfire effects is that individuals show reactance to information that conflicts with pre-existing beliefs, leading to even stronger beliefs (Lodge & Taber, 2000). This implies that content warnings might increase, rather than decrease, belief in false information (Pennycook et al., 2020), for example where individuals strongly believe that video content is generally authentic. We are aware of one deepfake study where content warnings increased disbelief in authentic videos, although the research context used political videos (Ternovski et al., 2022).

Exposure to content warnings could also lead to an “*implied truth effect*”, where people imply that untagged information is true simply because it does not have a warning (Pennycook et al., 2020, p.4945). The literature suggests that for content warnings to be effective, they must be applied as widely as possible (Martel & Rand, 2023). Practically, it is not possible to label all misleading content with warnings due to the speed and scale at which misinformation can be created and shared, considering the limited resources of fact-checkers

(Pennycook et al., 2020). People could, however, infer that unlabelled fake content is real because they expect false information to have been tagged already (Pennycook et al., 2020).

Consequently, we formulate our fourth research question:

RQ4: Could content warnings have any unintended consequences through backfire effects or implied truth effects?

3. Present Study

3.1. Overview

The goal of the present study was to measure susceptibility to deepfake content, and whether content warnings reduce susceptibility without having unintended consequences. We recruited participants for an online experiment during which they watched sixteen videos of celebrities, 8 of which were real videos and 8 were deepfakes. Participants were randomly allocated into one of two experimental conditions. In the treatment condition, warnings were added to 4 deepfakes, with the remaining deepfake and real videos having no warnings. This was intentional to test for an implied truth effect per Pennycook et al. (2020). In the control condition, participants watched the same videos but without warnings.

3.2. Hypotheses and dependent variables

Our overarching outcome of interest was susceptibility to deepfakes, which we measured from a truth discernment perspective and an overall beliefs perspective (Pennycook & Rand, 2021a).

3.2.1. Truth discernment

The truth discernment measure of susceptibility captured how accurate participants were at categorising videos as real or deepfakes. Based on our literature review, we formulated the following hypotheses:

***H1:** Individuals will perform above chance levels of accuracy at categorising all videos.*

***H2:** Adding content warnings to deepfake videos will increase accuracy judgments (i.e. warnings will improve truth discernment).*

The main dependent variables for truth discernment were a participant's video categorisation scores. The total score was a measure of the participant's score when categorising all videos as real or deepfakes with a maximum score of 16. We also computed a deepfake score as a more focussed score for exploratory analysis, measuring how participants fared at categorising specifically the 8 deepfakes videos, with a maximum score of 8.

3.2.2. Overall beliefs

The overall beliefs measure of susceptibility measured the overall number of videos that participants categorised as real, regardless of accuracy. We formulated the following hypotheses when measuring overall beliefs:

***H3:** Individuals will be biased towards categorising videos as real than deepfakes.*

***H4:** Adding content warnings to deepfake videos will reduce the overall number of videos categorised as real (i.e. warnings will correct overestimations of authentic content).*

***H5:** Adding content warnings to deepfake videos will decrease the number of real videos categorised as real (i.e. there will be a backfire effect where participants discredit the authenticity of real videos).*

***H6:** Adding content warnings to deepfake videos will increase the number of untagged videos categorised as real (i.e. there will be an implied truth effect).*

When analysing overall beliefs, we sought to identify a warning effect and an implied truth effect on overall perceptions about the authenticity of the videos, per Pennycook et al. (2020). We used the number of 'real' responses participants provided when categorising

videos as real or fake as the main outcome variable of interest, as opposed to accuracy scores. This allowed us to isolate the presence and absence of warnings on perceptions of the videos. It also allowed us to assess whether warnings lead people to discredit real videos.

3.3. Methodology

3.3.1. Ethics approval

This study was carried out with the approval of the LSE Research Ethics Committee and in accordance with the Research Ethics Policy and Procedures (LSE Research Ethics Committee, 2023) and Code of Research Conduct (LSE Research Ethics Committee, 2024).

3.3.2. Sample size and recruitment

A priori power analysis in G*Power (Faul et al., 2007) confirmed a sample size of at least 156 was needed to detect small to medium effects ($\alpha = 0.05$, $d = 0.40$) with a power of 0.80 and using two experimental conditions and one-sided t-tests. Participants had to be at least eighteen years of age and living in Australia, Canada, the EU, New Zealand, the UK or USA. The decision to recruit participants from these regions was to generate multi-region insights while limiting major cultural variations. We recruited participants through both convenience sampling and through Prolific. We paid Prolific participants a fair participant fee and controlled for recruitment source in our analysis.

3.3.3. Data cleaning and final sample

Data collection took place from 12 to 24 June 2024. Participants accessed and completed an online survey hosted on Qualtrics. Survey responses were anonymised, and no personal data was collected. Of the responses, 1 participant was removed for not providing consent. A further 20 participants did not complete the survey. A further 6 participants were removed for failing attention checks. A further 21 participants were removed because they

had professional expertise relating to our study or had participated in previous deepfake research.

The final sample size was $n=163$, with 34 participants recruited via convenience sampling (20.86%) and 129 participants recruited via Prolific (79.14%). The sample population skewed towards younger participants, with 75% of the sample aged 34 years or younger ($M_{age} = 29.97, SD_{age} \approx 9.77$). 52.15% of the sample identified as female, 42.94% as male, 3.07% as non-binary, 1.23% as transgender and 0.61% as other. The sample comprised mostly white participants (82.21%). At the time of the study, most participants were living either in the EU (38.65%) or the UK (30.06%). The sample featured a mixture of professional backgrounds, with 46.01% in full or part-time employment and 30.67% students. Most participants were aware of the concept of a deepfake (93.87%), and although most participants reported prior exposure to a deepfake (58.28%), a significant minority were unsure of this (33.74%). A summary of participant demographics groups is shown in Appendix A, Table A1.

3.3.4. Experimental design and conditions

The study used a simple between-subjects design with two experimental conditions. All participants were required to watch sixteen video clips, 8 of which were real, and 8 were deepfakes. In the treatment condition, a content warning was randomly added to 4 deepfake videos, warning participants that the authenticity of the video had been discredited by independent fact-checkers with a visual warning cue. Participants in the treatment condition were asked to answer additional questions about their interaction with the warnings. In the control condition, participants watched the same videos as the treatment condition, but there were no warnings. An overview of the experimental design is provided in Appendix B, Figure B1.

3.3.5. Experimental stimulus

Stimulus videos were sourced from the Celeb-DF dataset (Li et al., 2020). This is a large data set of real and deepfake videos of 59 different celebrities. The dataset was created to train machine learning models, although a recent study used this dataset for psychology research (Nas & Kleijn, 2024). The dataset is newer than other deepfake datasets such as the Deepfake Detection Challenge Dataset (Dolhansky et al., 2020) and FaceForensics++ (Rossler et al., 2019), and the deepfakes are harder to detect because they have fewer deepfake artifacts such as mismatched colour and poor resolution. The dataset features subjects from a variety of ethnic groups, genders and ages. For each celebrity, the dataset includes real videos of the celebrity in different interview settings. It then includes deepfakes of that celebrity where their face is swapped with the face of another celebrity from the dataset. In total, the dataset contained 590 real videos and 5,639 deepfakes. Appendix C, Figure C1 shows an example of a real video and a deepfake counterpart from the dataset.

There is no audio in any of the videos in the Celeb-DF dataset. This appeared common in the datasets we reviewed where the purpose is to train machine learning models to detect visual deepfake artifacts. Although there are some deepfake video datasets that include audio (Khalid et al., 2022), we found the visual content to be poor.

When deciding which videos to use, we randomly selected 16 of the 59 celebrities, checking they were balanced across genders. We randomly assigned the 16 selected celebrities to a ‘real’ or ‘deepfake’ condition, randomly selecting one of the celebrity’s real videos for each celebrity in the ‘real’ condition, as well as one of the deepfake videos for the celebrities in the ‘deepfake’ condition. We checked each selected video against pre-defined inclusion criteria per the approach of Somoray and Miller (2023) (e.g. consistent lighting, steady footage and single subject with no glasses).

The final set of stimulus videos comprised 16 videos of 16 different celebrities, with 8 of the videos being real and 8 of the videos being deepfakes. Seven of the 16 videos featured female subjects and three of the videos featured non-white subjects. The average length of each video was 12 seconds. Videos were embedded within the Qualtrics survey platform. We tried to mitigate bias from using well-known celebrities in our instructions to participants (see Section 3.4.8 below). Appendix C, Table C1 provides a summary of the stimulus videos.

3.3.6. Information sheet and informed consent

All participants read an information sheet that provided details of the study, the details of the researcher, inclusion criteria, how responses would be used, confidentiality and anonymity and details of the LSE Research Privacy Notice and LSE ethical guidelines. Participants were asked to provide informed consent. Participants who did not provide informed consent were directed to the survey exit page.

3.3.7. Experimental conditions and randomisation

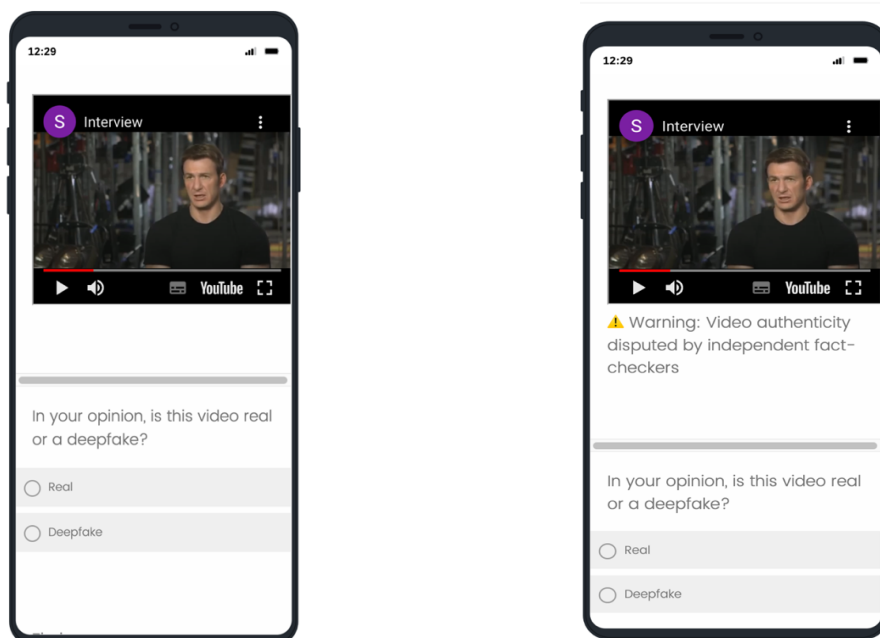
Participants were randomly allocated into the control or treatment conditions using Qualtrics' built-in randomisation feature. Successful randomisation should result in bias from differences in individual characteristics being equally balanced across control and treatment groups, allowing for an unbiased measure of the treatment effect (List et al., 2011). Appendix A, Table A2 reports the results of randomisation checks for balance. The differences in means between control and treatment groups was small across observed characteristics and were not significant. This confirmed successful randomisation and balance across treatment and control groups. We therefore do not control for observed characteristics in our main analysis (Vaccari & Chadwick, 2020).

3.3.8. Video categorisation task

All participants read an introduction to the video categorisation task. They were told that half of the videos would be real, and half would be deepfakes. Giving this information was deemed necessary for ethical reasons and to facilitate insights on whether participants categorised videos in these proportions. All participants were given a definition of a deepfake. Participants were instructed to make their categorisation decisions based on their overall impression of a video and not simply whether they recognised the video subject. Participants could replay each video multiple times. As participants navigated through each video, participants answered the question “*In your opinion, is this video real or a deepfake?*” with binary “Real” or “Deepfake” responses. Four videos contained attention checks. Figure 1 below shows an example of the video categorisation task.

Figure 1

Example of Video Categorisation Task



Note: Example of the video categorisation task within the Qualtrics survey platform as seen by participants in the control condition (left) and the treatment condition (right).

3.3.9. Treatment group insights

Participants in the treatment group answered questions about their awareness of the warnings, how the presence of a warning influenced their decisions and to what extent they trusted the warnings. Participants were also asked how the absence of a warning influenced their perception of a video's authenticity.

3.3.10. Demographic questions

Participants in both conditions answered standard demographic questions. We also asked participants to confirm whether they have worked in occupations relating to AI, investigative journalism or fact-checking (Köbis et al., 2021), as well as whether they had participated in academic research on deepfakes. We used responses to these questions as additional exclusion criteria to reduce bias. We also asked participants about their past awareness and exposure to deepfakes and gathered free-text insights on sentiment towards deepfakes.

3.3.11. Debriefing

All participants were debriefed within Qualtrics. The debriefing gave full details of the study purpose and design and a reminder that responses were anonymous and that we did not collect personal data. Participants were given the contact details of the primary researcher and encouraged to provide study feedback. Participants were reminded of their right to withdraw from the study and a unique participation ID to do so. We consulted Greene et al. (2023) on best practices for debriefing participants on misinformation experiments. Their guidance recommends explaining which information used in the study was incorrect and why it is incorrect. The guidance also recommends sharing resources from credible sources. We therefore presented participants with a score for the video categorisation exercise and personalised feedback on each video categorisation decision based on deepfake detection

strategies developed by MIT Media Lab (n.d.). We also shared educational resources on deepfakes and artificial intelligence. A sample of the debriefing is shown in Appendix G.

3.4. Results: Truth Discernment

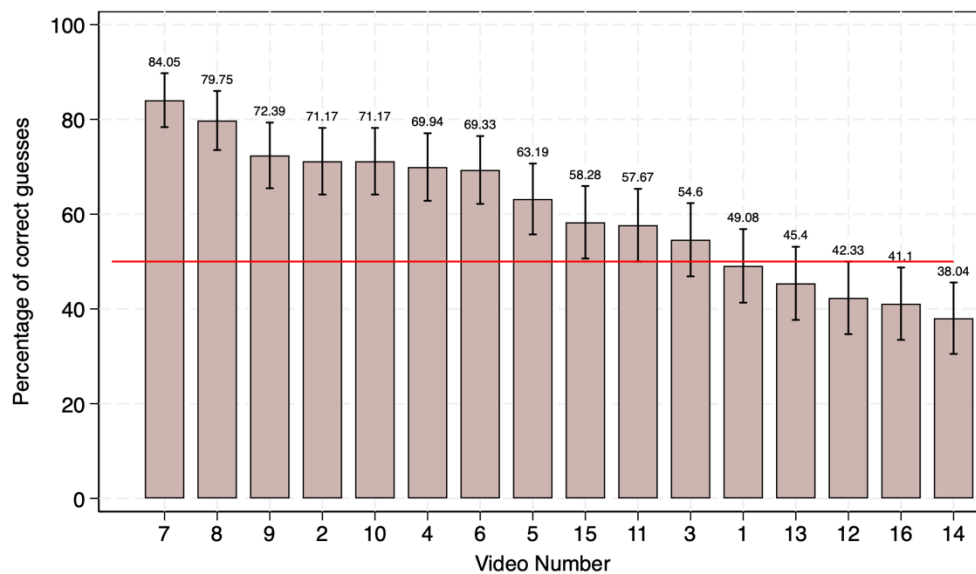
3.4.1. Overall accuracy

The mean total accuracy score for all participants was 9.67 out of 16 ($SD = 2.33$), or 60.46% ($SD = 14.55$) accuracy. Based on chance, we would expect individuals to correctly categorise 8 out of 16 videos (50% accuracy). A one-sample t-test confirmed that on average, participants' total scores were significantly greater than chance ($t = 9.18, p < .001$). The lowest score was 4 (25% accuracy) and the highest score was 15 (93.75% accuracy). The mean deepfake score for all participants was 4.2 out of 8 ($SD = 1.63$), or 53.3% ($SD = 20.24$) accuracy, which was also significantly above chance levels ($t = 2.06, p < 0.05$). The lowest score was 0 and the highest score was 8. Hypothesis 1 was therefore supported.

Figure 2 below shows the percentage of accurate scores per video for all participants independent of treatment condition, arranged in descending order and with a reference line indicating chance levels of accuracy. Participants performed significantly better than chance for most videos.

Figure 2

Overall Accuracy by Video Number



Note: Percentage of correct guesses for each video for all participations independent of treatment condition arranged in descending order per video number. Error bars denote 95% confidence intervals and the reference line in red indicates the probability of guessing correctly based on chance. Plot adapted from Figure 2 of Köbis et al., 2021 and created in Stata.

Appendix D, Figure D1 disaggregates performance according to video type independent of treatment condition. Visually, it shows how participants overall performed better at correctly categorising real videos than deepfakes and how accuracy judgments exceeded chance levels in 7 out of the 8 real videos, but only 4 out of the 8 deepfake videos. Together, this suggests uncertainty towards the deepfakes.

3.4.2. Effect of content warnings on accuracy

Table 1 displays the results of a one-tailed independent sample t-tests that were conducted to test the hypothesis that the mean total and deepfake accuracy scores for the treatment group are higher than those for the control group.

Table 1

Independent Samples t-tests Comparing Video Categorisation Scores Between Treatment and Control Groups

<i>Video Categorisation Score</i>	Control		Treatment		<i>t</i> (161)	<i>p</i>	Cohen's <i>d</i>
	<i>M</i>	<i>SD</i>	<i>M</i>	<i>SD</i>			
Total score	9.46	2.26	9.90	2.39	-1.19	0.883	0.18
Deepfake score	4.18	1.52	4.35	1.75	-0.68	0.753	0.10

At the participant level, mean accuracy for the treatment group was not significantly higher than the control group ($p > 0.05$) for either score. The t-tests therefore reveal no significant improvement in accuracy judgments due to the treatment. Appendix E, Tables E1 and E2 show results of linear regressions carried out at the participant level which confirm the treatment did not have a statistically significant effect on improving accuracy. The lack of effect at the aggregate participant level is likely due to the partial imposition of warnings. Consequently, we also analysed the effect of warnings at the individual video level using 'warned' 'untagged' and 'deepfake' dummy variables, similar to Pennycook et al. (2020).

At the video level, logistic regression analysis confirmed there was no significant effect of warnings on accuracy (Appendix E, Tables E3 to E5). We did, however, observe significant negative coefficients for the 'deepfake' variable, suggesting that accuracy scores were worse for deepfakes. Together, the t-tests and regression analysis provided insufficient evidence to support Hypothesis 2.

3.4.3. Accuracy by treatment condition

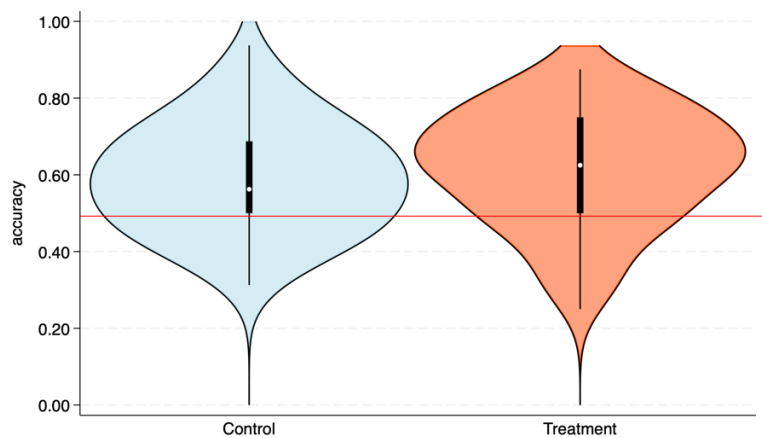
Although we did not observe significant warning effects on accuracy judgments, we were interested in accuracy patterns by treatment condition. The t-tests in Table 1 show that treatment group participants correctly categorised all sixteen videos in the set as real or

deepfakes 61.88% of the time whereas the control group correctly categorised the videos 59.19% of the time. For the eight deepfake videos, treatment group participants correctly categorised the deepfake videos in the set as deepfakes 54.38% of the time and the control group 52.25% of the time. Appendix D, Figure D2 compares treatment and control group performance for each video, showing an overall inconsistent picture on accuracy judgments. When disaggregating their accuracy judgments by video type, the treatment group outperformed the control group when categorising real videos, but neither group outperformed the other when categorising deepfakes, despite the treatment group being exposed to warnings (Appendix D, Figure D3).

Figure 3 presents violin plots showing the density distribution of the total scores for control and treatment conditions. Although the central tendency of the total score is similar in both experimental conditions, the distribution appears slightly more peaked in the treatment group around the median, suggesting less variability and consistent performance than the control group. Figure 4 presents further violin plots showing the density distribution of the deepfake specific scores for control and treatment conditions. A multi-modal distribution is apparent within the treatment group, which is likely due to the partial imposition of warnings. This may suggest warnings helped treatment group participants correctly identify deepfakes when tagged with warnings, but resulted in uncertainty and variability in accuracy judgments when deepfakes were not tagged with warnings.

Figure 3

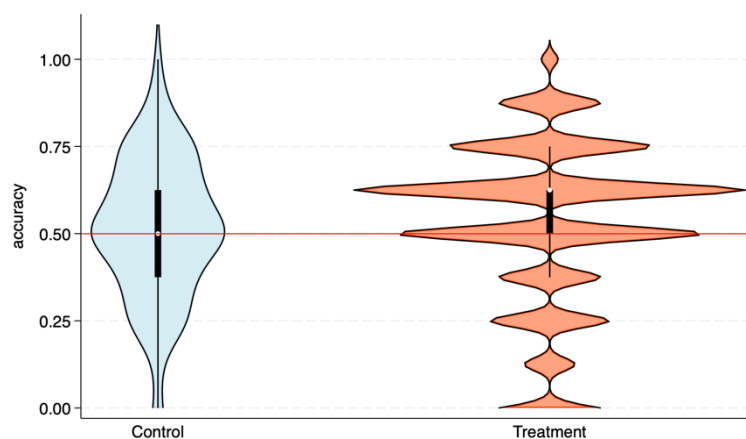
Distribution of total accuracy scores by condition



Note: Violin plots comparing the density distribution of a participant's total score (i.e. categorisation of all 16 videos as real or deepfakes) between control and treatment conditions. The y-axis is rescaled from 0 to 16 (the maximum possible score) to a range of 0 to 1.00 to present accuracy scores as a proportion of the maximum possible score. The interquartile ranges are represented by the black rectangles and the medians are indicated by the white circles within the black rectangles. The black whiskers extend to the smallest and largest values within 1.5 times of the interquartile range. The red reference line at 0.50 indicates chance levels of accuracy. Plot adapted from Figure 3 Somoray and Miller (2023) and created using violinplot in Stata.

Figure 4

Distribution of deepfake accuracy scores by condition



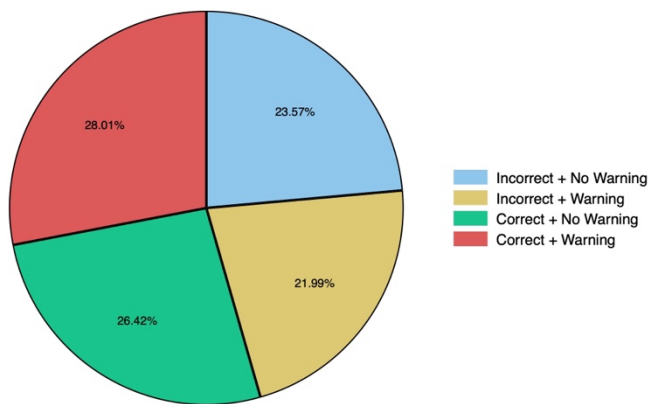
Note: Violin plots comparing the density distribution of a participant's deepfake score (i.e. categorisation of the 8 deepfake videos in the video set as real or deepfakes) between control and treatment conditions. The y-axis is rescaled from 0 to 8 (the maximum possible score) to a range of 0 to 1.00 to present accuracy scores as a proportion of the maximum possible score. The interquartile ranges are represented by the black rectangles and the medians are indicated by the white circles within the black rectangles. The black whiskers extend to the smallest and largest values within 1.5 times of the interquartile range. The red reference line at 0.50 indicates chance levels of accuracy. Plot adapted from Figure 3 Somoray and Miller (2023) and created using violinplot in Stata.

3.4.4. Accuracy by video warning status

Figure 5 shows the proportion of correct and incorrect guesses for the deepfake videos with and without warnings within the treatment group. It shows how most guesses were correct (54.47%), with the largest segment corresponding to correct guesses in the presence of a content warning (28.01%). However, treatment group participants still categorised a sizeable proportion of deepfakes videos as real, even when a warning was added to that video (21.99%), suggesting a failure to take the warning into account when making categorisation decisions.

Figure 5

Proportions of Correct and Incorrect Guesses for Deepfake Videos with and without Warnings in the Treatment Group



Note: Pie chart illustrating the aggregate proportion of correct and incorrect guesses for deepfake videos with and without warnings in the treatment group. Plot created in Stata.

3.4.5. Accuracy and heterogeneity

As exploratory analysis, a logistic regression was carried out at the video level to predict the likelihood of accurately categorising a video based on participant and video subject characteristics. As seen in Table 2, video categorisation accuracy was significantly higher when the video subject was male ($\beta = 0.517$, $p < 0.001$) compared to when the subject

was female, irrespective of the participant's own gender. Female participants were overall significantly more accurate ($\beta = 0.312$, $p < 0.05$) compared to non-female participants, irrespective of the video subject's gender. However, the negative and statistically significant interaction term between participant gender and the video subject gender shows a decrease in female participants' accuracy when the video subject is male compared to female ($\beta = -0.303$, $p < 0.05$). The video subject's ethnicity did not influence accuracy.

Table 2

Results of Logistic Regression Predicting Video Categorisation Accuracy Using Individual Video Scores as the Dependent Variable

Variable	Video Score Estimate	SE	p-value
MaleSubject	0.517***	0.104	0.000
Female	0.312**	0.111	0.005
MaleSubject#Female	-0.303*	0.142	0.032
WhiteSubject	-0.031	0.102	0.759
Constant	0.090	0.131	0.491

Note. Table E9 shows the logistic regression coefficients for predicting the likelihood of accurately categorising a video as real, where the dependent variable 'Video Score' is binary (1 = correct; 0 = incorrect). The analysis was conducted at the individual video level which means each video seen by each participant is treated as a separate observation in the model. 'MaleSubject' and 'WhiteSubject' are dummy variables to indicate the gender and ethnicity of the video subject respectively. The interaction term 'MaleSubject#Female' shows the combined effect of a male video subject and the participant being female. Standard errors are clustered by participant.

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

3.5. Results: Overall Beliefs

3.5.1. Overall number of videos categorised as real

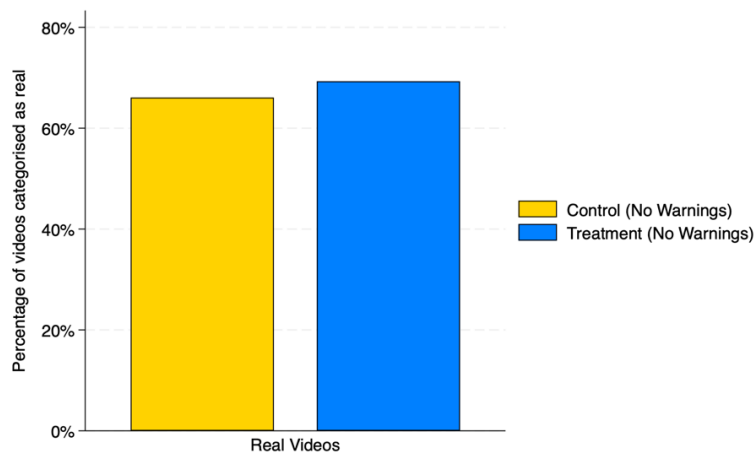
Participants categorised 57.18% of the sixteen videos as being real ($M = 9.15$, $SD = 2.14$), compared to just 42.82% as being deepfakes ($M = 6.85$, $SD = 2.13$). Given that only half of the videos were real, and half were deepfakes, this was a significant overestimation of real videos ($p < 0.05$) and a significant underestimation of deepfake videos ($p < 0.05$). In fact, 61.35% of participants thought there were more than 8 real videos, whereas only 23.31% thought there were more than 8 deepfakes. Only 15.34% guessed exactly 8 real videos and 8 deepfake videos as per the study instructions. Hypothesis 3 was therefore supported.

To examine whether content warnings predict overall authenticity perceptions, we followed the approach of Pennycook et al. (2020) by disaggregating the number of real responses by video type (i.e. real or deepfake), experimental condition (i.e. control or treatment) and whether the videos had warnings (i.e. no warnings in the control group, warnings present in 4 of the deepfake videos in the treatment group, warnings absent in the remaining 4 deepfakes in the treatment group).

Figure 6 first disaggregates the number of correct ‘real’ responses for the real videos. On average, the control group categorised 66.07% ($SD = 18.96$) of the real videos as being real, with this result being 69.30% ($SD = 18.85$) for the treatment group and the difference was significant ($p < 0.05$). Figure 7 disaggregates the number of ‘real’ responses for the deepfake videos according to experimental condition and, within the treatment group, whether a video had a warning.

Figure 6

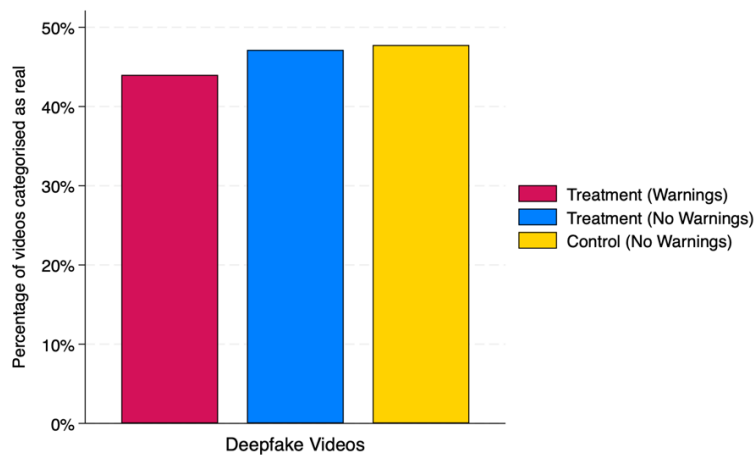
Percentage of Real Videos Categorised as Real by Condition



Note: Bar chart presenting the percentage of real videos categorised as being real by control and treatment groups (i.e. responded 'real' when asked to categorise each video during the experimental session). There were no warnings applied to real videos during the experiment meaning there are only two conditions to compare in this bar chart. Plot adapted from Figure 2, Pennycook et al. (2020) and created in Stata.

Figure 7

Percentage of Deepfake Videos Categorised as Real by Condition and Warning Status



Note: Bar chart presenting the percentage of deepfake videos categorised as being real by control and treatment groups (i.e. responded 'real' when asked to categorise each video during the experimental session). The bar chart disaggregates the videos according to treatment condition and whether a warning was present. Warnings were not applied to any of the deepfake videos watched by the control group, but in the treatment group half of the deepfake videos had warnings and half did not have warnings. As such, the bar chart compares three different conditions (i.e. control group (no warnings), treatment group (warnings) and treatment group (no warnings)). Plot adapted from Figure 2, Pennycook et al. (2020) and created in Stata.

As can be seen, the bar with the fewest ‘real’ responses is for the deepfake videos that were tagged with warnings in the treatment group (43.98%, $SD = 27.92$). In comparison, the treatment group categorised 47.16% ($SD = 27.57$) of deepfake videos without warnings as real, and the control group categorised 47.76% ($SD = 18.92$) of the deepfake videos without warnings as real. To calculate the effect of the warning on the likelihood of categorising the deepfake videos as real, we compared the proportion of tagged deepfake videos categorised as real by the treatment group with the proportion of deepfake videos categorised as real by the control group. The difference between these groups was statistically significant ($t = 4.07$, $p < 0.05$). We then calculated Cohen’s d to quantify the effect size which was found to be -0.16, indicating that the warnings reduced the likelihood of participants in the treatment group categorising the tagged deepfake videos as real compared to the control group, although effect sizes are small.

To test the robustness of our findings, we ran logit regressions at the level of the video rating with standard errors clustered at the video level. Per the approach in Pennycook et al. (2020), we tested for the presence of a warning effect with a ‘warned’ dummy variable that indicated when a deepfake video had a warning in the treatment group, and a ‘deepfake’ dummy to indicate when a video was a deepfake. Results of the logistic regression are summarised in Appendix E, Tables E6 to E7. As can be seen, warnings did not significantly reduce the likelihood of categorising a video as real when analysed at the individual video level, meaning people were still biased in their estimations. Hypothesis 4 was not supported.

3.5.2. Unintended consequences

We do not find evidence of a backfire effect in terms of warnings increasing the overall number of videos categorised as real. This is evident from the negative coefficient on the ‘warned’ dummy in Appendix E, Table E7. We also tested whether warnings led to a

decrease in the number of real videos being categorised as real. Logistic regression results in Appendix E, Table E9 do not show evidence of this effect, meaning Hypothesis 5 was not supported.

We tested for an implied truth effect with an ‘untagged’ dummy variable that indicated a deepfake video in the treatment group not having a warning, with a dummy variable indicating whether a video was a deepfake. As can be seen from Appendix E, Table E8, although the absence of a warning on a deepfake video was positively associated with the video being categorised as real, the effect was not statistically significant ($p > 0.05$). Consequently, we do not find evidence of an implied truth effect and Hypothesis 6 was not supported.

4. Discussion

4.1. Answers to research questions

RQ1: How do humans perform at accurately categorising videos as real or deepfakes, and what does this imply about susceptibility to deepfake content?

On average, participants accurately categorised 60.46% of the videos as real or deepfakes ($SD = 14.55$), which was significantly greater than chance levels of accuracy. Our literature review found that human accuracy ratings in deepfake detection studies are typically above chance, with most finding accuracy of between 57.6% to 62%. Our overall accuracy rating sits between the 60% accuracy found by Bray et al. (2023) and the 60.70% accuracy found by Somoray and Miller (2023), suggesting average human accuracy is clustered at $\approx 60\%$. Our findings contrast with the 80% accuracy rating found by Nas and Kleijn (2024), who also used the Celeb-DF dataset for their stimulus videos. This may be due to the smaller, younger and more motivated sample ($N = 130$, $M_{age} = 20$, females = 84.61%) of mostly university students, who received course credits for taking part in the study.

Our findings as they relate to overall accuracy suggest that while humans are better than chance at discerning between real versus fake videos, they make systematic errors nearly forty percent of the time, which by implication leaves them vulnerable to believing a deepfake is authentic when it is not. Categorisation errors of this magnitude ought to be of concern to policymakers given the increasing prevalence of deepfakes. Urgent additional research is needed on behavioural interventions that can significantly improve accuracy as a complement to advances in automated detection.

RQ2. Do individuals show an overall bias towards categorising videos as real or deepfakes, and if so, what does this imply about susceptibility to deepfake content?

In line with our hypothesis, our study found that participants were biased towards categorising videos as real. This implies susceptibility to believing a deepfake is authentic through overestimation of authentic content. Despite disagreement in the literature, our findings support the existence of a seeing is believing heuristic, where individuals assume that a video is real unless there is obvious evidence to the contrary, leading to biased perceptions (Köbis et al., 2021; Somoray & Miller, 2023). Our findings also support the existence of a realism heuristic (Sundar, 2008) given the overestimation of real videos.

Our finding of bias towards authenticity is interesting given that we expressly informed participants beforehand that half of the videos would be real, and half would be deepfakes. Most participants (61.35%) deviated from these instructions and thought there were more than eight real videos, whereas only a minority (23.31%) did so for the deepfakes. Only 15.34% of participants guessed exactly 8 real videos and 8 deepfakes. This may suggest a baseline assumption of authenticity in the context of visual media, with participants struggling to adjust to new conditions that do not assume total authenticity (Köbis et al., 2021). Future research could investigate this further, for example by using different

proportions of real and deepfake videos. Alternatively, participants could be asked to reflect on whether they categorised videos in the instructed proportions and examine their reasons for deviating from the study instructions.

RQ3. If individuals are susceptible to deepfake content, do content warnings reduce susceptibility?

Contrary to our hypothesis, content warnings did not significantly reduce people's susceptibility to deepfake material from either a truth discernment or overall beliefs perspective. From a truth discernment perspective, we hypothesised an increase in video accuracy scores. From an overall belief perspective, we hypothesised a decrease in the overall number of videos categorised as real. We did not observe either effect. This finding is consistent with the literature that suggests mixed effectiveness of warnings to reduce susceptibility to deepfakes. The only deepfake study we are aware of that found warnings to be effective tested a different outcome, namely belief in a message delivered by a deepfake celebrity, as opposed to detecting whether the video itself was a deepfake (Ahmed, 2021).

We do not believe lack of salience of warnings contributed to the absence of effects. As shown in Appendix F, Table F1, 93.67% of participants in the treatment group confirmed noticing warnings during the experiment. Part of the answer may lie in the mixed levels of trust in the warnings and an apparent preference for deliberation before decision-making, as shown in Appendix F, Tables F2 and F3. While reflective cognition is encouraging, it may have resulted in a backfire effect for some participants, offsetting the overall effect of the warnings. It is also possible that the warning we chose for our study is suboptimal. Research from the field of misinformation has mixed findings on the effectiveness of fact-checker labels, with some commentators finding them to be highly effective (Martel & Rand, 2023), and others suggesting they cause uncertainty, frustration and distrust (Guo et al., 2024). The

lack of effect of warnings may suggest a more conclusive warning about authenticity is needed for deepfakes and this should be investigated in future research.

RQ4: Could content warnings have any unintended consequences through backfire effects or implied truth effects?

We did not find evidence that content warnings led to a backfire effect in terms of increasing overall bias towards authenticity. Nor did we find evidence that warnings lead people to discredit authentic videos (i.e. a liar's dividend). These are promising findings since prior deepfake research found that warnings reduced belief in authentic political videos (Ternovski et al., 2022). Since our study did not feature any political or ideological, this may suggest warnings do not reduce belief in authentic non-partisan video content.

We did not find evidence of an implied truth effect in our study, despite our hypothesis that we would do so based on the findings of Pennycook et al. (2020) in the context of textual misinformation. We suggest this is due to reflection and deliberation, since most treatment group participants indicated that they still considered a video could be a deepfake despite the absence of a warning (Appendix F, Table F4). Very few (4.05%) participants felt that untagged videos were more likely to be real than a deepfake. Although these are reassuring findings, further research is needed. Pennycook et al. (2020) did not expand on the mechanisms that explain an implied truth effect and further research is needed on this type of unintended consequence. Additionally, although we do not find evidence of an implied truth effect, it was apparent from our data analysis that the absence of warnings in videos may still have driven uncertainty for untagged videos.

4.2. Additional findings

Our study provides additional insights outside of the scope of our research questions. An interesting nuance is found in the influence of video status on truth discernment and overall belief. Our analysis from a truth discernment perspective suggests participants

struggled to correctly categorise the deepfake videos compared to the real videos, resulting in lower accuracy scores. However, when analysed from an overall belief perspective, participants were overall less likely to categorise the deepfake videos as real compared to real videos. This is a subtle nuance suggesting that although participants often made inaccurate decisions, they were still more suspicious of the deepfake videos in terms of overall perceptions.

We found unexpected gender influences in our study, finding that female participants were more accurate overall at categorising videos, but their accuracy decreased when the video subject was male. We suggest this as an avenue for future research as it suggests heterogeneity in detection accuracy, meaning blanket interventions may not work for all demographics. Separately, although we found that most of our sample were previously aware of deepfakes (93.87%) and had been exposed to one (58.28%), we did not find any association between prior exposure and awareness on either accuracy or overall perceptions.

We find overwhelmingly negative sentiment towards deepfakes generally. Over three-quarters of participants who provided free-text responses to a question about their attitudes towards deepfake technology held either negative or very negative views. The most frequently cited concern was potential for deception, followed by misinformation and privacy concerns. Appendix H summarises our sentiment analysis and supports calls on a policy level for greater regulation of deepfakes.

4.3. Applied implications

Our findings have practical implications for policymakers and firms. First, interventions that may have worked to mitigate susceptibility to textual misinformation may not be as effective for multimodal content such as deepfakes. Our study finds evidence of bias towards perceiving multimodal content as real, with the literature hypothesising that

such content is perceived as more credible and trustworthy (Sundar, 2008). This may render contextual warnings less effective for this type of content. Consequently, legal interventions mandating the use of labelling requirements for deepfake content, such as those under the new EU AI Act, may be misguided. From a policymaking perspective, national campaigns that educate members of the public about bias towards authenticity may offer a way forward (Hwang et al., 2021), but these would need to be carefully designed to avoid creating oversensitivity to deepfakes and a *liar's dividend* problem (Chesney & Citron, 2018).

Second, policymakers must be alive to the risk of deepfake content bypassing traditional flagging systems. Our literature review suggests that people are increasingly sharing deepfakes through encrypted channels that may evade content flagging systems and fact-checkers (Pinhanez et al., 2022). Given our findings that people fail to accurately discern between real and deepfake videos approximately forty percent of the time, the use of private channels to share deepfake content is likely to worsen susceptibility to deceptive content. A way forward is for policymakers to collaborate with platform developers to embed behavioural interventions within such channels. Future research should test embedded interventions, but carefully measure unintended consequences since people likely use encrypted channels to avoid firm-level regulation in the first place. Consequently, they might exhibit strong reactance to within-channel interventions.

Third, the absence of an implied truth effect lends support for a pragmatic approach to content warnings by firms who choose to deploy them. This means fact-checkers could continue to prioritise the labelling of misleading content, but not take exhaustive steps to label authentic videos too. Excessive labelling could lead to oversensitivity to manipulated content, which as noted in our literature review could lead to overestimating the amount of deepfakes and a *liar's dividend* problem (Chesney & Citron, 2018).

4.4. Limitations

Although we believe our research advances the literature on human detection and perception of deepfake material, we acknowledge several limitations. First, we recruited participants using snowball sampling and from the recruitment platform Prolific and with geographic criteria. Whilst we controlled for recruitment source when running our analysis to control for difference in motivations due to Prolific participants being paid, non-probability samples may not generalise to wider populations (Pasek, 2015). Our study provides insights on human detection and perception of deepfake material in western countries but may not generalise to other regions.

Second, our stimulus videos featured high-profile celebrities and did not include audio. It has been suggested that the use of public figures in deepfake research could introduce bias into experiments due to familiarity and support for the celebrity outside of the experiment (Köbis et al., 2021). Although our literature review found that familiarity could influence truth judgments (Berinsky, 2017; Newman et al., 2015), we did not measure participants' familiarity with the video subjects as we did not wish to fatigue participants. We were therefore unable to make conclusions about the relevance of the fluency heuristic and cite this as an opportunity for further research. Audio may influence detection and perception of deepfake content (Ahmed & Chua, 2023; Nas & Kleijn, 2024), and the absence of audio limits our findings to purely visual modalities.

Third, while the exploration of detection and perception of deepfakes in a non-partisan context provides valuable insights, we acknowledge that deepfakes are often used as a vector of political misinformation. The literature has found individuals may believe information that aligns with their political beliefs (Kahan et al., 2017; Kunda, 1990) and we were unable to test whether this replicates to non-textual misinformation modalities in this study.

Fourth, although our stimulus videos were taken from the same dataset and we adopted rigorous methodology in selecting them, there may have been inherent differences in the artifacts and visual quality. This may have resulted in some of the videos being easier to categorise as real or deepfakes, although our findings do suggest that deepfake videos were harder to categorise than real videos overall.

Fifth, our research may lack ecological validity since the experimental conditions are not representative of how people encounter deepfake content online. For example, people are highly unlikely to be warned about deepfakes in advance (Nas & Kleijn, 2024), and they are unlikely to encounter real and deepfake videos in equal proportions. We therefore echo the modesty of Somoray and Miller (2023) that our study offers a conservative estimate of susceptibility to deepfake content given the controlled nature of the experiment.

5. Conclusion

Deepfakes are an increasingly realistic and democratised technology that have the potential to cause extreme harm. Policymakers must take urgent steps to reduce susceptibility to harmful deepfake material. Our study contributes to a growing body of research on susceptibility to deepfakes by measuring both video accuracy judgments and overall perceptions. We find that humans struggle to discern between real and deepfake media and are overall biased towards perceiving videos as real, leaving them susceptible to believing a manipulated video is real when it is in fact fake. Our study shows the limitations of adding content warnings to deepfakes, but reassuringly does not find evidence of unintended consequences through backfire or implied truth effects. Together, our findings could help inform future policymaking and complement the design of technical and legal interventions to reduce susceptibility to manipulated content.

References

- Ahmed, S. (2021). Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. *Personality and Individual Differences*, 182, 111074. <https://doi.org/10.1016/j.paid.2021.111074>
- Ahmed, S., & Chua, H. W. (2023). Perception and deception: exploring individual responses to deepfakes across different modalities. *Heliyon*, 9(10) e20383. <https://doi.org/10.1016/j.heliyon.2023.e20383>
- Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2), 211–236. <https://doi.org/10.1257/jep.31.2.211>
- Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General*, 149(8), 1608–1613. <https://doi.org/10.1037/xge0000729>
- Berinsky, A. J. (2017). Rumors and health care reform: experiments in political misinformation. *British Journal of Political Science*, 47(2), 241–262. <https://doi.org/10.1017/S0007123415000186>
- Bontrager, P., Roy, A., Togelius, J., Memon, N., & Ross, A. (2018, October). Deepmasterprints: Generating masterprints for dictionary attacks via latent variable evolution. In *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)* (pp. 1-9). IEEE. <http://doi.org/10.1109/BTAS.2018.8698539>
- Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect ‘deepfake’ images of human faces. *Journal of Cybersecurity*, 9(1), tyad011. <https://doi.org/10.1093/cybsec/tyad011>

- Bristow, T. (2023, October 9) *Keir Starmer suffers UK politics' first deepfake moment. It won't be the last*. POLITICO. <https://www.politico.eu/article/uk-keir-starmer-labour-party-deepfake-ai-politics-elections/>
- Burkhardt, J. M. (2017). Chapter 1. History of Fake News. *Library Technology Reports*, 53(8), 5-9. <https://doi.org/10.5860/ltr.53n8>
- Caldwell, M., Andrews, J. T., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. *Crime Science*, 9(1), 1-13. <https://doi.org/10.1186/s40163-020-00123-8>
- Chesney, R., & Citron, D. K. (2018). Deep Fakes: a looming challenge for privacy, democracy, and national security. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.3213954>
- Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., ... & Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*, 42(4), 1073-1095. <https://doi.org/10.1007/s11109-019-09533-0>
- Cook, J. (2019, June 23). Here's what it's like to see yourself in a deepfake porn video. *HuffPost UK*. https://www.huffpost.com/entry/deepfake-porn-heres-what-its-like-to-see-yourself_n_5d0d0face4b0a3941861fced
- Darnton, R. (2017, February 13). The True History of Fake News. *The New York Review of Books*. <https://www.nybooks.com/online/2017/02/13/the-true-history-of-fake-news/>
- Dobber, T., Metoui, N., Trilling, D., Helberger, N., & De Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, 26(1), 69-91. <https://doi.org/10.1177/1940161220944364>
- Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397* <https://doi.org/10.48550/arXiv.2006.07397>

- Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods*, 39(2), 175-191. <https://doi.org/10.3758/bf03193146>
- Fletcher, J. (2018). Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance. *Theatre Journal*, 70(4), 455-471. <https://doi.org/10.1353/tj.2018.0097>
- Frenda, S. J., Knowles, E. D., Saletan, W., & Loftus, E. F. (2013). False memories of fabricated political events. *Journal of Experimental Social Psychology*, 49(2), 280–286. <https://doi.org/10.1016/j.jesp.2012.10.013>
- Gilbert, D. T., Tafarodi, R. W., & Malone, P. S. (1993). You can't not believe everything you read. *Journal of Personality and Social Psychology*, 65(2), 221–233. <https://doi.org/10.1037/0022-3514.65.2.221>
- Graham. (2024, June 26). *Deepfakes: Federal and state regulation aims to curb a growing threat*. Thomson Reuters Institute. <https://www.thomsonreuters.com/en-us/posts/government/deepfakes-federal-state-regulation/>
- Greene, C. M., de Saint Laurent, C., Murphy, G., Prike, T., Hegarty, K., & Ecker, U. K. (2022). Best practices for ethical conduct of misinformation research. *European Psychologist*, 28(3), 139–150. <https://doi.org/10.1027/1016-9040/a000491>
- Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1), e2110013119. <https://doi.org/10.1073/pnas.2110013119>
- Guo, C., Guo, Z., Zheng, N., & Guo, C. (2024). All warnings are not equal: a user-centered approach to comparing general and specific contextual warnings against misinformation. In *Proceedings of the 57th Hawaii International Conference on System Sciences*.

<https://scholarspace.manoa.hawaii.edu/server/api/core/bitstreams/17efa1a4-f3c5-407f-b0ff-5536287f0c0b/content>

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

<https://doi.org/10.1080/10584609.2019.1674979>

Hao, K. (2020, October 2020). A deepfake bot is being used to “undress” underage girls. *MIT Technology Review*. <https://www.technologyreview.com/2020/10/20/1010789/ai-deepfake-bot-undresses-women-and-underage-girls/>

Herbert Smith Freehills. (2023). *AI-deep synthesis regulations and legal challenges: recent face swap fraud cases in China*.

<https://www.herbertsmithfreehills.com/notes/tmt/2023-07/ai-deep-synthesis-regulations-and-legal-challenges-recent-face-swap-fraud-cases-in-china>

Hernandez, I., & Preston, J. L. (2013). Disfluency disrupts the confirmation bias. *Journal of Experimental Social Psychology*, 49(1), 178–182.

<https://doi.org/10.1016/j.jesp.2012.08.010>

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

Iacobucci, S., De Cicco, R., Michetti, F., Palumbo, R., & Pagliaro, S. (2021). Deepfakes unmasked: the effects of information priming and bullshit receptivity on deepfake recognition and sharing intention. *Cyberpsychology, behavior, and social networking*, 24(3), 194-202. <https://doi.org/10.1089/cyber.2020.0149>

- Kahan, D. M., Peters, E., Dawson, E. C., & Slovic, P. (2017). Motivated numeracy and enlightened self-government. *Behavioural Public Policy*, 1(1), 54–86.
<https://doi.org/10.1017/bpp.2016.2>
- Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.
- Kelley, N. J., Hurley-Wallace, A. L., Warner, K. L., & Hanoch, Y. (2023). Analytical reasoning reduces internet fraud susceptibility. *Computers in Human Behavior*, 142, 107648. <https://doi.org/10.1016/j.chb.2022.107648>
- Khalid, H., Tariq, S., Kim, M., & Woo, S. S. (2022). FakeAVCeleb: A novel audio-video multimodal deepfake dataset (arXiv:2108.05080). *arXiv preprint arXiv:2108.05080*.
<https://doi.org/10.48550/arXiv.2108.05080>
- Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: people cannot detect deepfakes but think they can. *Iscience*, 24(11).
<https://doi.org/10.1016/j.isci.2021.103364>
- Korshunov, P., & Marcel, S. (2020). Deepfake detection: humans vs. machines. *arXiv preprint arXiv:2009.03155*. <https://doi.org/10.48550/arXiv.2009.03155>
- Kouzy, R., Abi Jaoude, J., Kraitam, A., El Alam, M. B., Karam, B., Adib, E., Zarka, J., Traboulsi, C., Akl, E. W., & Baddour, K. (2020). Coronavirus goes viral: quantifying the Covid-19 misinformation epidemic on Twitter. *Cureus*, 12(3), e7255.
<https://doi.org/10.7759/cureus.7255>
- Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, 108(3), 480–498. <https://doi.org/10.1037/0033-2909.108.3.480>
- Lawmaker uses AI voice clone to address Congress. (2024, July 25). *BBC News*.
<https://www.bbc.co.uk/news/videos/c728q850e5do>

- Lewis, A., Vu, P., Duch, R. M., & Chowdhury, A. (2023). Deepfake detection with and without content warnings. *Royal Society Open Science*, 10(11), 231214.
<https://doi.org/10.1098/rsos.231214>
- Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. *In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3207-3216).
<https://doi.org/10.48550/arXiv.1909.12962>
- Linden, S. van der. (2023). *Foolproof: Why Misinformation Infects Our Minds and How to Build Immunity*. W. W. Norton & Company.
- List, J. A., Sadoff, S., & Wagner, M. (2011). So you want to run an experiment, now what? Some simple rules of thumb for optimal experimental design. *Experimental Economics*, 14(4), 439–457. <https://doi.org/10.1007/s10683-011-9275-7>
- Lodge, M., & Taber, C. (2000). Three steps toward a theory of motivated political reasoning. In A. Lupia, M. D. McCubbins, & S. L. Popkin (Eds.), *Elements of Reason: Cognition, Choice, and the Bounds of Rationality* (pp. 183–213). Cambridge University Press. <https://doi.org/10.1017/CBO9780511805813.009>
- LSE Research Ethics Committee. (2023, December). Research Ethics Policy and Procedures. Retrieved from <https://info.lse.ac.uk/staff/services/Policies-and-procedures/Assets/Documents/resEthPolPro.pdf>
- LSE Research Ethics Committee. (2024, April). Code of Research Conduct. Retrieved from <https://info.lse.ac.uk/staff/services/Policies-and-procedures/Assets/Documents/codResCon.pdf>
- Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive research: principles and implications*, 5, 1-20.
<https://doi.org/10.1186/s41235-020-00252-3>

- Martel, C., & Rand, D. G. (2023). Misinformation warning labels are widely effective: A review of warning effects and their moderating features. *Current Opinion in Psychology*, 101710. <https://doi.org/10.1016/j.copsyc.2023.101710>
- Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM computing surveys (CSUR)*, 54(1), 1-41. <https://doi.org/10.1145/3425780>
- MIT Media Lab. (n.d.). *Project Overview < Detect DeepFakes: How to counteract misinformation created by AI*. Retrieved 19 August 2024, from <https://www.media.mit.edu/projects/detect-fakes/overview/>
- Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research*, 154, 113368. <https://doi.org/10.1016/j.jbusres.2022.113368>
- Nas, E., & De Kleijn, R. (2024). Conspiracy thinking and social media use are associated with ability to detect deepfakes. *Telematics and Informatics*, 87, 102093. <https://doi.org/10.1016/j.tele.2023.102093>
- Newman, E. J., Garry, M., Unkelbach, C., Bernstein, D. M., Lindsay, D. S., & Nash, R. A. (2015). Truthiness and falsiness of trivia claims depend on judgmental contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41(5), 1337. <https://doi.org/10.1037/xlm0000099>
- Nickerson, R. S. (1998). Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2), 175–220. <https://doi.org/10.1037/1089-2680.2.2.175>
- Nightingale, S. J., Wade, K. A., & Watson, D. G. (2017). Can people identify original and manipulated photos of real-world scenes?. *Cognitive research: principles and implications*, 2, 1-21. <https://doi.org/10.1186/s41235-017-0067-2>

- Nyhan, B., & Reifler, J. (2010). When corrections fail: the persistence of political misperceptions. *Political Behavior*, 32(2), 303–330. <https://doi.org/10.1007/s11109-010-9112-2>
- Office for Artificial Intelligence. (2023) *A pro-innovation approach to AI regulation*. <https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper>
- Online Safety Act 2023*, c.50. <https://www.legislation.gov.uk/ukpga/2023/50/section/187>
- Pasek, J. (2015). Beyond probability sampling: population inference in a world without benchmarks. *Available at SSRN 2804297*. <https://doi.org/10.2139/ssrn.2804297>
- Pennycook, G., Cheyne, J. A., Seli, P., Koehler, D. J., & Fugelsang, J. A. (2012). Analytic cognitive style predicts religious and paranormal belief. *Cognition*, 123(3), 335-346. <https://doi.org/10.1016/j.cognition.2012.03.003>
- Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12), 1865. <https://doi.org/10.1037/xge0000465>
- Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management science*, 66(11), 4944-4957. <https://doi.org/10.1287/mnsc.2019.3478>
- Pennycook, G., & Rand, D. G. (2021a). The Psychology of Fake News. *Trends in Cognitive Sciences*, 25(5), 388–402. <https://doi.org/10.1016/j.tics.2021.02.007>
- Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021b). Shifting attention to accuracy can reduce misinformation online. *Nature*, 592(7855), 590-595. <https://doi.org/10.1038/s41586-021-03344-2>

- Pinhanez, C. S., Flores, G. H., Vasconcelos, M. A., Qiao, M., Linck, N., de Paula, R., & Ong, Y. J. (2022). *Towards a New Science of Disinformation* (arXiv:2204.01489). *arXiv preprint arXiv:2204.01489*. <https://doi.org/10.48550/arXiv.2204.01489>
- Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 Laying down Harmonised Rules on Artificial Intelligence and Amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA Relevance) (2024). <http://data.europa.eu/eli/reg/2024/1689/oj/eng>
- Romero Moreno, F. (2024). Generative AI and deepfakes: a human rights approach to tackling harmful content. *International Review of Law, Computers & Technology*, 0(0), 1–30. <https://doi.org/10.1080/13600869.2024.2324540>
- Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 1-11). <https://doi.org/10.48550/arXiv.1901.08971>
- 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. <https://doi.org/10.1080/21670811.2022.2026797>
- Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior*, 149, 107917. <https://doi.org/10.1016/j.chb.2023.107917>
- Spring, M. (2024, March 4). Trump supporters target black voters with faked AI images. *BBC News*. <https://www.bbc.com/news/world-us-canada-68440150>

- Sundar, S. S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility (pp. 73-100). In M.J. Metzger & A.J. Flanagin (Eds.), *Digital media, youth, and credibility* (pp.73-100). The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning Initiative. The MIT Press.
<https://www.issuelab.org/resources/875/875.pdf>
- Swami, V., Voracek, M., Stieger, S., Tran, U. S., & Furnham, A. (2014). Analytic thinking reduces belief in conspiracy theories. *Cognition*, 133(3), 572-585.
<https://doi.org/10.1016/j.cognition.2014.08.006>
- Ternovski, J., Kalla, J., & Aronow, P. (2022). The negative consequences of informing voters about deepfakes: evidence from two survey experiments. *Journal of Online Trust and Safety*, 1(2). <https://doi.org/10.54501/jots.v1i2.28>
- Thaler, R. H., & Sunstein, C.R. (2008). *Nudge: Improving decisions about health, wealth, and happiness*. Yale University Press.
- Treen, K. M. d'I., Williams, H. T. P., & O'Neill, S. J. (2020). Online misinformation about climate change. *WIREs Climate Change*, 11(5), e665. <https://doi.org/10.1002/wcc.665>
- Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, 185(4157), 1124-1131. <https://doi.org/10.1126/science.185.4157.1124>
- Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social media+ society*, 6(1), 2056305120903408. <https://doi.org/10.1177/2056305120903408>
- Van Aelst, P., Strömbäck, J., Aalberg, T., Esser, F., de Vreese, C., Matthes, J., Hopmann, D., Salgado, S., Hubé, N., Stępińska, A., Papathanassopoulos, S., Berganza, R., Legnante, G., Reinemann, C., Sheafer, T., & Stanyer, J. (2017). Political communication in a high-choice media environment: A challenge for democracy? *Annals of the*

International Communication Association, 41(1), 3–27.

<https://doi.org/10.1080/23808985.2017.1288551>

Vizoso, Á., Vaz-Álvarez, M., & López-García, X. (2021). Fighting deepfakes: media and internet giants' converging and diverging strategies against hi-tech misinformation.

Media and Communication, 9(1), 291-300. <https://doi.org/10.17645/mac.v9i1.3494>

Wardle, C., & Derakhshan, H. (2017). Information disorder: toward an interdisciplinary framework for research and policymaking (Vol. 27, pp. 1-107). Strasbourg: Council of Europe. <https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html#>

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3), 129–140.

<https://doi.org/10.1080/17470216008416717>

Weikmann, T., Greber, H., & Nikolaou, A. (2024). After deception: how falling for a deepfake affects the way we see, hear, and experience media. *The International Journal of Press/Politics*, 19401612241233539.

<https://doi.org/10.1177/19401612241233539>

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology innovation management review*, 9(11), 39-52. <http://doi.org/10.22215/timreview/1282>

Yang, Z. (2024, May 7). *Deepfakes of your dead loved ones are a booming Chinese business*. MIT Technology Review.

<https://www.technologyreview.com/2024/05/07/1092116/deepfakes-dead-chinese-business-grief/>

Appendix A

Participant demographics and randomisation

Table A1

Participant Demographics (n = 163)

	Frequency	%
Gender		
Female	85	52.15
Male	70	42.94
Non-binary	5	3.07
Other	1	0.61
Transgender	2	1.23
Location		
Australia	6	3.68
Canada	19	11.66
New Zealand	5	3.07
The United Kingdom	49	30.06
The United States of America	21	12.88
Within the European Union	63	38.65
Ethnicity		
Asian / Asian British	14	8.59
Black/African/Caribbean/Black British	4	2.45
Mixed/Multiple ethnic groups	6	3.68
Other ethnic group	5	3.07
White / Caucasian	134	82.21
Occupation		
Full-time employment	52	31.90
Inability to work	5	3.07
Part-time employed	23	14.11
Retired	1	0.61
Self-employed / freelancer	6	3.68
Student	50	30.67
Unemployed	26	15.95
Education		
Associate degree	6	3.68
Bachelor degree	65	39.88
Doctorate degree	3	1.84
High school/college graduate, diploma or equivalent	39	23.93
Master degree	30	18.40
Some high school	9	5.52
Trade/technical/vocational training	11	6.75
Aware of deepfakes		
Yes	153	93.87
No	10	6.13
Exposed to a deepfake		
Yes	95	58.28
No	13	7.98
Unsure	55	33.74

Note: Table A1 summarises our sample of n = 163 participants by observed characteristics.

Table A2*Randomisation Balance Checks*

	Control (1)	Treatment (2)	Difference (3)
Age	30.750 (10.098)	29.139 (9.400)	-1.611 (1.531)
Gender	1.488 (0.611)	1.633 (0.787)	0.145 (0.110)
Ethnicity	4.476 (1.285)	4.481 (1.175)	0.005 (0.193)
Location	4.690 (1.439)	4.354 (1.545)	-0.336 (0.234)
Occupation	3.750 (2.434)	4.203 (2.404)	0.453 (0.379)
Education	3.512 (1.718)	3.633 (1.627)	0.121 (0.262)
Aware of deepfakes	0.929 (0.259)	0.949 (0.221)	0.021 (0.038)
Exposed to a deepfake	1.869 (0.954)	1.633 (0.894)	-0.236 (0.145)
Observations	84	79	163

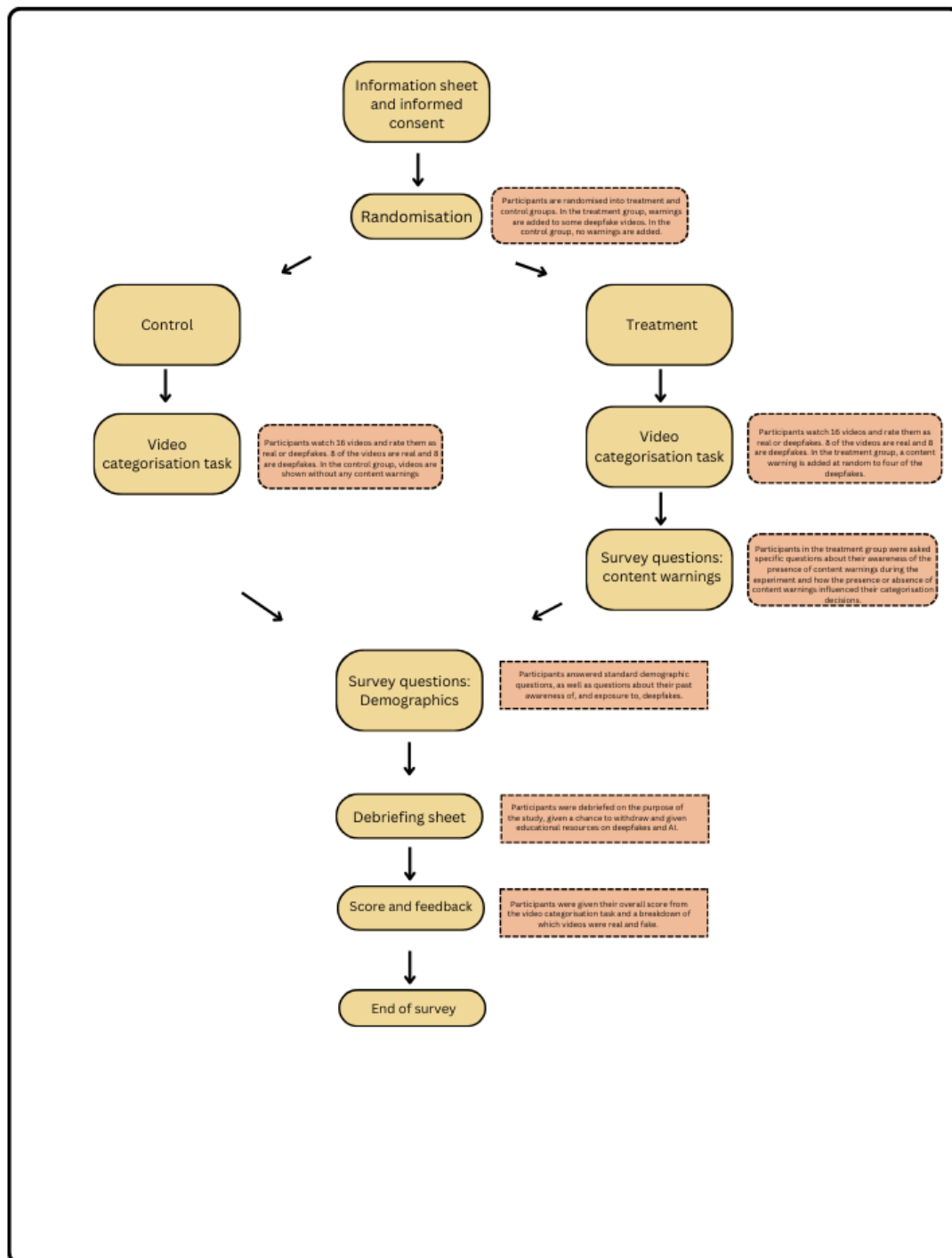
Note: Table A2 summarises the number of control and treatment group participants following randomisation. Columns 1 and 2 show the means for various observed characteristics within the control and treatment groups respectively. Column 3 shows the difference in means between the groups. Standard errors are in parenthesis. Separate t-tests confirmed that none of the differences were statistically significant at the five per cent level. Table created using balancetable package in Stata.

Appendix B

Overview of experimental design

Figure B1

Flowchart Summarising Experimental Design



Appendix C

Stimulus Videos

Table C1

Summary of Stimulus Videos

Video Number	Video Status	Subject Gender	Subject Ethnicity
1	Real	Male	White
2	Real	Male	White
3	Real	Male	White
4	Real	Female	White
5	Real	Female	White
6	Real	Male	Black
7	Real	Male	Black
8	Real	Male	White
9	Deepfake	Male	White
10	Deepfake	Female	White
11	Deepfake	Male	White
12	Deepfake	Female	White
13	Deepfake	Female	White
14	Deepfake	Female	White
15	Deepfake	Female	White
16	Deepfake	Male	Black

Figure C1

Example of Real Video and Deepfake Counterpart from the Celeb-DF Dataset



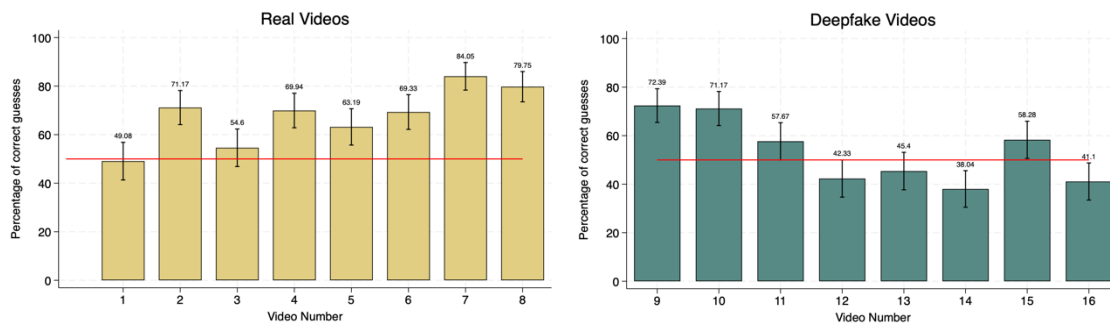
Note: Example videos from the Celeb-DF dataset (Li et al., 2020). On the left is an example of an authentic source video. On the right is an example deepfake counterpart.

Appendix D

Supplementary Plots

Figure D1

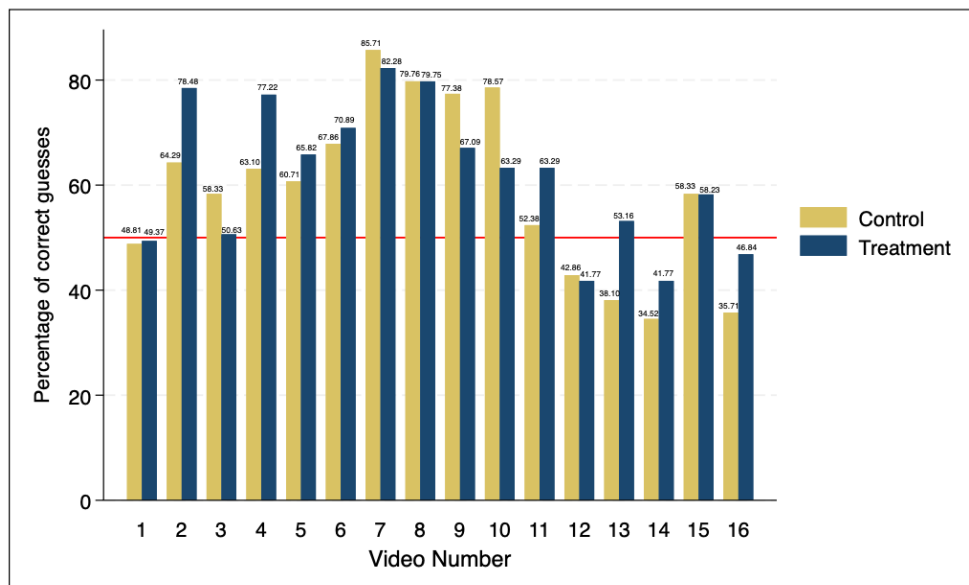
Overall Video Categorisation Accuracy by Video Type



Note: Percentage of correct guesses for each video for all participants per video type, independent of experimental condition. Error bars denote 95% confidence intervals and the reference line in red indicates the probability of guessing correctly based on chance alone. Plot adapted from Figure 2 of Köbis et al., 2021 and created in Stata.

Figure D2

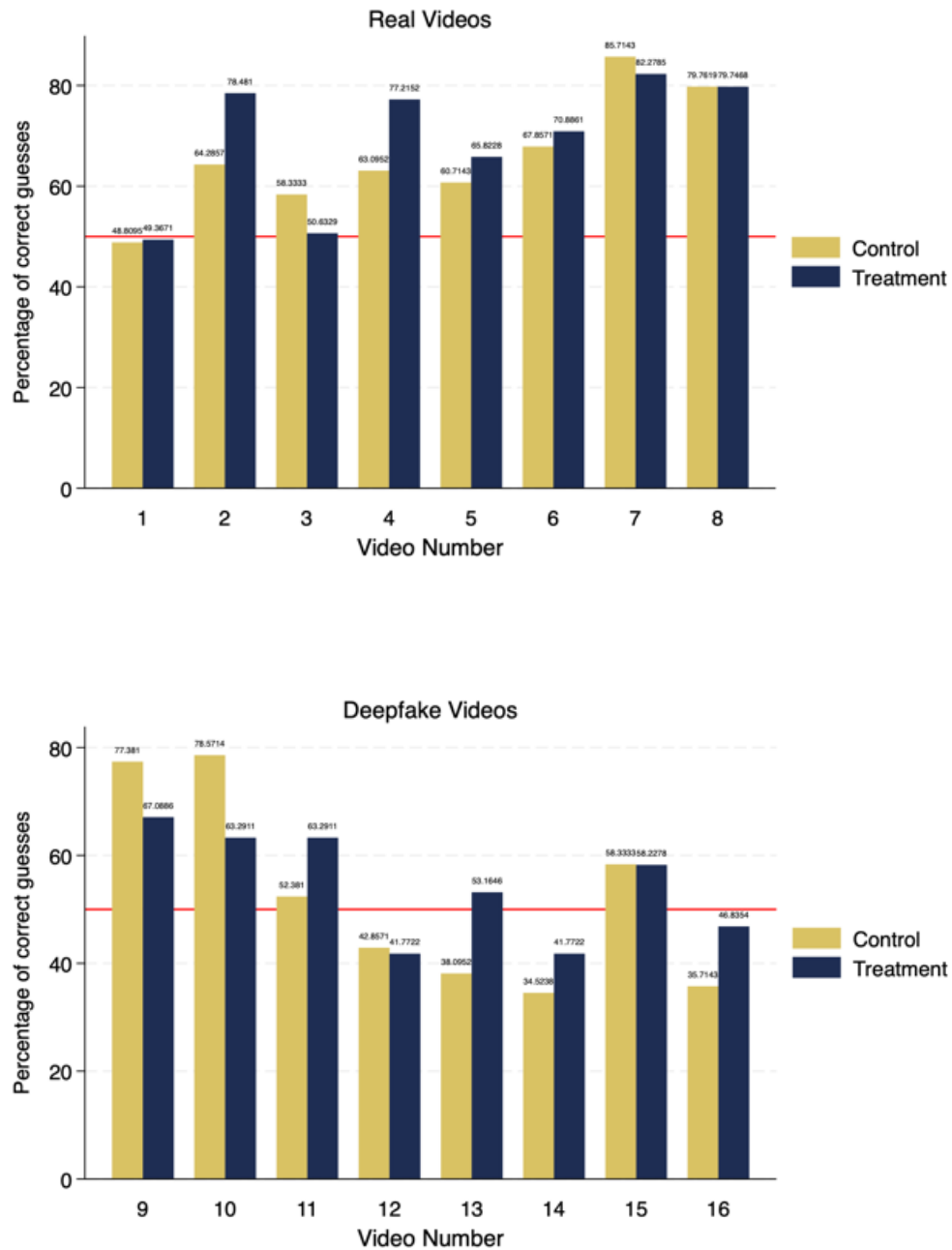
Video Categorisation Accuracy by Condition



Note: Comparison of percentage of correct guesses for treatment and control participants for each video. The reference line in red indicates the probability of guessing correctly based on chance alone. Plot adapted from Figure 2 of Köbis et al., 2021 and created in Stata.

Figure D3

Video Categorisation Accuracy by Condition and Video Type



Note: Comparison of percentage of correct guesses between treatment and control groups according to video type. The reference line in red indicates the probability of guessing correctly based on chance alone. Plot adapted from Figure 2 of Köbis et al., 2021 and created in Stata.

Appendix E

Regression Output

Table E1

Results of Linear Regressions using a Participant's Total Score as the Dependent Variable

Variable	<i>Total Score</i>		
	Estimate	SE	p-value
Treatment (T = 1)	0.397	0.367	0.280
Female	0.553	0.363	0.130
Age (1 x 10)	-0.338	0.186	0.071
White	0.486	0.478	0.311
UK	-0.748	0.463	0.108
Employed	-0.630	0.365	0.086
Student	0.418	0.408	0.307
Average Time on Videos	-0.004	0.010	0.700
Average Video Clicks	0.029	0.112	0.800
Aware of Deepfakes	0.075	0.762	0.921
Exposed to Deepfakes	0.386	0.370	0.298

Note. Table E1 shows the linear regression coefficients for the effect of various observed characteristics on a participant's total score, controlling for participant recruitment method. The total score is a continuous variable that measures how many of the 16 videos presented to participants during the experiment they accurately categorised as being either real or deepfakes. Treatment, Female, White, UK, Employed, Student, Aware of Deepfakes and Exposed to Deepfakes are binary variables taking on the value of 1 where that characteristic is observed. Age, Average Time on Videos and Average Video Clicks are continuous variables.

*** p<0.001, ** p<0.01, * p<0.05

Table E2

Results of Linear Regressions using a Participant's Deepfake Score as the Dependent Variable

Variable	<i>Deepfake Score</i>		
	Estimate	SE	p-value
Treatment (T = 1)	0.152	0.258	0.558
Female	0.578*	0.253	0.024
Age (1 x 10)	-0.146	0.132	0.268
White	1.023**	0.327	0.002
UK	-0.105	0.328	0.749
Employed	-0.099	0.259	0.702
Student	0.171	0.288	0.554
Average Time on Videos	-0.005	0.007	0.498
Average Video Clicks	0.046	0.079	0.564
Aware of Deepfakes	0.491	0.534	0.359
Exposed to Deepfakes	0.333	0.260	0.202

Note. Table E2 shows the linear regression coefficients for the effect of various observed characteristics on a participant's deepfake score, controlling for participant recruitment method. The deepfake score is a continuous variable that measures specifically how many of the 8 deepfake videos presented to participants during the experiment they accurately categorised as being either real or deepfakes. Treatment, Female, White, UK, Employed, Student, Aware of Deepfakes and Exposed to Deepfakes are binary variables taking on the value of 1 where that characteristic is observed. Age, Average Time on Videos and Average Video Clicks are continuous variables.

*** p<0.001, ** p<0.01, * p<0.05

Table E3

Results of Logistic Regression Predicting Video Categorisation Accuracy Using Individual Video Scores as the Dependent Variable

Variable	<i>Video Score</i>		
	Estimate	<i>SE</i>	<i>p-value</i>
condition	0.104	0.096	0.280
source	-0.114	0.119	0.336
Constant	0.466***	0.120	0.000

Note. Table E3 shows the logistic regression coefficients for predicting the likelihood of correctly categorising videos by condition, with video score (a binary variable where 1 = correct and 0 = incorrect) as the dependent variable and controlling for participant recruitment method. The analysis was conducted at the individual video level which means each video seen by each participant is treated as a separate observation in the model. Standard errors are clustered by participant.

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table E4

Results of Logistic Regression Predicting Video Categorisation Accuracy Using Individual Video Scores as the Dependent Variable

Variable	<i>Video Score</i>		
	Estimate	<i>SE</i>	<i>p-value</i>
Warned	0.080	0.146	0.584
condition	0.085	0.106	0.422
source	-0.117	0.112	0.336
Deepfake	-0.639***	0.121	0.000
Constant	0.477***	0.123	0.000

Note. Table E4 shows the logistic regression coefficients for predicting whether participants correctly categorised videos with video score (a binary variable where 1 = correct and 0 = incorrect) as the dependent variable and controlling for participant recruitment method and condition. The analysis was conducted at the individual video level which means each video seen by each participant is treated as a separate observation in the model. ‘Warned’ is a dummy variable indicating that a deepfake video in the treatment group had a content warning added to it. ‘Deepfake’ is a dummy variable indicating whether a video was a deepfake. Standard errors are clustered by participant .

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

Table E5

Results of Logistic Regression Predicting Video Categorisation Accuracy Using Individual Video Scores as the Dependent Variable

Variable	<i>Video Score</i>		
	Estimate	SE	p-value
Untagged	-0.126	0.143	0.377
condition	0.140	0.105	0.183
source	-0.117	0.121	0.336
Deepfake	-0.576***	0.095	0.000
Constant	0.476***	0.122	0.000

Note. Table E5 shows the logistic regression coefficients for predicting whether participants correctly categorised videos with video score (a binary variable where 1 = correct and 0 = incorrect) as the dependent variable and controlling for participant recruitment method and condition. The analysis was conducted at the individual video level which means each video seen by each participant is treated as a separate observation in the model. ‘Untagged’ is a dummy variable indicating that a deepfake video in the treatment group did not have a warning attached to it. ‘Deepfake’ is a dummy variable indicating whether a video was a deepfake. Standard errors are clustered by participant.

*** p<0.001, ** p<0.01, * p<0.05

Table E6

Results of Logistic Regression Predicting Likelihood of Categorising Videos as Real Using Individual Video Categorisation Decisions as the Dependent Variable

Variable	<i>Real Responses</i>		
	Estimate	SE	p-value
condition	0.024	0.086	0.783
source	0.032	0.100	0.753
Constant	0.252*	0.099	0.011

Note. Table E6 shows the logistic regression coefficients for predicting the likelihood of categorising a video as real (regardless as to actual video status) by treatment condition, with video categorisation (a binary variable where 1 = categorised real; 0 = categorised deepfake) as the dependent variable and controlling for participant recruitment method. The analysis was conducted at the individual video level which means each video seen by each participant is treated as a separate observation in the model. Standard errors are clustered by participant.

*** p<0.001, ** p<0.01, * p<0.05

Table E7

Results of Logistic Regression Predicting Likelihood of Categorising Videos as Real Using Individual Video Categorisation Decisions as the Dependent Variable

Variable	<i>Real Responses</i>		
	Estimate	SE	p-value
Warned	-0.195	0.147	0.186
condition	0.077	0.097	0.432
source	0.033	0.105	0.753
Deepfake	-0.823***	0.103	0.000
Constant	0.263*	0.103	0.011

Note. Table E8 shows the logistic regression coefficients for predicting the likelihood of categorising a video as real (regardless as to actual video status), with video categorisation (a binary variable where 1 = categorised real; 0 = categorised deepfake) as the dependent variable and controlling for participant recruitment method and condition. The analysis was conducted at the individual video level which means each video seen by each participant is treated as a separate observation in the model. ‘Warned’ is a dummy variable indicating that a deepfake video in the treatment group had a content warning added to it. ‘Deepfake’ is a dummy variable indicating whether a video was a deepfake. Standard errors are clustered by participant.

*** p<0.001, ** p<0.01, * p<0.05

Table E8

Results of Logistic Regression Predicting Likelihood of Categorising Videos as Real Using Individual Video Categorisation Decisions as the Dependent Variable

Variable	<i>Real Responses</i>		
	Estimate	SE	p-value
Untagged	0.011	0.147	0.942
condition	0.022	0.099	0.823
source	0.033	0.105	0.753
Deepfake	-0.872***	0.103	0.000
Constant	0.264*	0.104	0.011

Note. Table E9 shows the logistic regression coefficients for predicting the likelihood of categorising a video as real (regardless as to actual video status), with video categorisation (a binary variable where 1 = categorised real; 0 = categorised deepfake) as the dependent variable and controlling for participant recruitment method and condition. The analysis was conducted at the individual video level which means each video seen by each participant is treated as a separate observation in the model. ‘Untagged’ is a dummy variable indicating that a deepfake video in the treatment group did not have a warning attached to it. ‘Deepfake’ is a dummy variable indicating whether a video was a deepfake. Standard errors are clustered by participant.

*** p<0.001, ** p<0.01, * p<0.05

Table E9

Results of Logistic Regression Predicting Number of Real Videos Categorised as Real Using Individual Video Categorisation Decisions as the Dependent Variable

Variable	<i>Real Responses</i>		
	Estimate	SE	p-value
condition	0.140	0.137	0.308
source	-0.090	0.165	0.588
Constant	0.741***	0.000	0.011

Note. Table E7 shows the logistic regression coefficients for predicting the likelihood of categorising a real video as real (regardless as to actual video status) by treatment condition, with video categorisation (a binary variable where 1 = categorised real; 0 = categorised deepfake) as the dependent variable and controlling for participant recruitment method. The analysis was conducted at the individual video level which means each video seen by each participant is treated as a separate observation in the model. Standard errors are clustered by participant.

*** p<0.001, ** p<0.01, * p<0.05

Appendix F

Treatment Group Insights

Table F1

Treatment Group Insights on Awareness of Content Warnings

Warning noticed	Frequency	Percent
Yes	74	93.67
No	3	3.80
I don't remember	2	2.53
Total	79	100.00

Note. Table F1 summarises responses to the following question presented to $n = 79$ treatment group participants: “As you viewed the videos, did you notice any content warnings about the authenticity of a video?”

Table F2

Treatment Group Insights on the Effect of Warnings on Video Categorisation Decisions

Warning effect	Frequency	Percent
I took note of the warning and immediately categorised the video as a deepfake.	2	2.70
I took note of the warning, but I also used my own judgment when categorising the video.	47	63.51
I ignored the warning and solely relied on my own judgment when categorising the video.	25	33.78
Total	74	100.00

Note. Table F2 summarises responses to the following question presented to $n = 74$ treatment group participants who indicated they noticed content warnings during the experiment:

“Where a video had an authenticity warning, how did the presence of the content warning influence your decision to categorise that video?”

Table F3*Treatment Group Insights on Trust in Content Warnings*

Warning trust	Frequency	Percent
Completely	2	2.70
Mostly	5	6.76
Slightly	26	35.14
Somewhat	29	39.19
Not at all	12	16.22
Total	74	100.00

Note: Table F3 summarises responses to the following question presented to $n = 74$ treatment group participants who indicated they noticed content warnings during the experiment: “To what extent did you trust the content warnings provided about the authenticity of a video?”

Table F4*Treatment Group Insights on Videos Without Content Warnings*

Warning absence	Frequency	Percent
I felt the video was more likely to be real than a deepfake.	3	4.05
I considered the video could still be a deepfake despite the lack of a content warning	34	45.95
The absence of a content warning made no difference to how I perceived the video.	35	47.30
Other (free text response)	2	2.70
Total	74	100.00

Note: Table F4 summarises responses to the following question presented to $n = 74$ treatment group participants who indicated they noticed content warnings during the experiment: “When a video did NOT have a content warning, how did the lack of a warning influence your perception of the video's authenticity?”

Table F5*Treatment Group Insights on Videos Without Content Warnings*

Warning absence reasoning	Frequency	Percent
I assumed the video had been verified as real.	0	0
I assumed that a warning would only appear if a video was a deepfake.	0	0
I assume that video content is real unless I am told otherwise.	2	66.67
Videos without warnings appeared more realistic to me.	1	33.33
The lack of a warning made me less suspicious of the video's authenticity.	0	0
Total	3	100.00

Note: Table F5 summarises responses to the following question presented to $n = 3$ treatment group participants who indicated that videos that did not have content warnings seemed more likely to be real than deepfakes: “You have indicated that videos that did NOT have content warnings seemed more likely to be real than deepfakes. Please select the statement below that best explains your reasoning.”

Appendix G

Sample Debriefing Measures

Figure G1

Extract from Participant Debriefing Sheet Sharing Educational Resources

Access your score and further resources

Although the deepfakes used in this study were 'neutral' (i.e. there was no audio attached to the videos in which political ideologies or harmful content could be shared), it is still important to be aware of which videos used in the study were real and which were fake. You can view your score for the deepfake detection exercise and see which videos were real or fake by navigating to the next page.

You may also be interested in the following educational resources which provide more information on artificial intelligence and deepfake technology:

DetectDeepfakes: How to counteract misinformation created by AI

(MIT Media Lab: <https://www.media.mit.edu/projects/detect-fakes/overview/>)

Introduction to AI Guide

(UK Government: <https://www.gov.uk/government/publications/introduction-to-ai-with-a-focus-on-counter-fraud/introduction-to-ai-guide-with-a-focus-on-counter-fraud-html>)

Deepfakes and Audiovisual Disinformation

(UK Government: <https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-deepfakes-and-audiovisual-disinformation>)

Data Science and AI Glossary

(The Alan Turing Institute: <https://www.turing.ac.uk/news/data-science-and-ai-glossary>)

Figure G2

Sample Scorecard as part of Participant Debriefing Measures



THE LONDON SCHOOL
OF ECONOMICS AND
POLITICAL SCIENCE

Your Score

Your total score in the deepfake detection exercise is **9 out of 16**.
This means you correctly categorised 9 out of the 16 sixteen
videos as real or deepfakes.

Feedback on your performance for each video is provided below.
Note that the actual order of the videos below was randomised
during the experiment, meaning you will not have viewed videos
during the experiment in the same order they are presented
below.

Video 1 – Correct

You categorised this video as real. This video was REAL.



Figure G3

Sample Personalised Feedback as part of Participant Debriefing Measures

Video 13 – Incorrect

You categorised this video as real. This video was a DEEPPFAKE. In
this video, the subject's lips appear misaligned at times. If you
look carefully between time timestamps 0:04 to 0:05, you might
notice mild pixelation at the left cheek. The subject's hair also
moves in a way that appears unnatural.

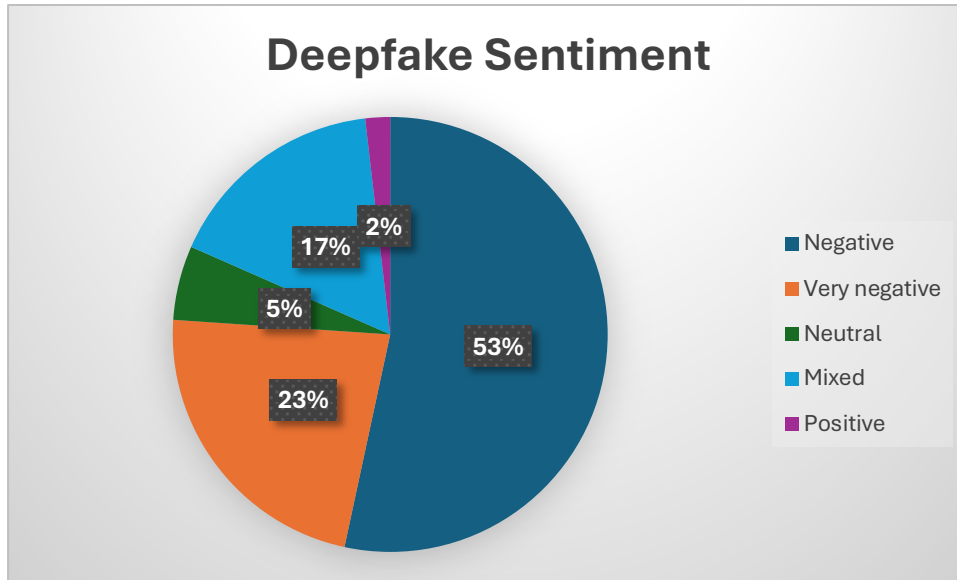


Appendix H

Deepfake Sentiment

Figure H1

Deepfake Sentiment Pie Chart



Note: Pie chart coding free-text responses according to sentiment scale, showing the percentage distribution of participant sentiment in response to the survey question “Please write one line of text that describes how you feel about deepfake technology. This could include concerns, benefits, personal experiences, or any other observations.” Free-text responses were optional meaning that not all participants provided a response. Participants could also have their concerns coded into more than one category, meaning total responses may exceed the number of unique participants. Pie chart created in Excel.

Table H1*Frequency Table of Deepfake Concerns*

Concern	Frequency	Example
Misinformation / fake news	19	<i>“I think that this is just one tool used by many to spread mis/disinformation”</i>
Privacy / consent	11	<i>“I think deepfake technology can bring nothing but trouble for humanity in general. Its potential benefits (such as sentimental value in recreating deceased loved ones or entertainment enrichment) are nothing but a cover up for its real, harmful purpose: exploiting unaware women and children for deepfaked porn...”</i>
Deception / impersonation	52	<i>“I think it’s extremely concerning for humanity as we lack critical thinking skills as a species and most will believe what they see without the thought of it being a fake”</i>
Crime / fraud	7	<i>“It could be used to scam people”</i>
Lack of regulation	4	<i>“It is a technology, if not regulated, can be disastrous in many ways.”</i>
Reputation	5	<i>“I think currently it’s a deeply concerning and unethical act as it violates privacy concerns and can defame an individual.”</i>
Potential for abuse	13	<i>“I think it’s impressive, but also scary and I’m afraid it can be used for bad things”</i>
Economy / society	8	<i>“I feel it could help make it easier to make content, but it also takes jobs away from a lot of people and can be an invasion of privacy if not done right.”</i>

Note: This frequency table groups sentiment-negative free-text responses sorted by concern type. Free-text responses were optional meaning that not all participants provided a response. Participants could also have their concerns coded into more than one category, meaning total responses may exceed the number of unique participants.