Deepfakes in Digital Discourse:
The Impact of Information Priming
and Truth Bias on Deception Detection

Kayleah Son

A THESIS

Submitted to the University of Michigan

Department of Communication and Media Studies

in partial fulfillment for the degree of

HONORS BACHELOR OF ARTS

April 2022

Advised by Professor Sitong Guo

Abstract

The amount of deepfake (DF) content online has proliferated rapidly: at the start of 2019, there were 7,964 deepfake videos on the web – and nine months later, that figure jumped to 14,678 (Ajder et al., 2019). A portmanteau of "deep learning" and "fake," deepfakes reference synthetically-constructed content through the manipulation of visual or auditory events, evoking a verisimilar reality – while increasing the efficiency of disinformation and causing palpable consequences as a result. The overall aim of this study is to determine whether simple priming of DF information significantly increases the ability to detect DF media and how truth bias influences this detection process. On this premise, this study addresses the following research questions: does priming with negative information on DFs increase user DF recognition while priming with positive information, on the other hand, abate the ability of users to recognize DFs? Additionally, does truth bias play a moderating role in the users' increased or decreased ability to detect DF after being primed, and does it have a direct influence on DF detection? Participants were placed in either 3 conditions (negative priming, positive priming, and control), watched a DF video of Tom Cruise, and evaluated their DF detection rate by asking respondents to determine whether the video was real or fake. Findings indicate that negative priming is most effective in participants' ability to detect DF, and truth bias has a significant moderating effect on the relationship between priming and DF detection. Ultimately, the study concludes that external priming effects and internal truth biases determine the effectiveness of DF material avoiding recognition. The current study also discusses the implications of the results, that the recognition of misleading material could profit from reasoned media literacy intervention.

Keywords: deepfakes, truth bias, priming, disinformation, deception detection

Acknowledgements

It brings me immense joy to see my thesis complete, but I could not have done it on my own without the countless individuals who have supported me throughout this process. I am indebted to each and every one of you.

First and foremost, I would like to thank my advisor Professor Sitong Guo. Despite her research emphasis being unrelated to my thesis, her dedication to challenging the boundaries of my research inspired me to continue to think beyond my limits. Whether responding to my frantic late-night emails about IRB proposals, or my general anxiety in tackling a thesis; Professor Guo has been the most supportive and encouraging mentor, and who is selflessly generous with her time. To Professor Guo; your expertise, insight, and patience are unparalleled. I am incredibly honored and lucky to have been able to grow under your tutelage during this process; this paper and I are both better off because of it. In all my future endeavors, I will always be able to point my passion for research back to you.

To Professor Scott Campbell, your passion for teaching and enthusiasm for discussions was infectious; thank you so much for filling our classes with your excited energy and uplifting spirit. Your desire for the challenge was the very push we needed to achieve more.

A special thank you to Professor Brian Weeks for your willingness to mentor and support my thesis despite recently returning from paternity leave. Your feedback and insight are the backbones of this paper.

To Professor Tony Bushner, despite our brief time together I found our chat (and your course) to be so valuable and integral in my work; you have truly inspired my academic interests and I have learned so much from your class. Your receptivity and ability to make everyone's voices feel valued were my source of confidence as I embarked on this journey.

To my family – especially my parents – for your unconditional love and positivity during this process. You were all undoubtedly my strongest anchors and my comfort during this time, I could not have asked for a better team to tackle this challenge with. Likewise, I want to thank you, Kaylene, for listening to my endless rants and for providing me with the comic relief I needed during stressful times.

Lastly, to my friends at Michigan, I am eternally grateful for your constant support for my thesis, but also beyond this campus. Seeing everyone's genuine interest and eagerness to help made me so grateful and honored to be your friend; you guys were my strength to get to where I am today, both in and out of the classroom. Happy graduation, and –

Forever Go Blue!

Contents

Table of Contents: Figures and Tables

List of Figures:

List of Tables:

CHAPTER 1:

INTRODUCTION

The proliferation of fake news emerges as one of the preeminent technological threats in this growing digital economy (Albright, 2017). Leveraging personal biases and emotions, misleading content has been adept at disproportionately impacting and manipulating susceptible populations of democratic societies (Chadwick & Vaccari, 2020). As a product of an increasingly noxious and inordinate spread of information, fake news has expanded its latest frontier to include AI-based cybercrime. Revolutionary on the technological front, University of Washington's Steven Seitz and his colleagues published an algorithm that enabled a high-quality, lip-syncing video in 2017. A few hours of rudimentary audio clips synthesized a photorealistic semblance of Obama speaking on topics including terrorism, fatherhood, and job creation (Suwajanakorn et al., 2017). Ostensibly bona fide, the primary premise of the video was that it was entirely fake – on purpose. Canonical in the emerging disinformation landscape, the advent of AI-supported, deceptive content has become increasingly difficult for humans and machines alike to distinguish as artificial, having severe consequences arising from the inability to recognize what is real and fake.

Enabled by AI technology called Generative Adversarial Networks (GANs; Goodfellow et al., 2016), Seitz and his colleagues hastened the dialogue around “deepfakes,” which are synthetic yet hyper-realistic audio and video content of events that never occurred. Derived from AI deep learning and neural architecture, an average individual has a predictable range of facial movements that correspond with the sounds made when forming specific words (Hancock & Bailenson, 2021). GANs utilize authentic video footage as a training set to create competition between two neural networks, likened to an arms race between deception and detection. Results are remarkably convincing; this avant-garde approach to deceit is no longer dependent on the quality of manipulated media, but on easy accessibility (Langguth et al., 2021). With consumer-grade software equipment, foundational technical skills, and a repository of available footage, nearly anyone can create synthesized videos in the modern day (Beridze & Butcher, 2019; Yang et al., 2019; Hall, 2018). Therefore, the capacity to create DFs is proceeding faster than the capacity to detect them.

Ranked as the most pervasive and epistemic AI crime threat in 2020 (Caldwell et al., 2020), DFs facilitate a variety of problems in a manner classified as the "newest way to commit the oldest crime[s]," (Spivak, 2019, p. 339), including blackmail, intimidation, and identity theft (Citron & Chesney, 2019; Citron & Chesney, 2019), ideological manipulation (Fletcher, 2018), and cyberbullying (Albahar & Almalki, 2019). With broader implications for trust (Fallis, 2020; Maras & Alexandrou, 2019), examining the forthcoming impact of deepfakes within digital discourse seems to be inevitable. The already-existing planes of disinformation, frequent and distinct throughout mainstream media, suggest the malicious prospect of deepfakes toward authentic, online discourse. Prior research reveals the persuasive power of images and video over text, and the comparatively fragile defenses users have against visual deception of high caliber (Newman et al., 2015; Stenberg, 2006). In a marketplace of ideas disrupted with deepfake videos, the truth will face difficulty emerging from the scrum. As this capacity spreads, individuals will increasingly encounter a dilemma: when faced with video or audio evidence of a newsworthy event, can its authenticity be trusted? Although the question in and of itself is not a novel inquiry, it will become harder to answer with the proliferation of deepfake media. Even more insidiously, as disinformation becomes more widespread and accessible, it is not far off to fathom the rate at which "fake" news will expand into "deep–fake" news.

Previous research on deepfakes denotes the consequences in which discourse was undermined by the dissemination of fake news (MacKenzie & Bhatt, 2020; Zannettou et al., 2019), micro-targeting (Zuiderveen Borgesius et al., 2018), and cyber subversion and information warfare (Paterson & Hanley, 2020; Polyakova & Boyer, 2018). With the unique efficacy of deepfakes in spreading disinformation, previous studies have raised critical concerns about the legitimacy of disinformation that can subjugate individuals to misinform their perceptions and exacerbate social divisions. In particular, after being exposed to a DF, citizens had lower trust and confidence toward the politician depicted in the deepfake, or toward the politician's party (Dobber et al., 2017). Similarly, other studies have revealed that without proper and extensive media literacy education, disinformation messages including a deepfake video resulted in greater "vividness, persuasiveness, credibility, and intent to share the message" (Hwang et al., 2021, p. 190).

Framed as a media literacy issue in addition to being established as a technological issue, there is a growing need to tackle the problem from an educational perspective, and equip citizens with the discernment necessary to question even the most realistic digital content. This current study extends this approach by analyzing how individuals respond to specific priming information on deepfakes and the user's ability to recognize deceptive content as a result. By emulating real-life situations and considering external factors in conjunction with critical thinking, this research allows for a new understanding of how caution can be implemented in parallel to the growth of deceptive media in online discourse. As both legislative and technological solutions have yet to match the speed of deception, it is the citizens' responsibility to be articulate within the digital ecosystem and learn how to decipher disinformation, in hopes that we may dismantle this infrastructure of misleading content together.

CHAPTER 2:

REVIEW OF CURRENT LITERATURE

2.1 Background

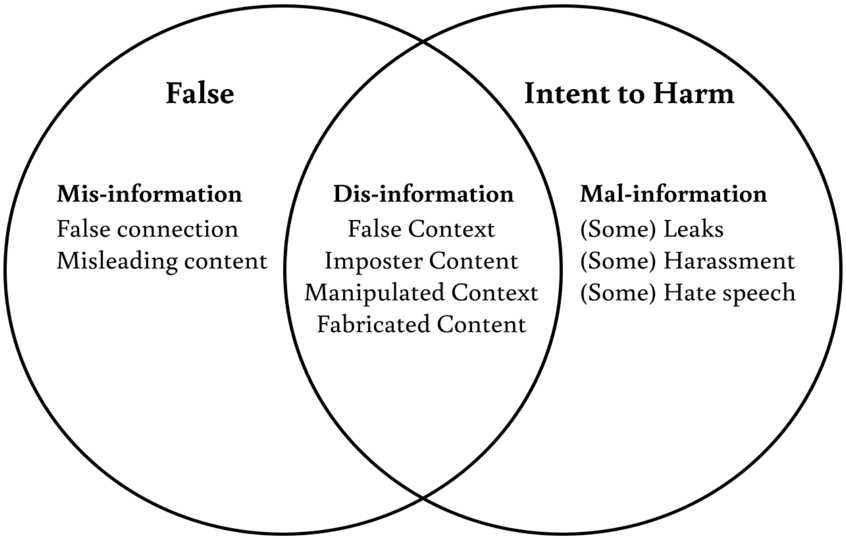
The “metacognitive experience” is oftentimes referenced when audiovisual content or images are simpler to process in contrast to written text. This term is used to describe the experientially derived feelings-toward-our-thinking that develop responses to menial tasks, including processing new information (Schwarz et al., 2007). A sub-experience, labeled “fluency,” is particularly integral in understanding the reasoning behind people believing in false information. According to Berinsky, individuals have a higher likelihood of accepting messages as true if the content is perceived to be familiar (2017). When the content or messages seem familiar, it elicits a “truthiness effect” – or in other words, a sense of “fluency” that allows the material to become easier to digest, and thus more compelling (Newman et al., 2015). With their technical ability to imitate realism, particularly when the video depicts well-known individuals, deepfake videos carry the potential to augment harmful content when fluency becomes generated through familiarity, irrespective of the authenticity of the video’s content.

2.1.1 Deepfakes as a Distinctive Form of Visual Disinformation

Within the context of information disorder, it is important to distinguish how false information conflates three categories: disinformation, misinformation, and malinformation (Wardle & Derakhshan, 2017), shown in Figure 1 below. Deepfakes, disseminated with the intention of deliberate harm to a person, social group, or society, originate under disinformation; which is bound by “imposter content,” manipulated content,” and “fabricated content” (Wardle & Derakhshan, 2017, p. 20). In other words, disinformation is defined as “intentional behavior that purposely misleads" (Chadwick et al., 2018, p. 4). In contrast to disinformation, misinformation is demarcated by the difference of intent – the latter implying the lack of deceiving nature (Jack, 2017), and malinformation is slightly distinct in the sense that although it has the intention to harm, it is based on semi-factual realities. However, with their cyclic nature, deepfakes have the potential to become misinformation if circulated online by those who mistakenly believe the content to be truthful. For the objective of this study, this aspect is not germane because we do not seek to study the factors contributing to the sharing of deepfake content.

Figure 1

Depicting misinformation, disinformation, and malinformation (First Draft in Asanov, 2019).



As an agent of visual disinformation, deepfakes encompass several dimensions of modality, including audio, video, & graphics; where the realism approach translates to "seeing is believing." The power of visual media to carry information is in line with how much signal is conveyed by a message. Due to the dominance of the visual system, videos have a high information-carrying capacity – therefore, there is a tendency for videos to become the reliable standard for truth. In a study by Appiah (2006), audio-visual testimonies were favored by participants over textual testimonies on a website; this favoritism was attributed to the visual vividness associated with audio-visual modality. The Modality Agency Interactivity Navigability Model (MAIN; Sundar, 2008) reinforces this concept, arguing that enriching the mode of presentation from text to audio-video provides realistic approximations of conveyed messages, consequently enhancing credibility visually. This framework implies that message receivers typically forego the presumed cognitive filters for assessing the veracity of the visual source and content. Supplementing the MAIN Model, the Limited Capacity Model also suggests that textual content tends to be processed more "systematically," since the receiver experiences cognitive overload when processing visual information (single modality text versus several modalities of audio, video, and graphics). In addition to the need to process the message, there are in-place structural components (e.g. animation, background) that are peripheral to the message itself (Lang, 2000). This depletion of cognitive capacity forces message receivers to depend, by default, on superficial cues that evoke heuristics when evaluating such content. In line with the previous Appiah (2006) study, according to the Heuristic-Systematic Model of persuasion (HSM; Eagly & Chaiken, 1993), there is an explanation behind this innate preference for heuristic processing: it "requires less cognitive effort and fewer cognitive resources than systematic processing" (p.327).

When organizing information heuristically, individuals make conclusions based on simplistic rules (e.g. long messages are credible, messages by trustworthy sources are reliable). Given the information ecosystem of the modern-day media environment, not only is it more difficult to systematically process messages, there is a shift toward heuristic processing when observing content (Metzger, Flanagin, & Medders, 2010). Furthermore, videos are effective at eliciting a more emotional than rational response (Nelson-Field et al., 2013). Given this framework within disinformation of trust and credibility, this study further considers the deceptive nature of deepfakes in the context of being primed and concerning an individual's truth bias, which corresponds to the significant research questions addressed in this study:

RQ$_1$: Does priming with a negative notion of DFs – in contrast to a positive notion or no prior notion – increase user recognition of DFs?

RQ$_2$: Does truth bias play a moderating role in the user's increased or decreased ability to detect DFs after being primed?

RQ$_3$: Does truth bias have a direct, negative effect on the user's ability of DF recognition?

2.2 Deepfake Detection and Accuracy

Based on earlier research exhibiting humans' visual processing abilities when it comes to facial recognition, there may be an expectation for humans to maintain that standard performance when it comes to identifying synthetic face manipulation in DF material. Visual neuroscience and perpetual psychology research demonstrated that the human visual system is armed with dedicated mechanisms for face perception (Sinha et al., 2006); particularly, there is a region of the brain solely trained to process faces (Kanwisher et al., 1997). Thus the human visual system is quicker and systematically more efficient at recognizing human faces than other objects, including objects with illusory faces (Keys et al., 2021). Whether an innate ability or learned expertise honed through experience, human visual recognition of faces seems to proceed holistically for a vast majority of the human population (Young & Burton, 2018; Richler &

Gauthier, 2014). However, compared to infants (Reid, 2017), adults are less cognizant when recognizing faces where the images include inverted or misaligned features (Richler et al., 2011; Rhodes et al., 1993; Yin, 1969).

Several studies have examined human accuracy in recognizing deception or disinformation, and there is a strong consensus in research echoing the inability of individuals to distinguish deception in interpersonal interactions (Burgoon et al., 1994; Kraut, 1980), let alone deepfakes. Rössler et al. (2018) found people correctly identifying deepfakes in only 50% of the total cases – which is statistically equivalent to random guessing. This data point was further aggravated by digital compression caused by online sites housing these deepfakes, resulting in ambiguity by smearing and blockiness (Vaccari & Chadwick, 2020). However, these findings primarily involve the resulting sum of truthful and deceptive messages when there is an equivalent distribution of truths and lies, and when considering factors of susceptibility, suspicion, and predictive utility (Miller & Stiff, 1993). In a more recent study comparing DF detection between human observers, computer vision deepfake detection models, and machine-informed human observers; results revealed that participants with access to the model's prediction were more accurate than those without, but inaccurate model predictions likewise decreased respondents' accuracy (Groh et al., 2022). In most cases, human participants in the study were hindered by manipulations designed to disrupt the visual processing of faces, suggesting a role for special cognitive capacities to explain human deepfake detection performance.

2.2.1 Priming and Deepfake Recognition

Priming research corroborates substantial evidence that individuals are indeed primed by messages salient in the media. By foregrounding and frequently sharing specific material, whether incidentally or through deliberate coverage, studies have demonstrated that exposing individuals to intentional cues can subtly influence, or prime, their conscious and subconscious responses (Molden, 2014). Likewise, such research has also indicated that after processing information, individuals develop "activation tags" that connect concepts in the mind (Collins & Loftus, 1975, p. 409). According to Tversky & Kahneman (1973, p. 208), these activation tags are easily accessible and impact the way subsequent information is evaluated since they remain at the "top of the head." Specifically, individuals are "primed" when the information is shared, stored at their immediate disposal, and recalled when encoding and evaluating subsequent information (Chawarski, 1996). Moreover, priming is distinct compared to other theories involving cognitive media effects, including framing, in that it focuses on the prominence of information, rather than focusing on the structure of the information and how it is consequently processed (Moy, Tewksbury, & Rinke, 2016; Chong & Druckman, 2007; Scheufele, 2000). Priming effects are known to transfer easily to relevant topics (Petty & Jarvis, 1996), and thus it is highly likely that such effects apply to evaluations of deceptive media in general.

Inaccurate information is highly difficult to discern (Flynn et al., 2017); therefore several legs of research have investigated approaches that may lessen susceptibility to falsehoods, particularly fake news. Pennycook et al. (2019)'s work investigated priming as a factor to minimize an individual's predisposition to believing false information. Two groups of participants were either exposed to headlines that were labeled as "disputed," or headlines that omitted the presence of a "disputed" tag. Findings suggested that the caution label moderately reduced the susceptibility of participants to fake news. Meanwhile, separate studies by Clayton et al. (2019) and Ecker et al. (2010) respectively evaluated similar priming methods; headlines were attached with general ("disputed") and specific ("rated false") tags to news headlines, along with providing respondents with general and specific warning instructions on the perceived credibility of fake news stories. Results for both studies revealed that both general and specific tags reduced vulnerability to fake news, specific tags were more effective in reducing susceptibility compared to general warnings; however, the extent to which the priming was effective in both cases were small and not completely effective. Most recently, Selezneva (2021) examined whether a caution, either general or specific, would prime participants to be broadly critical of both false and true news claims. The study indicated that both priming cautions did not have significant correlations with lower receptivity to fake news. Rather, the conclusion emphasized that the most important predictor for deception detection was an individual's analytical ability.

Specific to the realm of deepfakes, Iacobucci (2021) tested whether simple priming of DF information, in contrast to a control condition, would significantly increase an individual's ability to recognize DF media. Results exhibited that there was no extensive difference between the prime caution and control, but rather a discussion around countering deceitfulness of DFs through an educational standpoint – more so effective for those who naturally have a lower susceptibility to believe willfully misleading claims. Considering literature between priming and deception material expanding into deepfakes, this study proposes the following hypothesis:

H_1: Priming participants with a negative notion of DF will lead to greater DF recognition compared with the positive (primed with the positive notion of DF) and control (not primed with the notion of DF) groups.

2.3 The Role of Truth Bias in Deception Content Believability

In a meta-analysis by Levine et al. (1999), the study argued on the incompleteness of deception detection without the prime determinant, coined "the veracity effect." The study revealed that with the assumption of individuals being more often truth-biased, the level of presumed veracity in a given message or content was an important factor – resulting in detection accuracy to be substantially above 50% for truths, but well below 50% for lies. Given this specific adaptation, the “message veracity” of deepfakes are inherently considered “lies,” which allows us to advance with the premise that detection accuracy for such content will be relatively low. There are several reasonable explanations for this poor detection performance across various forms of content: when a receiver interacts with a message, they experience verbal and nonverbal cues of deception that they isolated from the sender (Burgoon, Buller, & Woodall, 1989; Miller & Stiff, 1993), and within this subset of behaviors there is no stable correlation that infallibly discerns deception from truth-telling (Kraut, 1980). Therefore, flawless deception detection accuracy is not possible. However, research indicates that veracity judgments are affected by systematic encounters with errors and biases. Rather than proactively evaluating the message or message agent for deceit, receivers rely on cognitive heuristics to compensate for clarity.

2.3.1 Truth Default Theory

Within the bounds of deception and detection, the Truth-Default Theory posits a passive acceptance of honesty when interacting with others (TDT; Levine, 2014). Under this framework, individuals tend to operate on a default presumption that the other person is truthful. Doing otherwise (such as consciously considering the possibility of deceit) requires some form of prompting or stimulus to propel a person out of their “truth-default” state. This theoretical structure stems from the earlier idea that individuals are typically “truth-biased” (McCornack & Parks, 1986; Zuckerman et al., 1981); this presumption functions under the objective that by assuming veracity, communication and cooperation will reach peak efficacy, and will almost always lead to correctness if honesty is maintained (Levain & Clare, 2014). However, this blind acceptance of truth occasionally leaves individuals vulnerable to deceit.

The existence of a truth-default has also been asserted by the previous works of Gilbert (1991) and Grice (1989) in the areas of mental representations of information and social linguistics, respectively. Gilbert (1991) considered several models of individuals mentally constituting true and false information. His Spinozan model implied that the most basic comprehensive of received information presupposes an initial belief, and thereafter some portions of that information are subsequently and actively unbelieved. As a result, Gilbert proposed truth-default not only under the extremities of incoming communication but generally for all types of information that is received through human cognition. Correspondingly, Grice (1989) advocated a system to make sense of what others say, even in situations where people do not exactly say what they mean. Grice suggested that the mere understanding of what is meant necessitates the presumption that the message is by default, true. This outlook of communicating in good faith not only allows individuals to make sense of interactions that would otherwise be incomprehensible or ambiguous but also makes them vulnerable to duplicity (McCornack & Levine, 1992). Although Grice's mutual principle and the truth-default are not nearly as identical, they are functionally similar in their implications. Collectively, each framework contributes to the tendency that initial belief is intrinsic and automatic, whereas disbelief requires deliberate and conscious mental effort.

Despite this general attention, however, there is a shortage of studies focused on exploring this latter concept thoroughly, and even the most notable of those have seldom examined the concept from a new, sociotechnical perspective. This study aims to quantify the social impact of DFs, specifically on the role of priming and truth bias on individuals attending to DF content. Perhaps feeling more uncertainties than being misled by DFs, the primary aim of this research is to assess whether the types of priming can augment deception detection, but more importantly, whether perceptions of truth and falsity are affected by individuals' preconceived levels of truth bias. Finally, the study considers whether truth bias is also a formidable variable that elicits a reluctance to recognize DFs as misleading content.

2.3.2 Truth Bias

The existence of truth bias in deception detection literature is a commonly accepted form of bias, and one of the most well-documented findings in research (Bond & DePaulo, 2006; McCornack & Parks, 1986). The truth bias refers to the tendency “to judge more messages as truths than lies,” or the uncritically accepting belief of honesty, independent of the actual honesty (Anderson, Ansfield, & DePaulo, 1997, p. 23). Truth bias also signifies that people's accuracy at detecting truths are greater than their accuracy in detecting lies (DePaulo et al., 1997; Zuckerman, DePaulo, & Rosenthal, 1981). Across the literature, the idea of a truth bias suggests that when a trigger stimulus is inert, individuals do not explicitly consider the veracity of a message. Likewise, in experiments that concentrate on detecting deception, truth bias pervades participants who are required to make truth-lie judgments (Bond & DePaulo, 2006; Levine et al., 1999). For example, in the Bond & DePaulo (2006) meta-analysis, a sizable majority (72.4%) of prior experiments found that subjects reported more than 50% of observed messages as truthful. Similarly, Vrij (2000) found that the mean accuracy rate was 67% for detecting truths, and 44% for detecting lies (with the chance accuracy being 50%).

Occasionally, deception detection scholars purposefully conceal the purpose of research and place veracity assessments as filler units. In McCornack & Levine (1990), they experimented with a low-suspicion condition; results indicated that truth bias increased from

64% under conditions of explicit prompting to 80% when such stimuli were less active. The finding demonstrated that with less explicit triggers, truth bias has a greater conspicuous presence. A potential explanation for these findings is the availability heuristic that relates to the “availability” of truthful and deceptive behavior (O’Sullivan, Ekman, & Friesen, 1988); in everyday life, people are more likely to interact with truthful statements than deceptive statements. Therefore, this lack of daily deceptive statements allows individuals to adopt a bias that true statements are more likely to occur, and judge others as being truthful. Another explanation is the phenomenon of social conversion rules that inhibit individuals from becoming suspicious of others (Vrij, 2000). Due to social culture that discourages distrust in others, people may have a tendency to assume things as being true rather than untrue.

It is also important to note that when detecting lies, people are guided by the plausibility, consistency, repeatability, and types of statements made (Vrij, 2000; Vrij et al., 1999; Bell & Loftus, 1988; Wells & Leippe, 1981). Individuals gravitate toward statements that sound plausible, are internally consistent, frequently mentioned, and contain more details that are true. Most statements likely fulfill most of these criteria and are therefore judged as truths. Since the detection of truth is intrinsically higher for individuals based on previous studies (Levine et al., 1999; Miller & Stiff, 1993), people are more likely to ascribe truth to messages, rather than deceit (Buller & Burgoon, 1996; Burgoon et al., 1994). This bias is particularly present within “conditions of high contextual suspicion” (McCornack & Levine, 1990). Therefore, truth bias plays a prominent role in how people perceive deceptive messages, or rely on cognitive heuristics (Stiff et al., 1992).

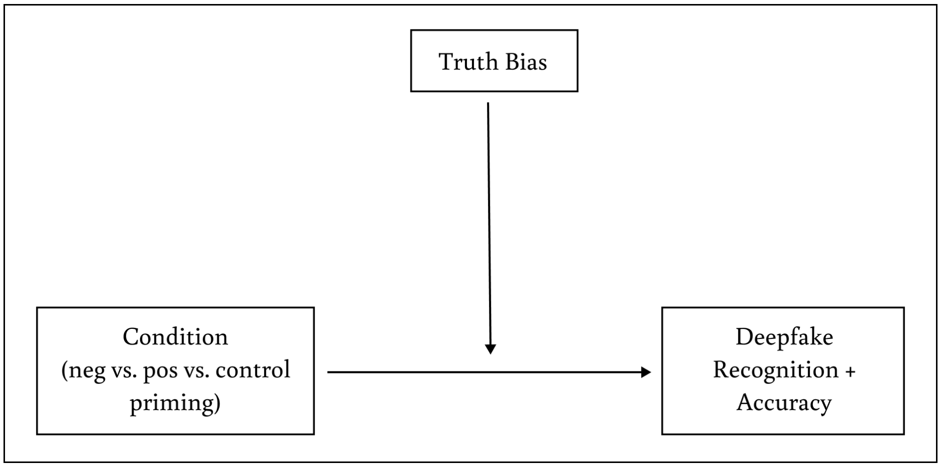
In the domain of DFs, simple priming correlated positively with recognizing deceptive content, and truth biases negatively with recognizing deceptive content. We expect the same process to occur within more specific conditions of simple priming (positive vs. negative), while also taking into consideration that individuals with higher truth bias will apply less cognitive processing compared to those with positive priming and higher truth bias in assessing the veracity of DF material. As depicted in Figure 1, this study predicts a two-way interaction where negative priming DF information should have a higher positive effect on DF recognition for individuals with truth bias, than for those experiencing positive priming with truth bias. With the overarching framework of truth bias that situates individuals as natural agents of truth in deceptive environments, this study proposes the following hypothesis:

H_2: Truth bias (TB) moderates the relationship between priming and DF recognition, such that the less individuals demonstrate TB, the more they will be able to recognize DF after being negatively primed, and vice versa.

H_3: Truth bias (TB) has a separate effect on user ability to detect DF material, the higher an individual's truth bias, the more positive attitude toward DF content, and decreased ability to recognize DF content.

Figure 2

The Hypothesized Moderation Model.


Truth Bias
Condition (neg vs. pos vs. control priming)
Deepfake Recognition + Accuracy

CHAPTER 3:

METHODOLOGY

3.1 Design

To research these theories, an online survey was conducted through Qualtrics and published using the Amazon Mechanical Turk platform.

Centered around DF-specific media literacy, the research adopted a single factor between-participant design in which priming with information on DFs was manipulated at three levels (conditions: negative priming vs. positive priming vs. no priming). In line with Iacobucci et al.'s (2021) priming experiment, this study took a step further and analyzed the specific differences within priming in itself, largely between negative and positive groups. Recognition containing these three treatments were expressed in the form of a short excerpt that either: exposed the negative effects of deepfakes (negative priming condition), elaborated on the positive consequences of GAN technology (positive priming condition), or omitted the excerpt completely (control condition). The order of the articles was randomly determined but was equally distributed across participants through a Qualtrics randomization feature.

3.1.1 Participants

Respondents were recruited using the Amazon Mechanical Turk (MTurk) platform; only those older than 18 years old were allowed to participate, and participants provided informed consent. 61.9% of respondents reported their age on a grouped average of 26-39 years old (n = 130, *M* = 27.8, *SD* = 0.67). Demographic data showcased 64.8% (n = 136) of participants identified as male, 34.3% female (n = 72), and 1% non-binary/other (n = 2). Participants were evaluated on their prior knowledge regarding deepfakes, resulting with a mean score of 4.08 on a

5-point scale (*M* = 4.08, *SD* = 1.07). However, a meta-analysis reported that age, education level, work experience, and sex have little impact on deception detection (Aamodt & Custer, 2006).

Additionally, a pre-screening question was included that assessed the types of social media platforms respondents interacted with the most; this question allowed for us to understand the context in which types of content participants interacted the most (e.g. Facebook algorithmic timeline versus Snapchat conversations). Results indicated that the most popular platforms were Instagram (22.59%), Facebook (21.39%), and Youtube (16.44%), respectively. Recruitment was restricted to randomized individuals in an online participant pool to minimize the possibility that they have previously engaged in deception detection research, coursework, or have been exposed to deception research content in their everyday lives. Data was collected from April 9-11, 2022, and an additional sample of only positive and negative priming participants was collected from April 13-15, 2022.

3.2 Procedure

All participants began the survey with a brief, high-level introduction to deepfakes through a short description:

> **For background**: "Deepfakes"—a hybrid of the terms "deep learning" and "fake"—are videos, photos, or audio recordings that have been modified to make false content appear real. Deep learning algorithm technology can replace faces and speech to make it appear as if someone said or did something that never happened. The most sophisticated deepfake videos require thousands of images to train algorithms to recognize and then manipulate a face.

Regardless of their comfortability with the concept, the brief explanation provided context that either reaffirmed or introduced the topic to participants. Respondents were then randomly branched into three treatments: negative priming, positive priming, and a control group. Negative group assignments were directed to an excerpt of the article, "Responding to Deepfakes and

Disinformation" by *The Regulatory Review*, a publication under the University of Pennsylvania Law School, shown in Appendix A. Published in 2021, the article assessed how regulation may encounter problems with deepfakes and disinformation. The excerpt contained several issues outlined by legal experts that spoke negatively about the future development of deepfakes. Positive group assignments read an excerpt in an article by ThinkAutomation, a business process automation startup in the UK, shown in Appendix B. Titled "Yes, Positive Deepfake Examples Exist," the piece organized the plethora of benefits across industries that were impacted positively by the creation and application of deepfakes: education, language, entertainment, art, and medicine. After reading the article, participants were required to answer a question about their respective content that corroborated that they had fully read the excerpt, shown in Appendix C. Control group assignments were not shown an article and were omitted from reading any content or answering questions about the respective content.

Respondents were then asked to rate their attitude toward DF technology with a six-item, 5-point semantic differential scale that included the following traits: bad/good, unpleasant/pleasant, helpless/helpful, boring/interesting, negative/positive, foolish/insightful. Participants were also asked to rate their perception of celebrity actor, Tom Cruise, through a 5-point semantic differential scale (negative/positive) to determine their level of truth bias. Evaluating the respondents' attitudes post-article and pre-DF video (containing Tom Cruise) is also important to understand any changes in or effects of trust bias after engaging with the DF video and priming treatment.

Prior to watching the video, participants were informed of its' alleged purpose: "The following video is of Tom Cruise, who had recorded a series of TikToks for an ad campaign that will be launching in the next year." In the survey, the video was treated and labeled as legitimate content in order to mimic the deceptive nature of deepfakes on the Internet. The video itself was uploaded by the Youtube channel, “Vecanoi,” which depicted VFX/A.I. artist Chris Ume as Tom Cruise. In a series of short clips uploaded as a compilation, the DF video (1 minute, 37 seconds) had a seemingly realistic Tom Cruise (emulated by Chris Ume’s Tom Cruise impression and DF software) doing mundane tasks while talking to the camera, such as slipping on a wet floor or playing golf. The video algorithm for Ume’s fake face used over 13,000 images of Cruise that captured him from almost every angle. Included in Appendix D is a screenshot of the advanced DF algorithm that mirrored Cruise’s features onto Ume’s face in the Youtube video.

After watching the video, DF recognition is measured with a single question: “Tom Cruise had actively participated in this ad series, and not due to digital video editing technologies.” Participants were prompted to answer this question with 5-point, Likert-type items (1 = Not At All; 5 = Absolutely). Perceptions of Tom Cruise were also re-evaluated through a 5-point semantic differential scale (negative/positive) in order to compare whether the DF video coupled with the respective priming conditions had elicited any change in trusting attitudes and biases toward the actor.

3.2.1 Data Fabrication

The survey took roughly 11.2 minutes to complete, and participants were compensated $0.80 for their time. Per reliability of data, the total intended sample population was 250 responses, with 80 participants per treatment group. Despite completing the survey with 405 attempted responses, 210 responses were ultimately validated and collected through the Amazon Mechanical Turk platform due to 187 invalid and 8 fraudulent survey respondents.

As mentioned in the survey layout prior, the questionnaire included two authorization features intended to eliminate those who did not: 1) comprehensively read the article excerpts, or

2) authentically fill out the survey. Since the primary objective of this study was to cross-examine the effect of the different priming conditions on the ability to detect DF material, respondents needed to have thoroughly read and understood their respective article excerpts utilized in the survey.

For participants randomized to read a positive or negative excerpt, they were initially greeted with the definition of a deepfake (aforementioned earlier in the survey protocol), alongside a clarification note that notified participants of the importance of reading the excerpt, and a heads-up of an evaluative question:

> **For the primary assignment, you will be asked to read an article and answer a follow-up question to ensure you read the excerpt. Please read thoroughly and answer correctly to the best of your abilities to be able to proceed to the rest of the survey and receive credit on M-Turk.**

For those who continued with the survey, after reading their randomized excerpt, participants were asked a single question that evaluated their cognizance to avoid instances such as skimming content. Negative priming participants were asked “What is NOT a problem mentioned in the article on the consequences of deepfake technology?” while positive priming participants were asked “What is NOT a positive application mentioned in the article on the benefits of deepfake technology?” Among the four available answers, three of which were short sentences copied from the excerpt itself, one answer was a completely false and contrived negative consequence or positive benefit. Those who were able to determine the misleading answer were allowed to proceed with the rest of the survey, while those who answered incorrectly were immediately taken to the end of the survey, and their submissions null. For those in the control group, since no excerpt was needed, there was no required question to check their understanding to complete the survey.

Additionally, the questionnaire embedded a message that was only visible to those who had completely answered all the mandatory questions, and correctly answered the article authentication question (if assigned). This message shared a validation code ("DEEPFAKE2022") for respondents to enter in the Amazon Mechanical Turk platform to receive credit for the completed survey. However, those that did not enter the correct code were flagged as fraudulent answers, and thus removed from the participant pool. It is important to note that this participant removal protocol bargained with the evenness of responses among the three treatment groups, however surprisingly, of 210 respondents, 61 were exposed to condition 1 (negative priming), 60 were exposed to condition 2 (positive priming), and 89 were exposed to condition 3 (control).

CHAPTER 4:

RESULTS & ANALYSIS

4.1 H_1: Negative Priming Increasing User DF Recognition

Hypothesis 1 predicted that priming participants with the negative notion of DF will lead to greater DF recognition compared with positive (primed with the positive notion of DF) and control (not primed) groups. For this initial hypothesis, a univariate analysis of variance (ANOVA) indicated that priming did play an influential role in a user's ability to detect DF, and among the types of priming those engaging with the negative priming conditions recognized DFs the most. As shown in Table 1, results indicated that there was a statistically significant difference between groups within priming and DF detection (F(2,207) = 3.47, *p* = .033).

Table 1

One-Way ANOVA of DF Detection and Priming

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	8.517	2	4.259	3.469	.033
Within Groups	254.078	207	1.227		
Total	262.595	209			

This information was revealed by a Tukey test conducted as a part of a post hoc test of multiple comparisons between conditions. As seen in Table 2, the analysis revealed that the negative priming group had more successful detection rates than the control group that received no priming material (*p* = .036). Thus, hypothesis 1 was supported by the data. However, there was no statistically significant difference between the negative priming and positive priming groups (*p* = .809) or between the positive priming and control groups (*p* = .172).

Table 2

Multiple Comparisons Between Priming Conditions

(I) Priming	(J) Priming	Mean Difference (I-J)	Std. Error	Sig	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	-.125	.201	.809	-.60	.35
	3	-.458*	.184	.036	-.89	-.02
2	1	.125	.201	.809	-.35	.60
	3	-.333	.185	.172	-.77	.10
3	1	.458*	.184	.036	.02	.89
	2	.333	.185	.172	-.10	.77

Note: The mean difference is significant at the 0.05 level.

4.2 H_2: Truth Bias as a Moderating Variable Between Priming and DF Recognition

Hypothesis 2 speculated that truth bias moderated the relationship between priming and DF recognition, such that the fewer individuals demonstrate TB, the more they will be able to recognize DF after being negatively primed, and vice versa. As shown in Table 3, a two-way ANOVA determined a statistically significant relationship between priming and truth bias (F(7,209) = 3.030, *p* = .005). The significance indicates that truth bias has a moderating effect on the relationship between priming and DF detection; therefore Hypothesis 2 is supported by the data.

Table 3

Two-way ANOVA of Truth Bias Between Priming and DF Detection

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	87.299*	13	6.715	7.508	<.001
Intercept	637.948	1	637.948	713.296	<.001
Priming	11.458	2	5.729	6.405	.002
Tom_Cruise_TB	39.875	4	9.969	11.146	<.001
Priming*Tom_Cruise_TB	18.968	7	2.710	3.030	.005
Error	175.296	196	.894		
Total	3197.000	210			
Corrected Total	262.595	209			

Note: R Squared = .322 (Adjusted R Squared = .288).

4.3 H_3: Truth Bias Decreasing User DF Recognition

Hypothesis 3 anticipated that truth bias (TB) would have a separate effect on user ability to detect DF material, where the higher an individual's truth bias, the more positive attitude toward DF content, and therefore decreased ability to recognize DF content. Shown in Table 4, a bivariate correlation calculated a strong, positive relationship between DF detection and truth bias, where the more truth bias an individual had (higher positive attitude toward Tom Cruise), the more frequent these respondents believed the DF video to be true (r = .481). Additionally, Table 4 reveals that this correlation is highly statistically significant (*p* = <.001), supporting Hypothesis 3.

Table 4

Pearson Correlation Coefficient of DF Detection and Truth Bias

		DF Detection	Truth Bias
DF Detection	Pearson Correlation	1	.481*
	Sig. (2-tailed)		<.001
	N	210	210
Truth Bias	Pearson Correlation	.481	1
	Sig. (2-tailed)	<.001	
	N	210	210

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

CHAPTER 5:

DISCUSSION

5.1 Discussions, Limitations, and Future Research

First, the overall findings of this study contributed to the assumption that anti-deception strategies should consider users' prior knowledge, along with education and crowd literacy on the existence of manipulated online material. On this premise, results on H1 (the relationship between priming and DF recognition is statistically significant, specifically negative priming conditions performed the best in DF recognition) confirmed that interacting with negative information on DF content primed users' ability to successfully identify such content, compared to groups that did not receive negative information priming. Similar to the function of fake news in Procedural News Knowledge (PNK), this process of negative information priming served as a subtle form of inoculation, creating a segment of "umbrella protection" against the threat of being deceived by DFs (Jones-Jang, Mortensen, & Liu, 2021). This protection effect allows users to arm themselves with the necessary information and skills to favorably recognize misleading content, which aligns with the functions of Amazeen and Bucy's PNK theory (2019).

Second, findings report that truth bias not only moderated the relationship between priming and DF detection, but also independently influenced users' ability to recognize deceptive content. Although it is not a necessary or sufficient condition for ideal DF recognition, negative priming aided recognition abilities that may have overridden those who had higher levels of truth bias. On this matter, this study investigated the extent to which truth bias may hamper DF detection, and the findings revealed a positive correlation between higher truth bias and a higher rate of believing DF content. More importantly, the study recorded interaction between simple priming and truth bias: negative information priming about DFs enhanced a participants' ability to recognize DF deception, but primarily for those individuals with already low levels of truth bias. Therefore, the study confirmed that negative information priming elicited higher levels of DF detection, and this detection stimulated resistance to persuasion, cooperating with the PKM model. The results highlight how priming with a negative perception of disinformation technology aids in developing a reluctance to accept and content with misleading content. Moreover, consistent with the strong correlation between negative priming and lower truth bias posited by the TDT, when engaging with negative content, users develop a more negative attitude toward the deceptive material, therefore allowing recognition to perform at a successful level.

Although hopeful in the realm of deception detection, this study is not exempt from limitations: the primary obstacle was attributed to the study's insufficient sample size due to data fabrication. Out of 405 potential participants, only 210 were determined as valid to be included for statistical analysis. Ultimately, the reduced viable results were limited to 61 and 60 respondents in the negative and positive groups respectively, versus 89 respondents in the control group. Though the amount is ample for this study's objective, the sample size is still relatively small to confidently confirm significant relationships within the data set, especially in determining how (beyond simply whether) truth bias is indeed a moderating variable between priming and DF detection. Building the study on a larger sample size could have yielded more accurate results in answer to the "how" aspect of this result. A further sample-related limitation is in regards to gender; male participants represented 64.8% of the sample, therefore further investigation – in addition to gender-balanced samples, but also with the objective of analyzing the effects of gender on DF recognition – is beneficial, in light of prior studies that point out that women are more likely to evaluate deceptive content as more truthful than are men (Wu et al., 2020; Buchan, Croson, & Solnick, 2008).

In retrospect, there were also several limitations encountered concerning the method survey material. First, the current study omitted a specific measurement that would have allowed the evaluation of whether the length of the DF material played a role in DF detection (Elaad, 2010; Masip, Garrido, & Herrero, 2009). After completing the findings, not differentiating the DF video (short and long) in the survey inhibited the study's ability to conduct a thorough analysis of the results, and whether exposure time to deceptive material is another variable that affects DF recognition. Second, participants were exposed to a brief and general introduction to DFs, in place of a specific refutational-same inoculation message. Future studies should assess whether dissimilar, in-depth inoculation methods (rather than types of priming) can more usefully immunize individuals who have higher levels of truth bias in trusting false content (Roozenbeek & Van Der Linden, 2019). Moreover, an extension of inoculation strategy research may include disclosure characteristics (Amazeen & Wojdynski, 2018; Evans & Park, 2015), where such studies may build upon disclosure-related topics in developing labels that signal misinformative material.

Additionally, although prior research has exposed the accuracy of detecting truth-lie statements to increase and truth bias to decrease over time, such phenomena have yet to be tested with DF content. Future research should consider involving a method of evaluating exposure length in deception detection studies. With DF being a relatively new frontier on the disinformation terrain, current studies on deception detection prioritize fake news and misinformation on social media platforms and less on DF technology. As is often the case in experimental research, the lack of DF-related research in social sciences and, specifically, from the perspective of media psychology, effectiveness, and its social impact, should be attempted in future research. Furthermore, the survey disregarded the impact of source credibility when interacting with DF material – the video itself was hosted on Vimeo, but an indication should have been embedded into a tailored Facebook newsfeed from a specific source to induce an environment most natural to participants interacting with DFs in real-life. Future research should examine source credibility as a determinant of deception detection.

CHAPTER 6:

CONCLUSION

To understand the role of disinformation within a contemporary, tech-saturated environment, it is important to also consider the social implications of its prominence in digital discourse. Sociotechnical infrastructure values and benefits from the blend of artificial intelligence and crowd wisdom, however, content moderation through supportive tools are necessary to carefully design a thoughtful and valuable space for online conversations (Hwang, Ryu, & Jeong, 2021; Pennycook et al., 2021). In detecting real-world deepfakes, results have demonstrated that deep-learning architectures are reliable at distinguishing between real and fake images; however, recognition of minimal inaccuracies remains a critical aspect that differentiates algorithms from human cognition (Taeb & Chi, 2022). Moreover, as much as technical tools can aid accurate identification, crowd wisdom can become augmented with more explainable and supportive AI.

This study focused on the primary role that disinformation, through rising mediums that included DFs, plays in digital discourse. Although the effectiveness of DFs is studied across many disciplines, the strategy utilized to understand the extent to which information priming (negative, positive, or control conditions) can provide ample cognitive protection in recognizing false information. Particularly within the case of DFs – where deep-learning architectures are becoming increasingly deceptive and realistic – recognizing how to best arm the digital demographic with discernment will be beneficial for discourse as it proliferates beyond the physical realm. As much as textual cues aid in DF detection, ultimately, identifying authentic videos involves much more than visual processing – human context, their perception of the world, and the ability of critical reasoning amalgamate to provide a cohesive infrastructure to fight disinformation. Most importantly, an inability to recognize falsehoods and the truth poses dire consequences to the extent that it impedes communities and the information ecosystem with the ability to develop trust and accurate discourse (Tsafi et al., 2020; Allen et al., 2020; Sundar, Molina, & Cho, 2021). Overall, this study was meant to highlight that DFs should be regarded as a clandestine threat to online users, particularly by encouraging scholars to further explore users' prior knowledge on the issue or utilizing priming techniques with the knowledge to strategize against DF and its negative effects.

References

Aamodt, M. G., & Custer, H. (2006). Who can best catch a liar. Forensic Examiner, 15:6–11.

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

Albahar, M., & Almalki, J. (2019). Deepfakes: threats and counter- measures systematic review. Journal of Theoretical and Applied Information Technology, 97:3242–3250.

Albright, J. Welcome to the era of fake news. (2017). Media and Communication, 5:87–89.

Allen, J., Howland, B., Mobius, M., Rothschild, D., Watts, D. J. (2020). Evaluating the fake news problem at the scale of the information ecosystem. Sci. Adv. 6.

Amazeen, M.A., & Bucy, E.P. (2019). Conferring resistance to digital disinformation: the inoculating influence of procedural news knowledge. Journal of Broadcasting & Electronic Media. 63:415–432.

Amazeen, M.A., & Wojdynski, B.W. (2018). Reducing native advertising deception: revisiting the antecedents and consequences of persuasion knowledge in digital news contexts. Mass Communication and Society. 22:222–247.

Anderson, D.E., Ansfield, M.E., & DePaulo, B.M. (1997). The Accuracy-Confidence Correlation in the Detection of Deception. Sage Journals.

Appiah, O. (2006). Rich media, poor media: The impact of audio/video vs. text/picture testimonial ads on browsers' evaluations of commercial web sites and online products. Journal of Current Issues & Research in Advertising, 28(1), 73–86.

Asanov, T. (2019). "Fake News in Modern News Media: Disinformation, Misinformation and Malinformation", Medium. https://medium.com/@tasanoff/fake-news-in-modernnews -media-disinformation-misinformation-and-malinformation-e4fdfa2ab571.

Bell, B.E. and Loftus, E.F. (1988). “Degree of Detail of Eyewitness Testimony and Mock Juror Judgments,” Journal of Applied Social Psychology, 18:1171–1192.

Beridze, I., & Butcher, J. (2019). When seeing is no longer believing. Nat. Mach Intell. 1(8), 332–334.

Berinsky, A. J. (2017). Rumors and health care reform: Experiments in political misinformation. British Journal of Political Science, 47(2), 241–262.

Bond, C. F., Jr., & DePaulo, B. M. (2006). Accuracy of deception judgments. Personality and Social Psychology Review, 10:214–234.

Buchan, N., Croson, R., & Solnick, S. (2008). Trust and Gender: An Examination of Behavior and Beliefs in the Investment Game. Journal of Economic Behavior & Organization. 68(3-4):466-476. doi: 10.1016/j.jebo.2007.10.006

Buller, D.B., & Burgoon, J.K. (1996). Interpersonal Deception Theory. Communication Theory, 6(3): 203 - 242.

Burgoon, J. K., Buller, D. B., Buslig, A. L. S., & Roiger, J. F. (1994). Interpersonal deception: VIII. Further analysis of nonverbal and verbal correlates of equivocation from the Bavelas et al (1990) research. Journal of Language and Social Psychology, 13(4), 396–417.

Burgoon, J.K., Buller, D.B., Ebesu, A.S., & Rockwell, P. (1994). Interpersonal deception: V. Accuracy in deception detection, Communications Monographs, 61:4, 303-325.

Burgoon, J.K., Buller, D.B., & Woodwall, W.G. (1989). Nonverbal Communication: The Unspoken Dialogue. NY: Harper & Row.

Caldwell, M., Andrews, J.T.A., Tanay, T., & Griffin, L.D. (2020). AI-enabled future crime. Crime Science; 9 (1).

Chadwick, A., Vaccari, C., & O'Loughlin, B. (2018). Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. New Media & Society, 4255–4274.

Chesney, R. & Citron, D. K. (2019). Deepfakes and the new information war. Foreign Affairs, January/February, 147–155.

Chong, D., & Druckman, J. N. (2007). A theory of framing and opinion formation in competitive elite environments. Journal of Communication, 57(1), 99–118.

Citron, D. K., & Chesney, R. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. California Law Review, 107, 1753–1819.

Chong, D., & Druckman, J. N. (2007). A theory of framing and opinion formation in competitive elite environments. Journal of Communication, 57(1), 99–118.

DePaulo, B.M., Charlton, K., Cooper, H., Lindsay, J. L. & Muhlenbruck, L. (1997). "The Accuracy-Confidence Correlation in the Detection of Deception." Personality and Social Psychology Review 1: 346–357.

Dobber, T. & Trilling, D. & Helberger, N., & de Vreese, C. H. (2017). Two crates of beer and 40 pizzas: the adoption of innovative political behavioural targeting techniques. Internet Policy Review, 6(4). https://doi.org/10.14763/2017.4.777

Eagly, A. H., & Chaiken, S. (1993). The psychology of attitudes. San Diego, CA: Harcourt, Brace, & Janovich.

Ecker, U. K. H., Lewandowski Y. S. & Tang, D. T. W. (2010). Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition, 38*(8), 1087–1100.

Elaad, E. (2010). Truth bias and regression toward the mean phenomenon in detecting deception.

Psychol Rep. 106(2):641-2. doi: 10.2466/PR0.106.2.641-642.

Evans, N.J., & Park, D. (2015). Rethinking the persuasion knowledge model: schematic antecedents and associative outcomes of persuasion knowledge activation for covert advertising. Journal of Current Issues & Research in Advertising. 36:157–176.

Fallis, D. (2020). The epistemic threat of deepfakes. Philosophy & Technology.

Flynn, D. J., Nyhan, B., & Reifler, J. (2017). The nature and origins of misperceptions: Understanding false and unsupported beliefs about politics. Advances in Political Psychology*,* 38, 127–150.

Fletcher, J. (2018). Deepfakes, artificial intelligence, and some kind of dystopia: The new faces of online post-fact performance. Theatre Journal, 70, 455–471.

Goodfellow I. Nips 2016 tutorial: Generative adversarial networks 2016; arXiv preprint arXiv:1701.00160.

Gilbert, D. T. (1991). How mental systems believe. American Psychologist, 46, 107–119. doi:10.1037/0003-066x.46.2.107

Grice, H. P. (1989). Studies in the way of words. Cambridge, MA: Harvard University Press.

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

Hall, H. K. (2018). Deepfake videos: when seeing isn't believing. Cath. UJL Tech. 27, 51.

Hancock, J., & Bailenson, J. (2021). The Social Impact of Deepfakes. Cyberpsychology, Behavior, and Social Networking. 24(3), 149-152.

Hwang, Y., Ryu, J.Y., & Jeong, S.H. (2021). Cyberpsychology, Behavior, and Social Networking, 188-193.

Iacobucci, S., De Cicco, R., Michetti, F., Palumbo, R., & Pagliaro, S. (2021). Deepfakes Unmasked: The Effects of Information Priming and Bullshit Receptivity on Deepfake Recognition and Sharing Intention. Cyberpsychology, Behavior, and Social Networking. 194-202.

Iyengar, S., & Kinder, D. R. (1987). News that matters: Agenda-setting and priming in a television age. University of Chicago Press.

Jack, C. (2017) Lexicon of Lies: Terms for Problematic Information. Data & Society Research Institute.

Jones-Jang, S.M., Mortensen, T., & Liu, J. (2021). Does media literacy help identification of fake news? Information literacy helps, but other literacies don’t. American Behavioral Scientist. 65:371–388.

Kanwisher, N., McDermott, J., & Chun, M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. J. Neurosci. 17, 4302–4311.

Keys, R., Taubert, J., & Wardle, S.G. (2021). A visual search advantage for illusory faces in objects. Attention Perception Psychophysics. 83, 1942–1953.

Kraut, R. (1980). Humans as lie detectors. Journal of Communication, 30, 209-216.

Lang, A. (2000). The limited capacity model of mediated message processing. Journal of Communication, 50(1), 46–70.

Langguth, J., Pogorelov, K., Brenner, S., Filkuková, P., & Schroeder, D.T. (2021). Don't Trust Your Eyes: Image Manipulation in the Age of DeepFakes. Front. Commun. 6:632317.

Levine, T. (2014) Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection. Journal of Language and Social Psychology.

Levine, T., & Clare, D. D. (2014). Documenting the Truth Default: The Low Frequency of Spontaneous, Unprompted Veracity Assessments in Deception Detection. Human Communication Research, 45(3).

Levine, T., McCornack, S., & Park, H. (1999). Accuracy in Detecting Truths and Lies: Documenting the “Veracity Effect.” Communication Monographs 66 (2):125-144.

MacKenzie, A., & Bhatt, I. (2020). Lies, bullshit and fake news: Some epistemological concerns. Postdigital Science and Education, 2, 9–13.

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

Masip, J., Garrido, E., & Herrero, C. (2009). Heuristic versus systematic processing of information in detecting deception: questioning the truth bias. Psychol Rep. 105(1):11-36. doi: 10.2466/PR0.105.1.11-36.

McCornack, S., & Levine, T. (1990). When lies are uncovered: Emotional and relational outcomes of discovered deception. Communication Monographs, 57(2), 119–138.

McCornack, S., Parks, M. (1986). Deception Detection and Relationship Development: The Other Side of Trust. Annals of the International Communication Association 9(1):377-389.

Metzger, M. J., Flanagin, A. J., & Medders, R. B. (2010). Social and heuristic approaches to credibility evaluation online. Journal of Communication, 60(3), 413–439.

Miller, G. R., & Stiff, J. B. (1993). Deceptive Communication. Sage.

Molden, D. C. (2014). Understanding priming effects in social psychology: What is “social priming” and how does it occur? Guilford Press. *Understanding priming effects in social psychology,* 3–13.

Moy, P., Tewksbury, D., & Rinke, E. M. (2016). Agenda-setting, priming, and framing. The International Encyclopedia of Communication Theory and Philosophy.

Nelson-Field, K., Riebe, E., & Newstead, K. (2013). The emotions that drive viral videos. Australas. Marketing J. 21(4), 205–211. doi:10.1016/j.ausmj.2013.07.003.

Newman, E. J., Garry, M., Unkelbach, C., Bernstein, D. M., Lindsay, D., & Nash, R. A. (2015). Truthiness and falsiness of trivia claims depend on judgmental contexts. Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(5), 1337–1348.

O’Sullivan, M., Ekman, P. & Friesen, W.V. (1988) “The Effect of Comparisons on Detecting Deceit.” Journal of Nonverbal Behavior 12: 203–216.

Paterson, T., & Hanley, L. (2020). Political warfare in the digital age: Cyber subversion, information operations, and “deep fakes.” Australian Journal of International Affairs, 74(4), 439–454.

Pennycook, G., & Rand, D. G. (2019). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. Journal of Personality, 88(2), 185–200.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A.A., Eckles, D., & Rand, D.G. (2019). Shifting attention to accuracy can reduce misinformation online. https://doi.org/10.31234/osf.io/3n9u8

Petty, R. E., & Jarvis, W.B.G. (1996). An individual differences perspective on assessing cognitive processes. N. Schwarz & S. Sudman (Eds.). Answering questions: Methodology for determining cognitive and communicative processes in survey research, 221–257.

Polyakova, A. & Boyer, S. (2018). The future of political warfare: Russia, the West, and the coming age of global digital competition. Brookings Institute, 1–18.

Reid, V.M. (2017). The human fetus preferentially engages with face-like visual stimuli. Curr. Biol. 27, 1825–1828.e3.

Rhodes, G., Brake, S., & Atkinson, A.P. (1993). What's lost in inverted faces? Cognition. 47, 25–57.

Richler, J.J., Cheung, O.S., & Gauthier, I. (2011). Holistic processing predicts face recognition. Psychol. Sci. 22, 464–471.

Richler, J.J., & Gauthier, I. (2014). A meta-analysis and review of holistic face processing. Psychol. Bull. 140, 1281–1302.

Roozenbeek, J., & Van Der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. Journal of Risk Research. 22:570–580.

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. Cornell University.

Scheufele, D.A. (2000). Agenda-setting, priming, and framing revisited: Another look at cognitive effects of political communication. Mass Communication & Society, 3(2–3), 297–316.

Schwarz, N. (2010). Meaning in context: Metacognitive experiences. In B. Mesquita, L. F. Barrett, & E. R. Smith (eds.), The mind in context (pp. 105 -125).

Selezneva, E. (2021). Priming Caution Does Not Decrease Receptivity to Fake News. Revue YOUR Review (York Online Undergraduate Research), 8.

Sinha, P., Balas, B., Ostrovsky, Y., & Russell, R. (2006). Face recognition by humans: Nineteen results all computer vision researchers should know about. Proc. IEEE 94, 1948–1962.

Spivak, R. “Deepfakes”: the newest way to commit one of the oldest crime. (2019). Georgetown Law Technology Review; 3:339–400.

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

Stiff, J., Kim, H.J., & Ramesh, C.N. (1992). Truth Biases and Aroused Suspicion in Relational DeceptionTruth Biases and Aroused Suspicion in Relational Deception. Sage Journals. 19(3).

Sundar, S.S. (2008). The MAIN model: A heuristic approach to understanding technology effects on credibility. The MIT Press. M. J. Metzger & A. J. Flanagin (Eds.), Digital media, youth, and credibility, 72–100.

Sundar, S.S., Molina, M.D., & Cho, E. (2021). Seeing Is Believing: Is Video Modality More Powerful in Spreading Fake News via Online Messaging Apps? Journal of Computer-Mediated Communication. 26(6), 301–319. https://doi.org/10.1093/jcmc/zmab010

Suwajanakorn, S., Seitz, S., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama: Learning Lip Sync from Audio. ACM Trans. Graph. 36, 4, Article 95, 1–13.

Tsfati, Y., Boomgaarden, H. G., Strömbäck, J., Vliegenthart, R. , Damstra, A., & Lindgren, E. (2020). Causes and Consequences of Mainstream Media Dissemination of Fake News: Literature Review and Synthesis.” Annals of the International Communication Association 44 (2): 157–173. https://doi.org/10.1080/23808985.2020.1759443

Taeb, M., & Chi, H. (2022). Comparison of Deepfake Detection Techniques through Deep Learning. J. Cybersecur. Priv. 2(1), 89-106; https://doi.org/10.3390/jcp2010007

Tversky, A., & Kahneman, D. (1973). Availability: A heuristic for judging frequency and probability. Cognitive Psychology, 5(2), 207–232.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. Social Media and Society 2020; 6:1–13.

Vrij, A. (2000). Detecting Lies and Deceit: The Psychology of Lying and Implications for Profession Practice. Wiley and Sons.

Vrij, A., Harden, F., Terry, J., Edward, K. & Bull, R. (1999). The Influence of Personal Characteristics, Stakes and Lie Complexity on the Accuracy and Confidence to Detect Deceit.

Wardle, C., & Derakhshan, H. (2017). Information Disorder: Toward an interdisciplinary framework for research and policymaking. Council of Europe, 20–25.

Wells, G.L., & Leippe, M.R. (1981). "How Do Triers of Fact Infer Accuracy of Eyewitness Identification? Using Memory of Peripheral Details Can be Misleading. Journal of Applied Psychology 66: 682–687.

Wu, Y., Hall, A., Siehl, S., Grafman, J., & Krueger, F. (2020). Neural Signatures of Gender Differences in Interpersonal Trust. *Front. Hum. Neurosci.* 14:225. doi: 10.3389/fnhum.2020.00225

Yang, X., Li, Y. & Liu, S. (2019). Exposing deepfakes using inconsistent head poses. IEEE international conference on acoustics, speech, and signal processing.

Yin, R.K. (1969). Looking at upside-down faces. J. Exp. Psychol. 81, 141–145.

Young, A.W., & Burton, A.M. (2018). Are we face experts? Trends Cogn. Sci. 22,100–110.

Zannettou, S., Sirivianos, M., Blackburn, J. & Kourtellis, N. (2019). The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. Journal of Data and Information Quality, 1(3).

Zuckerman, M., DePaulo, B.M. & Rosenthal, R. (1981). Verbal and Nonverbal Communication of Deception. Academic Press. Advances in Experimental Social Psychology, 14, 1–57.

Zuiderveen Borgesius, F. J., Möller, J., Kruikemeier, S., Ó Fathaigh, R., Irion, K., Dobber, T., Bodo, B. & de Vreese, C. (2018). Online political microtargeting: Promises and threats for democracy. Utrecht Law Review, 14(1), 82–96.

Appendix A

Negative Priming Survey Article

The excerpt read by those who received negative priming on DF:

Please read in entirety the following article that discusses deepfakes.

Published by The Regulatory Review, a publication under the University of Pennsylvania Law School:

Deepfakes are uniquely effective at spreading **disinformation**, which raises critical concerns for democracy and national security. Effective democratic discourse requires that voters start from the same foundation of facts, but deepfakes can lead individuals to live in their own subjective realities and exacerbate social divisions.

Deepfake videos depicting public figures making incendiary comments or behaving inappropriately could also alter election outcomes. As deepfakes become more well known, public officials caught on camera can exploit a "liar's dividend" and claim that a real video is a deepfake. Without clear methods to distinguish what is real from what is not real, the public may lose trust in the media and other public institutions.

Meanwhile, scholars worldwide have explored the problems created by deepfakes and the challenges for regulators seeking to address these problems:

- Deepfake videos raise serious concerns, because videos are inherently credible, interact with cognitive biases, and travel quickly on social media platforms, argue Danielle K. Citron of Boston University School of Law and Robert Chesney of the University of Texas School of Law.
- Tom Dobber of the University of Amsterdam and several coauthors in a report in the International Journal of Press/Politics find that deepfakes can substantially impact viewers' beliefs about a political candidate and influence their views of the candidate's political party. Dobber and his coauthors warn that, as technology advances, the precision and effectiveness of deepfakes are likely to increase. They argue that deepfakes are a "potential new frontier of disinformation warfare" that requires policy action.
- Deepfakes create major challenges to the integrity of the democratic process in the United States, argues Richard L. Hasen of University of California, Irvine School of Law. But because governmental efforts to control deepfakes could implicate free speech rights, Hasen warns that such efforts must be

justified as serving a compelling state interest and then designed in the most narrow fashion possible.

- In an article for the Catholic University Journal of Law and Technology, Holly Kathleen Hall of Arkansas State University explains that government authorities will find it difficult to take action against those individuals who create deepfakes. Deepfake videos often spread anonymously in public forums, Hall notes.
- In an article in the Journal of Intellectual Property Law & Practice, Edvinas Meskys of Vilnius University School of Law and three co-authors developed a taxonomy of deepfakes, classifying them into four use categories: revenge porn; political campaigns; commercial uses; and creative uses.
- Although some lies may receive free speech protection under the First Amendment, a deepfake spreading false information may lose First Amendment protection if it is “not only a falsity, but a forgery as well,” suggests Marc Jonathan Blitz of Oklahoma City University School of Law. Blitz argues that deepfakes presented as news can create a “war of all against all” where every source of information could be false and impact individual decisions or collective democracy negatively.

Appendix B

Positive Priming Survey Article

The excerpt read by those who received positive priming on DF:

Please read the following article that discusses deepfakes.

Published by ThinkAutomation, a business process automation startup in the UK:

Although alarming to some, positive deepfake examples also exist. So, let’s look at some of the beneficial applications of the technology.

Educational deepfakes: Deepfake technology holds positive potential for education. It could revolutionize our history lessons with interactivity. It could preserve stories and help capture attention. How? With deepfake examples of historical figures.

- For instance, in 2018 the Illinois Holocaust Museum and Education Centre created hologrammatic interviews. So, visitors could talk to and interact with Holocaust survivors. They could ask questions and hear their stories. As deepfake technology advances, this kind of virtual history could become achievable on a much wider scale.
- Another example comes from CereProc, a company that ‘resurrected’ JFK in voice. This deepfake made it possible to hear the late president deliver the speech he would have delivered, if not for his assassination. In this way, deepfake technology could help us preserve not just the facts in history books, but the impact historical events had on real people.

Reaching worldwide audiences: Positive deepfake examples also show how the technology can make language barriers (and bad dubbing) a thing of the past.

- Take, for instance, the David Beckham malaria announcement. By using AI technology, David Beckham was shown to speak nine different languages in order to share a message for the Malaria Must Die campaign. Because deepfakes can replicate voices and change videos, it can allow for translated films that use the original actors. The voices sound like the original ones. Crucially, the lip movements even match the words spoken.
- So, with the help of positive deepfakes, we can better share thoughts, films and other creative works on a worldwide basis. Even those with lower budgets. This stands to improve the diversity of our entertainment and content consumption. In other words, deepfakes can shatter language barriers, making content more accessible.

Deepfake in media: In fact, there are a lot of positive deepfake examples for use in the entertainment industry. Deepfakes can keep film characters consistent.

- Consider the times that an actor has passed away. Deepfake technology can fill the role of CGI, recreating the likeness of unavailable past actors. So, the character doesn't have to pass away with their actor. For example, the recreation of the late Peter Cushing in Star Wars: Rogue One (2017), who passed away in 1994. Or, consider times when a character needs to be older or younger than their actor.
- For example, the late Carrie Fisher's character, Princess Leia. Even though the actress herself was not available, her young likeness was recreated. This also demonstrates another positive use of deepfake technology: aging and de-ageing.

Deepfakes in the art world: It's also possible to find positive deepfake examples for the art world.

- AI technology could help us create virtual museums. This would allow access to the world's masterpieces for people that otherwise might not be able to experience them in person. We could share convincing, deepfake artwork across the world. Deepfake technology could even allow us to resurrect dead artists.
- For instance, Salvador Dali at the Salvador Dali Museum in Florida. Perhaps a more novel use of deepfakes could be to bring art to life. Samsung's AI research laboratory has allowed the Mona Lisa to move her head, eyes and mouth. So, if you thought she was watching you before…

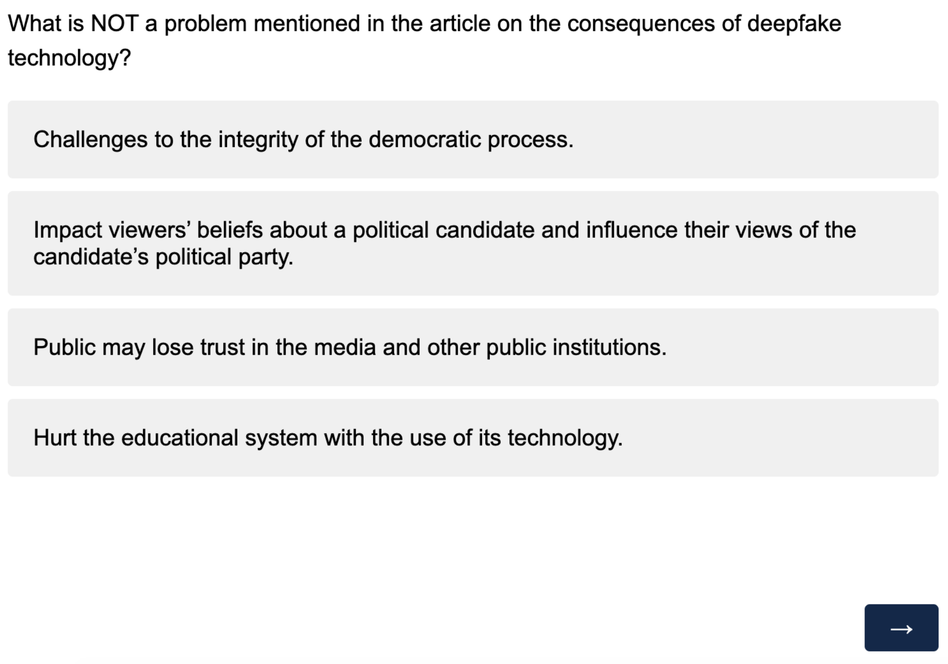
Deepfakes in medicine: The technology behind deepfakes can also provide benefits to the healthcare industry. Specifically, it can provide a boost to data privacy, while helping with the development of new diagnosis and monitoring practices.

- By using the technology behind deepfakes, hospitals can create deepfake patients. That is, patient data for testing and experimentation that's realistic, but doesn't put real patients at risk. So, researchers can use true-to-life deepfake patients, instead of real patient data. From this, there's room to test new methods of diagnosis and monitoring. Or even train other AI to assist with medical decision making.

Appendix C

Negative/Positive Article Evaluation Questions

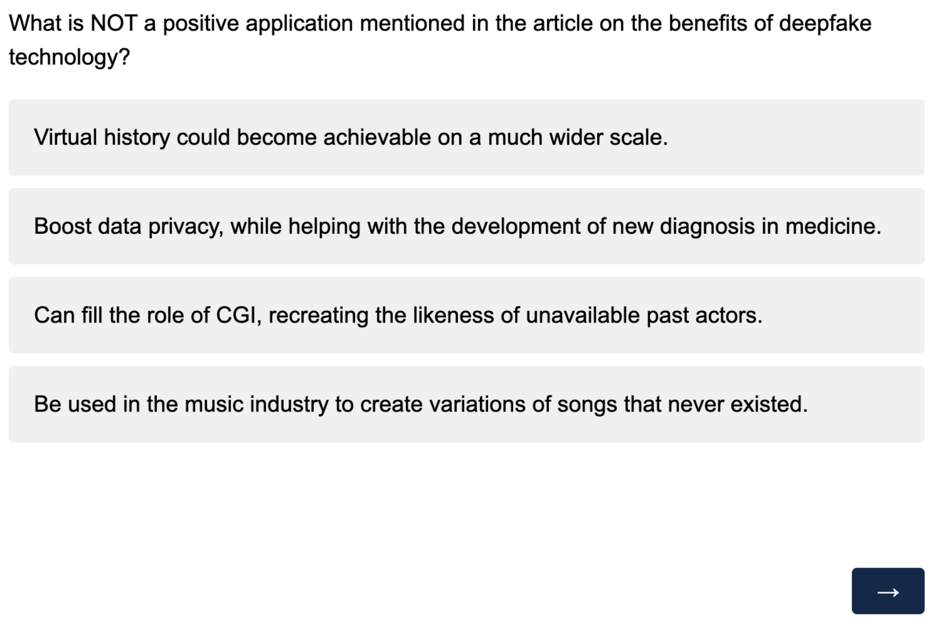
Negative Article Evaluation Question:



What is NOT a problem mentioned in the article on the consequences of deepfake technology?

- Challenges to the integrity of the democratic process.
- Impact viewers’ beliefs about a political candidate and influence their views of the candidate’s political party.
- Public may lose trust in the media and other public institutions.
- Hurt the educational system with the use of its technology.

Positive Article Evaluation Question:



What is NOT a positive application mentioned in the article on the benefits of deepfake technology?

- Virtual history could become achievable on a much wider scale.
- Boost data privacy, while helping with the development of new diagnosis in medicine.
- Can fill the role of CGI, recreating the likeness of unavailable past actors.
- Be used in the music industry to create variations of songs that never existed.

Appendix D

Tom Cruise Deepfake Video Thumbnail

This thumbnail was retrieved from the deepfake video from a cut of the Youtube "Very realistic Tom Cruise Deepfake | AI Tom Cruise," from the user "Vecanoi."