"He Looks Very Real": Media, Knowledge and Search-based Strategies for Deepfake Identification

Dion Hoe-Lian Goh
Wee Kim Wee School of Communication and Information,
Nanyang Technological University, Singapore

Author Note

Correspondence concerning this article should be addressed to Dion Goh, Email: ashlgoh@ntu.edu.sg; Phone: +65-6790 6290.
Address: Wee Kim Wee School of Communication and Information, Nanyang Technological University, 31 Nanyang Link , Singapore 637718.

Abstract

Deepfakes are a potential source of disinformation and the ability to detect them is imperative. While research focused on algorithmic detection methods, there is little work conducted on how people identify deepfakes. This research attempts to fill this gap. Using semi-structured interviews, participants were asked to identify real and deepfake videos, and explain how their decisions were made. Three categories of deepfake identification strategies emerged: the use of surface video and audio cues, processing of the messages conveyed in the video, and the searching of external sources. Participants often used multiple strategies within each category. However, identification challenges occurred due to participants' preconceived notions of deepfake characteristics and the message embodied in the video. This work contributes to research by shifting the focus from the algorithmic detection of deepfakes to human-oriented strategies. Practically, the findings provide guidance on how people can identify deepfakes, which can also form the basis for the development of educational materials.

Keywords: deepfake videos, identification strategies, information credibility, elaboration likelihood model, disinformation

"He Looks Very Real": Media, Knowledge and Search-based Strategies for Deepfake Identification

Introduction

The Web has allowed people access to a wide variety of information sources but there are credibility issues to contend with. A recent development is the deepfake video, or simply, deepfake. Deepfakes use deep learning to synthetically create media that appear to be actual video recordings (Chesney & Citron, 2019).

Although deepfakes have positive uses, a recent study (Ajder et al., 2019) suggested that a significant number were ill-intentioned, such as for pornographic purposes, cyberbullying, political attacks, and other forms of malfeasance (Karasavva & Noorbhai, 2021). Unsurprisingly, concerns about deepfakes being a source of disinformation that can negatively impact trust in media and political institutions have surfaced (Vaccari & Chadwick, 2020). To address this concern, the objective of this paper is to uncover the various ways in which people identify deepfakes. Our findings reveal a range of identification strategies but also demonstrate potential problems in their use.

There is now an emerging body of work on identifying deepfakes where current efforts are mostly focused on machine learning techniques (e.g. Lyu, 2020; Taeb & Chi, 2022; Tolosana et al., 2020). Algorithmic research has yielded useful findings, but the human perspective of deepfake identification has not yet been investigated sufficiently. Doing so is important because the accuracy of algorithms is not sufficiently adequate to be relied on exclusively (Kurohi, 2021). There is an "arms race" between deepfake generators and deepfake detectors. As well, online platforms where deepfakes are found have not yet incorporated automated identification. Hence, human judgement (Carlson, 2017; Gieseke, 2020), despite its flaws, is still needed. Finally, established work in disinformation detection has tended to focus on non-video contexts such as text and images (e.g. Barakat et al., 2021; Baptista & Gradim, 2020). Such findings may not be entirely applicable to deepfakes because the richness of video content, comprising audio, visual and textual elements, may make them more persuasive than other formats, inducing people to perceive them as more credible and shareable (Ajukhadar et al., 2010; Cao et al., 2021), and hence cause more harm to people (Burkell & Gosse, 2019).

There is some emerging relevant research involving deepfakes. Notable is a study on journalists' visual mis/disinformation detection practices that include deepfakes (Thomson et

al., 2022). However, such work is not representative of lay users who often do not have the training or skills of specialized groups, and who serendipitously encounter a questionable video online. Next, in exploratory studies that the present work will build on (Thaw et al., 2020; 2021), participants identified features in videos that characterized them as deepfakes as well as those that led to misidentification. These prior studies were atheoretical, inductive, and did not examine other identification strategies apart from these media-based cues.

To address these gaps in current deepfake research, the present study aims to obtain a better understanding of the strategies that people undertake to identify deepfakes. More specifically, our research questions are: (1) What types of strategies do people use to verify if a video is real or a deepfake? (2) What are the challenges associated with the use of these strategies? In answering these questions, we adopt the Elaboration likelihood Model (ELM; Petty & Cacioppo, 1986) as a theoretical lens. Briefly, the ELM proposes that an individual processes information through two routes. Central route processing describes information processing in a thoughtful manner, where a message's elements are carefully analyzed. It is cognitively demanding and effortful. In contrast, the peripheral route uses unconscious thinking and heuristics where simple cues are used, such as the author's reputation. Further, this study builds upon prior research in information credibility assessments (e.g. Hilligoss & Rieh, 2008; Metzger et al., 2010) that has uncovered a variety of evaluation strategies. While influential, these works do not explicitly deal with video, which, as mentioned above, have different properties when compared to other media types, possibly leading to different identification strategies.

Literature Review

Frameworks for Information Credibility Assessments

Information credibility assessment is a rich area of research with many theoretical frameworks that have been reported in the literature. Such frameworks typically describe the processes or steps undertaken, specific strategies employed and potential factors that may impact such assessments. This section reviews a few prominent examples.

The 3S model proposed by Lucassen and Schraagen (2011) and later revised in Lucassen et al. (2013) describes three major strategies for credibility evaluation of online information: semantic features - accuracy of the content; surface features - how the information is presented; and source features - the characteristics of the information content creator. The use of these strategies is influenced by one's domain expertise and information skills. Those who had domain expertise, knowledge of the topic of the content being

evaluated, would tend to rely more on semantic features, while those with less expertise would focus more on surface features. Further, those with better information skills used a wider array of surface features when compared to those with poorer skills.

Next, Metzger et al. (2010) proposed five heuristics that people use to evaluate online information credibility. First, reputation of source examines the name recognition of the Website where the content is found rather than the content itself. Second, endorsement-based heuristics confer credibility to the content if they perceive that others do likewise. Again, the actual content is not subject to much examination. Third, the consistency heuristic refers to the cross-validation of content across different Websites to ensure that the content is consistent. This heuristic requires more effort than the first two but does not delve deep into the semantics of the content unlike the 3S model. Fourth, the expectancy violation heuristic refers to credibility assessments based on whether a Website meets or fails a person's expectations in terms of presentation or content. If expectations are not met, the content is judged as not credible. Fifth, the persuasive intent heuristic deals with advertising or the presence of commercial content. If this is present, the content is viewed as less credible.

Finally, Hilligoss and Rieh (2008) developed a framework for general information credibility assessment applicable for both online and offline content. Three levels of credibility judgments are defined. The construct level concerns how a person conceptualizes credibility - truthfulness, believability, trustworthiness, objectivity and/or reliability. The specific conceptualization would impact the type of assessment strategies used. The heuristics level are the rules of thumb used to assess credibility which comprise media characteristics, source characteristics, endorsements, and aesthetics of the content. Lastly, interaction level credibility judgments are based on assessments of characteristics of content itself which are more specific and context-dependent than the heuristics level. These include the actual content, peripheral source cues including affiliation and reputation, and peripheral information object cues such as appearance of the content and the emotional effect on the reader.

Related Deepfake Research from the Human Perspective

As discussed, research in deepfake detection strategies from a human-oriented perspective is nascent, especially those focusing on general online users. First, journalistic practices for detecting visual mis/disinformation were investigated by Thomson et al. (2022). A set of five common manipulation techniques and their corresponding detection methods were uncovered. These comprise cloning pixels and reproducing them elsewhere, artificial

blurring of words or objects, saturation/desaturation of colors, retouching to emphasize or obfuscate parts of an image, cropping to remove unwanted areas, and misattribution of content. Detection techniques primarily involve a close examination of the pixels to detect noise and other inconsistencies, as well as statistical and machine learning algorithms to uncover artifact manipulation. Clearly however, these detection techniques are sophisticated, require effort, and likely beyond the reach of an average online user.

To address this issue, two related exploratory studies (Thaw et al., 2020; 2021) sought to elicit media-based features associated with correct and incorrect deepfake identification from participants who were typical Internet users comprising students or working adults. Media-based features are akin to the media characteristics at the heuristics level in Hilligoss and Rieh's (2008) framework. Examples include blurred faces and unnatural voice tones. Interestingly, there were overlaps in the types of features associated with correct and incorrect deepfake identification. Put differently the very features that people used to correctly identify deepfakes may also result in false positives or false negatives, where an authentic video is incorrectly flagged as fake and vice versa. For example, while participants associated low resolution videos with deepfakes, many authentic videos also suffered from such a characteristic. Other examples include poorly perceived unnatural behavior and lack of emotions conveyed by the person in the video. However, these could be actual characteristics of the person in the video, rather than an indication of a deepfake.

Next, Tahir et al. (2021) conducted a survey-based study where users watched videos and were asked to determine which were deepfakes. However, unlike the prior studies above, the videos in the dataset had no audio and only visual features were examined. The most commonly used features were focused on the face of the person in the video, comprising eyes, forehead, lips, cheeks, nose and facial expressions. Results suggest that real videos were easier to identify than deepfakes, with accuracy rates at 88% and 58% respectively. Once again, this indicates that alternative strategies are required to improve on deepfake identification performance other than surface visual cues.

A common finding among the studies above is that deepfake identification performance was surprisingly mediocre at best and suggests the relative sophistication of deepfake generators when compared with the elicited human identification strategies. Further, a shortcoming of these studies was that they did not examine other identification strategies apart from media-based features such as those that were more cognitively demanding. Doing so is important due to the richness of the video medium as mentioned previously. Additionally, these studies were atheoretical, and the strategies were purely inductively

derived. While helpful, incorporating a theoretical lens provides a foundation to systematically guide the conceptualization of this research, organize and analyze the strategies identified, and provide explanations for potential challenges in their use (Rasmussen, 2017).

The Elaboration Likelihood Model

The Elaboration Likelihood Model (ELM; Petty & Caioppo, 1981; 1986) explains attitudinal changes in individuals as they encounter messages and their sources. In the model, people process information through two routes. The central route, also known as systematic processing, describes an effortful, cognitively demanding and logical evaluation of a message. People using this route tend to put more thought in scrutinizing and evaluating messages. In terms of assessing credibility, examples include quality of the arguments and the consistency of the message. This would also imply perhaps searches or consultations with external sources as part of the evaluation process. On the other hand, the peripheral route, or heuristic processing, employs unconscious, automatic thinking and rules of thumb. The reliance is on heuristic cues and cognitive shortcuts for decision-making. This route is therefore less effortful when compared with the central route. In the context of credibility assessments, examples include presentation style, presence of endorsements and reputation of the source. The use of external sources would likely not feature prominently.

The ELM has been influential in the information credibility literature, for example, serving as the framework of choice for organizing Hilligoss and Rieh's (2008) levels of credibility assessment as well as the basis for Metzger's (2007) dual processing model for Web credibility assessment. More recent examples of its use include the contexts of health information (Choi, 2020), online reviews (Aghakhani et al., 2022), and fake news (Chen, Kearney, & Chang, 2021), but not yet in the study of deepfakes whose characteristics are different from these information types.

The present study adopts the ELM to address the atheoretical nature of prior work. First, as a theory of information processing, the ELM serves as a means to explain the ways in which people process a video to decide whether it is real or a deepfake. Second, because disinformation-oriented deepfakes may be said to persuade people of its supposed authenticity, the ELM, as a theory of persuasive communication, is also used to identify the challenges associated with the identification strategies that people employ.

Methodology

Participants

A total of 38 participants were recruited for the present study through poster and email advertisements placed in a local university as well as through snowball sampling. Our sample thus comprised undergraduate or graduate students in which 21 were females and 17 were males. There were 26 participants who were between 21 to 30 years, 11 who were between 31 to 40, and 1 who was 41 to 50 years. Note that age ranges are provided as many participants were comfortable with providing only this information rather than their actual ages. Participants majored in a variety of disciplines including computer science, business, engineering, humanities, and the social sciences.

All participants watched online videos with the majority (31) doing so on a daily basis, two watching at least 4-6 times a week, and five watching at least 1 to 3 times a week. Next, participants were asked about the platforms they used to watch videos and the top three were YouTube (38 participants), Facebook (30) and Instagram (28). Trailing at fourth and fifth places were TikTok (8) and Dailymotion (7) respectively. Finally, the top five video genres watched by participants were entertainment (33), music (27), news (19), gaming (17), and travel (16).

Although these participants were not representative of the entire population of users who may encounter deepfakes, their profiles constitute an important segment of people who would likely encounter them. Specifically, our participants were young, educated adults who were frequent users of the Internet. Interestingly, research suggests that young adults encounter challenges in various aspects of online information retrieval including assessing the reliability of the information (Wineburg & McGrew, 2019).

Data Collection

The present study employed individual semi-structured interviews to answer the two research questions. This qualitative method was used because there is currently little work done on deepfake identification from the human perspective and interviews would facilitate the collection of richer and more in-depth data when compared with questionnaires that collect quantitative data (Hennink, Hutter, & Bailey, 2020). As this was during the height of the COVID-19 pandemic, the interviews were conducted online.

Eight videos were prepared for this study, four of which were real, and four were deepfakes. All videos were sourced from the Internet and were of similar resolution/quality

and length of between 60-80 seconds. The real videos were: (A) a speech by former United States president Barack Obama urging Kenyans to take responsibility in their upcoming national election to make a difference in the nation; (B) a speech by Mark Zuckerberg supporting the investigation of the Russian government's interference in the 2016 United States presidential election; (C) another speech by Barack Obama who called for Republicans in Congress to do their jobs; (D) Will Ferrell and Red Hot Chili Peppers' drummer, Chad Smith, having a drum-off session on The Tonight Show Starring Jimmy Fallon. The four deepfakes were: (E) a speech by Indian politician, Manoj Tiwari criticizing an opposing political party in English and encouraging people to vote for his party; (F) Amazon CEO, Jeff Bezos plays a Talosian alien from the Star Trek movie series, and Tesla CEO, Elon Musk, plays Captain Christopher Pike; (G) actress Lynda Carter, who played Wonder Woman in the 1970s, replaces Gal Gadot in the 2017 Wonder Woman movie trailer; and (H) a speech by UK politician Jeremy Corbyn encouraging voters to support his opponent, Boris Johnson as UK prime minister. These videos were selected as they were non-local and thus increased the likelihood that our participants were not familiar with them, allowing for a more unbiased study.

The study comprised two parts spaced one day apart. In the first part, each participant was assigned to watch two real and two deepfake videos drawn from our dataset. Four links were provided in an email sent to participants, and each link led to a Web page that contained only the video with no other contextual information. The purpose was to focus the participants on the video itself and not rely on peripheral cues. Put differently, the present study sought to extract baseline identification strategies from participants, simulating a situation such as when a video was received through email or a messaging app like WhatsApp or Telegram with no other cues other than a message cajoling the recipient, "*you have to watch this!*" Participants did not know the videos' authenticity but were informed in the email message that they had to identify which were real and which were fake in the second part of the study. They could use whatever identification strategies including Web searches. The "deepfake" term was not used in this part to avoid biasing participants; only "fake" was used. The sequence of the links to the assigned videos were randomized for each participant to minimize order effects (Lavrakas, 2008). Once participants received the email, they could begin watching and identifying their assigned videos.

At an agreed upon time the following day, the second part of the study began. Participants were asked a series of questions in the semi-structured interview. These include demographic questions, and those pertaining to video consumption behavior described in the

previous section. Next, participants were asked whether they had watched the videos before, and to identify which were real or fake. For each video, questions were then posed to understand how the participants arrived at each conclusion, with probes inserted to elicit more details when necessary. Finally, the authenticity of the videos were revealed, and the "deepfake" concept was explained for the benefit of participants who were unfamiliar with this term. Participants were then requested to reflect on why their decisions were incorrect as well as how this experience would change the way the identified real or deepfake videos in the future. Participants' responses were recorded.

There were a couple of reasons for the one-day gap between the two parts of the study. The first was to allow time for participants to use various identification strategies that they felt were appropriate. Second, this gap would allow participants to use these strategies in a presumably more familiar environment that they were comfortable with, thus simulating a naturalistic setting. Conversely, if participants were shown the videos and immediately asked to verify their authenticity, they might only rely on a limited number of strategies.

Data Analysis

All interview responses were transcribed and content analyzed, focusing on the strategies that the participants utilized to identify the real and deepfake videos. The analysis comprised two parts, consistent with prior information credibility assessment research such as Hilligoss and Rieh (2008) and Klawitter and Hargittai (2018).

First, a set of preliminary codes representing known identification strategies was created. These were divided into two categories according to the ELM. Codes for peripheral route processing were derived from the prior exploratory deepfake work described earlier (Thaw et al., 2020; 2021). Due to the limited work done thus far, codes for central route processing were derived from information credibility assessment work in other areas including Hilligoss and Rieh (2008) and Lucassen and Schraagen (2011). Taken together, the initial set of codes involved deepfake identification through video cues, audio cues, and content cues. Next, two research assistants trained in information science separately coded the interview transcripts according to the preliminary list, but also inductively generated new codes where necessary. Following this, the research assistants met to evaluate their coding schemes to derive a final version. The transcripts were then independently recoded and subsequently compared. Any disagreements at that point were resolved through discussion.

Results

Surprisingly, our participants did not perform well in identifying which videos were real or deepfake. Only 13 of 38 participants made all correct identifications, with the majority (16) having only two correct identifications. Four participants correctly identified three videos, while five participants incorrectly identified all videos. It should be noted that all participants reported that they had not watched their assigned videos prior to the study.

Three major categories of diverse strategies emerged from the analysis of the interviews as shown in Table 1, addressing the first research question. However, in answering the second research question and as can be seen in the more detailed strategy descriptions below, challenges were identified in their effective usage. Media-based strategies examine the surface characteristics (Lucassen & Schraagen, 2011) of the audio and video. This corresponds to peripheral route processing (Petty & Cacioppo, 1986) in the ELM. Here, the focus on authenticity was not on the content in the video but its visual and auditory cues. Knowledge-based strategies on the other hand, rely on the understanding of the message conveyed in the video and this is often coupled with one's personal knowledge. Finally, search-based strategies employ resources beyond the video and personal knowledge. Typically online resources are consulted but potentially knowledgeable people may also be approached. These final two categories are more cognitively demanding, and correspond to central route processing (Petty & Cacioppo, 1986) in the ELM.

Media-Based Strategies

By far, the largest number of strategies uncovered were media-based, and all participants reported using at least one of them. In terms of surface video characteristics, most participants relied on detecting the presence of **graphical anomalies** or imperfections in the video such as discoloration, blurred or pixelated edges, and distortions. This strategy represents the lowest level of visual cues where groups of pixels are examined. One participant (P13, female, 21-30 years) noted for example in the Star Trek deepfake, "the character's face is unnatural... the edges are patchy and unclear". Next to be used were **behavioral anomalies**, which participants deem as the perceived unnaturalness of the actors' facial and bodily movements in the video. This encompassed lip synchronization, movement of body parts, and facial expressions. The use of these cues is neatly summarized by one participant (P30, female, 21-30 years) who watched the Mark Zuckerberg video, "... the lips did not close, the neck doesn't look natural and was too straight, the words doesn't match the lips". Finally, the **production quality** of a video was a matter of consideration and

comprises what participants deem are characteristics of professionally produced videos. These include editing quality, video resolution, mood, lighting, and camera work. Examples of responses across various videos include, "Wonder Woman video also had many strange edits" (P3, male, 25 years), "bad video quality, background is blurred" (P11, female, 27 years), and "the zooming effect looks not smoothly done" (P15, female, 32 years).

Table 1. Deepfake Identification Strategies.

Category	Description	Strategies
Media	Use of surface video and audio cues. Message understanding is not used as part of ascertaining authenticity.	• Graphical anomalies • Behavioral anomalies • Production quality • Voice inflection • Sound quality
Knowledge	Messages conveyed by the person(s) in the video are processed to understand the information being conveyed. Often combined with other personal knowledge.	• Message evaluation • Knowledge of the actor • Knowledge of the topic • Knowledge of deepfakes
Search	Sources external to the video are consulted to ascertain authenticity. Includes search and use of known information sources.	• General search • Ask others

Note: The term "actor" is used in this paper to refer to the people depicted in the video, whether real or fake.

In contrast, participants employed fewer surface audio characteristics. One was **voice inflection**, or changes in pitch and tone of a speaker's voice. Participants honed in on whether inflections sounded natural, and for actors that were recognizable, whether their inflections matched. For the latter situation, there was no attempt to process the message and only the tone of voice was assessed. A participant (P29, female, 21 to 30 years) who correctly identified the Mark Zuckerberg video as real noted, "The voice seems to match the usual person on other videos." The second characteristic was more general and termed as s**ound quality**, relating to the ease of listening and understanding of the speech as well as the level of sound distortion and noise. For example, one participant (P18, female, 32) correctly identified the Wonder Woman video as a deepfake because she felt that the background

music was cutting in and out at times, "The sound quality for Wonder Woman was a bit off, the sound was not smooth".

An interesting finding that arose from the interviews is that participants appeared to have certain preconceived notions of the characteristics of real and deepfake videos which guided how these strategies were employed. Specifically, there was an expectation that deepfakes would have (1) graphical and behavioral anomalies, (2) unnatural voice inflections and poor sound quality, and (3) poor production quality. Conversely, a real video would not exhibit these characteristics.

Knowledge-Based Strategies

Participants employed a range of knowledge-based strategies that were often combined to ascertain video authenticity. The most frequently reported strategy was **message evaluation**, in which participants attempted to ascertain if what the actor(s) said was comprehensible although not necessarily truthful or accurate. A participant (P38, male, 25 years) attempted to digest what Manoj Tiwari supposedly said but felt that the content was implausible and therefore concluded that the videowas a deepfake. He explained that, "What he is saying doesn't make sense. I didn't understand what he said." Message evaluation is a basic knowledge-based strategy that was rarely used in isolation but often employed in conjunction with others.

One approach taken by participants was to combine message evaluation with **knowledge of the actor**. Put differently, participants used their personal knowledge of the person in the video to ascertain if he/she was capable of conveying the message. If not, then the video was considered a likely deepfake candidate. One participant (P2, male, 28 years) who correctly identified Obama's speech to Republicans as real noted that, "I think Obama would give that kind of speech based on my understanding of him as a public figure". A second strategy was to use **knowledge of the topic** encapsulated within the message. The goal was to ascertain if there was consistency between the message content and what is known about the topic in general. For example, a participant (P34, male, 21-30 years) remarked that the Jeremy Corbyn video was a deepfake because, "I do not recall Boris Johnson being strongly supported as Prime Minister". Similarly, another participant (P27, female, 31-40 years) said that the Mark Zuckerberg video was real because. "... it really happened. There was quite a debate on that issue". In the first example, the participant determined the video was a deepfake because there was an inconsistency between Jeremy Corbyn's message and what was known about Boris Johnson. In the second example, the

participant established that Mark Zuckerberg's message was aligned with current events and therefore the video was real.

Interestingly, some participants had prior **knowledge of deepfakes** and applied it to identify their assigned videos. These participants felt that deepfakes could be used to influence people on popular or significant topics ("if its a popular topic, people will use it to create deepfakes to influence people"; P26, male, 21-30 years) and hence, videos with perceived low-significant content would not be deepfake candidates ("people won't create deepfake videos for those non-famous people"; P24, male, 25 years). There seemed to be an impression among participants that there were technological limitations in deepfake generators. They felt that deepfakes would be of poorer quality than real videos ("the quality is OK but it doesn't look as good as the others"; P20, female, 41-50 years), and complex scenes ("they are too many movie effect to deepfake"; P15, female, 32 years) were harder to create deepfakes from than more static ones ("he did not have much body movement which makes it easier for editing and face swapping"; P14, female, 21-30 years).

Search-Based Strategies

Two search-based strategies were reported. The first and most popular was **general search**, where participants issued queries to search engines to either obtain information about the video in question or check if there were similar videos for comparison. A participant (P23, male, 21-30 years) who was unsure about the Star Trek video said, "The first thing I do is to go to Google to see if there are articles about whether the video is a deepfake". Through the Google search, the participant eventually found information about this video and established it as a deepfake.

A second strategy was to **ask others**, people whom participants knew, such as friends or colleagues. A typical response was provided by a participant (P21, female, 31-40 years) who correctly identified the Jeremy Corbyn deepfake, "I asked my friends if they had heard the news before, especially news on current affairs that need verification". A few participants also mentioned making a general query to those who may be familiar with the video or its topic. This consultation was done online, as in the case of a participant (P19, female, 41-50 years) who correctly identified the Manoj Tiwari deepfake, "Call people? That's not possible. To me, its Twitter".

Use of Multiple Strategies

Although the strategies are described separately, many participants reported using multiple strategies to ascertain the authenticity of the videos they watched. In fact, the use of a single strategy was rare. Some participants focused only on media-based strategies, such as the following example, "The lighting is inconsistent, compared with other characters, the angle and shadow of his face are not in line with the scene. For example, in terms of facial features, his neck is blurry as compared to the pictures around him. In the Star Trek video, the character's face is unnatural, the edges of the transposed faces are patchy and unclear." Here, the participant (P13, female, 21-30 years) used the production quality (inconsistent lighting) and graphical anomalies strategies (patchy and unclear faces) to correctly identify her assigned deepfakes.

In contrast, other participants used a range of media, knowledge and search-based strategies to arrive at their decisions. One participant (P17, female, 33) described how deepfake identification was done, "I'm most confident that Wonder Woman is fake because of the video itself. The female character's face, it looks awkward, not real, there's a blurry effect around her face when she turns. For Star Trek, I'm not as confident because I didn't watch it. I identified it as deepfake because I saw it when I Googled... I also think some people are producing deepfakes for some Star Trek movies." In this example, the participant employed the graphical anomalies (blurred face) strategy as well as the knowledge of deepfakes (people produce Star Trek deepfakes) and general search strategies (Google search was used).

Post-Study Responses

For participants who did not correctly identify all their assigned videos, many expressed surprise at their performance and in their reflections, comments primarily centered on how real all their assigned videos looked and the difficulty of telling them apart. Two typical examples include, "... but in comparison I think the fake ones are quite well done, I cannot confidently say those two are fake" (P16, female, 33) and "all of them look very real... look at their facial features, their voice " (P22, male, 31-40 years).

Participants were also asked how their deepfake identification strategies would change after knowing their performance. Unsurprisingly, those who correctly identified all their videos would not change their strategies. In all cases, such participants had used a mix of media-based and at least one of the knowledge-based or search-based strategies. One of these participants emphasized the need for the latter two strategies, "I think its always good to

check with multiple sources to verify the information. I will always pay attention to what the video says and check whether it matches with my knowledge. I suggested people to keep up with the latest news and stay informed" (P3, male, 25 years).

Surprisingly, three participants remained adamant about using only media-based strategies despite making incorrect judgments. An inspection of the participants' coded interview responses revealed that all three participants had used only a single strategy for each video identification but would increase the variety of strategies to be used in the future. This is exemplified by the comment of one of these participants (P8, female, 21-30 years), "Pay attention to facial features movement and quality of video. Look at the lip movements". Here, the graphical anomalies, behavioral anomalies and production quality strategies were referred to.

It was encouraging to note that apart from these three participants, all other participants who had least made one incorrect identification in the study reported that they would be using multiple categories of strategies in the future, consistent with those who made all correct identifications. It appears that these participants understood the difficulty of spotting deepfakes and doing so accurately would require a multi-pronged approach. Importantly, the use of search-based strategies featured prominently in all these responses. One example was the use of the general search strategy, "I will look for related content and the background of the video to verify the contents of the video" (P12, male, 26). Another example focuses on asking others, "I will share it with my friends and ask for their opinion" (P14, female, 21-30 years).

Finally, a few participants mentioned that they would use all three categories of strategies in the future, and is captured by the following response (P2, male, 25 years):

> Firstly, I will look at whether this video violates common sense and my recognition or understanding of people who are featured in the video. The recognition includes background, speaking style, and content's logic.
> Secondly, I will check whether the pictures in the video are natural, whether there is jagged rendering, and any rendering anomalies such as shadow mis-coordination. I will also check the sound and the picture is synchronized.
> Lastly, I will do some research on the source of this video.

Here, the participant covered numerous identification strategies. This includes message evaluation (understanding of the content's logic), knowledge of the actor (understanding of people), graphical anomalies (jagged rendering and shadows), behavior anomalies (synchronization of sound and picture), and general search (research on video source).

Discussion

As deepfakes become more entrenched and more accessible, it is imperative that people are equipped with the skills and knowledge to detect they so that they do not fall prey to potential sources of disinformation. Harnessing the ELM and existing information credibility assessment frameworks, the present study sought to uncover the types of strategies that people used to identify deepfakes and the potential challenges encountered.

Eleven deepfake identification strategies organized into three categories were uncovered through interviews with participants. The media-based category of strategies corresponds to peripheral route processing (Petty & Cacioppo, 1986) in the ELM, while the strategies found in the knowledge-based and search-based categories correspond to central route processing. Although the strategies in the former category were easier to execute because it simply entailed participants looking out for visual and auditory cues without need for deeper cognitive processing, it was telling that after the correct answers were revealed, the more cognitively demanding knowledge-based and/or search-based categories featured in all responses about remediating steps for future deepfake identification tasks. This suggests that participants recognized that a more robust approach involving both peripheral and central route processing is required to improve deepfake identification performance. The qualitative data collected in this study is not able to definitively ascertain this and future work involving quantitative data collection would be required.

An important observation from our findings is that our participants often had preconceived notions about deepfake and read video characteristics and set out to confirm them. Put differently, participants employed a form of the expectancy violation heuristic (Metzger et al., 2010) in their assessments. This was especially noticeable in the media- and knowledge-based strategies. Participants expected deepfakes to exhibit graphical and behavioral anomalies, possess low production and sound quality, and feature unnatural voice inflections. Further, the messages conveyed in a deepfake would normally not be aligned with the participant's knowledge of the topic and expectations of the actor. Finally, the content of the video is also assessed against the participant's perceptions of why people create deepfakes and how they are produced. If these expectations are met, the video is assessed to be a deepfake while if not, the video is assessed to be real. This finding is can also be explained by biased elaboration in the ELM (Petty & Cacioppo, 1986) where prior knowledge, which in this case is about deepfakes, influences one's information processing towards an initial opinion.

However, it is easy to see when the use of this method of assessment could break down. Deepfake creation tools are becoming easier to use (Mustak et al., 2023), allowing for better quality and realistic content to be produced. Hence when accompanied by a current topic of interest and a plausible message, there is a likelihood that people who have certain expectations about deepfakes may believe that such a video is authentic. For example, one participant who thought that the Star Trek deepfake was real remarked that, "the quality doesn't make me think that it's a fake. They look natural and the content is believable" (P21, female, 30-40 years). Conversely, there are also real videos that are not of high resolution or production quality, or whose actors and their messages do not align with certain expectations. In such cases, people may mistake them for deepfakes. This was the case of a participant (P9, female, 21-30 years) who thought that Obama's speech to Republicans was not real because of the audio quality and content, "Obama video's audio sounds a little off; Did not believe that a politician like Obama make such a comment". She did not perform any search-based strategies.

Relatedly, an examination of the strategies that led to correct and incorrect video identifications had overlaps. Intuitively, the sole reliance on media-based strategies is insufficient and knowledge-based and search-based strategies are required for better identification accuracy. However, the analysis of the interviews indicates that there were participants who used a mix of strategies across categories and performed perfectly while others using a similar mix did not. In fact, of the five participants who incorrectly identified all their assigned videos, two of them used a mix of strategies. Put differently, the use of more cognitively demanding strategies, as described in the ELM, does not guarantee identification success. Here perhaps, unfamiliarity with the videos could be a factor. All our participants had not yet encountered the videos they watched, and once again, high quality deepfakes with seemingly plausible content may be convincing enough to be considered real. This was exemplified by the lament of the following participant who misidentified the Manoj Tiwari video as genuine despite using both media and knowledge-based strategies, "He looks very real. Whatever he is saying, he looks very authentic" (P38, female, 31-40 years).

Another point of note is that the present study was conducted in a setting in which controls were put in place, including the types of videos watched, instructions for task completion, and a specific timeframe to complete the tasks. Thus, participants had knowledge that they were in a study and were aware that they were watching a mix of real and falsified content even though the term “deepfake” was never mentioned. Participants were thus primed to look out for fakes. Despite this, it is concerning that less than half could correctly identify

all their assigned videos as deepfake or real. Extended into real-world situations, such as viewing a video encountered in social media or receiving one via a Whatsapp chat from a friend, identification performance may be worse if someone were to encounter a deepfake. This is because unlike a study setting, there may be no contextual clues to suggest whether a person should be vigilant about the possibility of deepfake content (Wagner & Blewer, 2019) and thus trigger an investigation of authenticity. Our findings thus call for the need for digital literacy programs as well as naturalistic studies of deepfake identification strategies.

Taken together, the findings showed that our participants demonstrated a potentially useful range of deepfake identification strategies. However as discussed, the findings also highlight challenges in using them.

Conclusion

The following implications may be drawn arising for the findings. First, this work contributes to research by shifting the focus from the algorithmic detection of deepfakes to human-oriented strategies. Deepfake generation algorithms are increasingly more sophisticated and the resulting videos more realistic (Yu et al., 2021). Coupled with the widespread use of the social media and mobile phones for communication and information access, the concern among many segments of society is that deepfakes will be used for spreading disinformation and other forms of falsehoods, bringing harm to various groups and individuals (Cover et al, 2022). Based on extant work on information credibility assessment and the ELM, this work extends information credibility research into the video medium. A framework comprising three categories of deepfake identification strategies are derived. These strategies are often combined in complementary ways as part of the deepfake identification process.

Second, the outcomes of this research show how identification of false information is performed differently in the video medium when compared to other types of media. At an abstract level, the strategies uncovered are aligned with prior research (e.g. Choi & Stvilia, 2015) including the use of media cues and content (Hilligoss & Rieh, 2008), the search for consistency (Klawitter and Hargittai, 2018), and the application of expectancy violation (Metzger et al., 2010) and biased elaboration (Petty & Cacioppo, 1986). The present study contributes to the literature by explicating concrete, deepfake-specific strategies. For example, media-based strategies include behavioral anomalies and product quality, while knowledge-based strategies include the actor and the topic. A complete description is found in Table 1. Additionally, the application of expectancy violation and biased elaboration

centers around preconceived notions of deepfakes and the need for consistency between the message, knowledge of the topic and expectations of the actor.

Next, this work provides practical guidance for identifying deepfakes through the three categories of strategies. In particular, the authenticity of a video cannot be simply evaluated based on a single strategy, and for better results, multiple strategies across categories should be employed. However, the findings also suggest that people should be mindful of the expectancy violation heuristic and how its breakdown may cloud their deepfake assessments. For example, if falsified content was presented in a way that was realistic enough, it could tilt the balance and persuade people to believe in its authenticity. Likewise, creators of videos that aim to inform the public should aim for high production quality that minimizes graphical and behavioral anomalies. If the content of the message is inconsistent with conventional expectations, there should be a deliberate effort to convince potential audiences of the video's credibility.

Finally, the findings from this study can help in the development of educational materials about the deepfakes. Such materials may include what deepfakes are and how they are created. The uses and especially, misuses of deepfakes and their consequences on individuals, organizations and society should be emphasized. Importantly, people should also learn how to spot deepfakes, and the strategies described in this study will be a useful starting point. Here, the strengths and weaknesses of each strategy could be covered as well as the need to use multiple categories for better identification performance. As well, the notion of information credibility in the video context should be discussed. This is important because a video is a relatively complex media object in which there could be many points and methods of falsification. Finally, lessons about the benefits and pitfalls of the consistency and expectation violation heuristics should be highlighted.

Despite the potentially useful results, there are limitations of this study that warrant future work. First, qualitative data was collected from a small number of participants via interviews. Although the data was rich and detailed, the findings may not be applicable to the larger population of people who encounter deepfakes. Two, the participant profiles focused on educated young adults due to the sampling method used. In particular, our young adult participants were those who regularly watched videos online and who were full-time students or working professional who were studying part-time. Other subgroups were not included in the study. Third, the study deliberately sought to elicit strategies based on examination of the video itself. Cues external to the video such as endorsements, source characteristics and other peripheral information were not investigated. Fourth, the one-day lag between the viewing of

the videos and the interview may potentially lead to unreliable recall. Even though mitigation measures were put in place such as clear instructions to participants at the start of study, the relatively short time gap, and probes used during the semi-structured interview itself, this remains a possibility.

Future work could augment the present work by a longitudinal naturalistic study of how people encounter deepfake content while conducting their everyday online activities, what triggers their suspicions, and how they respond to such videos. Through such a study, changes in behaviors in response to suspected deepfake content may also be tracked. Additionally, it would be helpful to adopt quantitative methodologies such as large-scale surveys to verify the stability of the deepfake identification strategies. Participants should also be recruited from a wider range of profiles including age, digital literacy and domain knowledge of video content. The incorporation of external peripheral cues could also be studied, perhaps qualitatively through interviews and quantitatively through experiments. Finally, it may be helpful to use videos from a greater variety of genres and topics to improve on the generalizability of the findings.

Acknowledgements. This research was supported by a Ministry of Education (Singapore) Tier 2 grant (T2EP40122-0004). The author would like to thank the students who helped with the data collection.

References

Aghakhani, N., Oh, O., Gregg, D., & Jain, H. (2022). How review quality and source credibility interacts to affect review usefulness: An expansion of the elaboration likelihood model. *Information Systems Frontiers*. Available at: https://link.springer.com.remotexs.ntu.edu.sg/article/10.1007/s10796-022-10299-w.

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

Ajukhadar, M., Senecal, S., & Ouellette, D. (2010). Can the media richness of a privacy disclosure enhance outcome? A multifaceted view of trust in rich media environments. *International Journal of Electronic Commerce*, *14*(4), 103-126.

Baptista, J. P., & Gradim, A. (2020). Understanding fake news consumption: A review. *Social Sciences, 9*(10), 185.

Barakat, K.A., Dabbous, A., & Tarhini, A. (2021). An empirical approach to understanding users' fake news identification on social media. *Online Information Review, 45*(6), 1080-1096.

Burkell, J., & Gosse, C. (2019). Nothing new here: Emphasizing the social and cultural context of deepfakes. *First Monday, 24*(12). Available at: https://journals.uic.edu/ojs/index.php/fm/article/download/10287/8297.

Cao, D., Meadows, M., Wong, D., & Xia, S. (2021). Understanding consumers' social media engagement behaviour: An examination of the moderation effect of social media context. *Journal of Business Research, 122*, 835-846.

Carlson, M. (2017). Automating judgment? Algorithmic judgment, news knowledge, and journalistic professionalism. *New Media & Society, 20*(5), 1755-1772.

Chen, C.Y., Kearnet, M., & Chang, S.L. (2021). Belief in or identification of false news according to the elaboration likelihood model. *International Journal of Communication, 15*, 1263-1285.

Chesney, R., & Citron, D. (2019). Deepfakes and the new disinformation war. *Foreign Affairs, 98*(1), 147-155.

Choi, W. (2020). Older adults' credibility assessment of online health information: An exploratory study using an extended typology of web credibility. *Journal of the Association for Information Science and Technology, 71*(11), 1295-1307.

Choi, W., & Stvilia, B. (2015). Web credibility assessment: Conceptualization, operationalization, variability, and models. *Journal of the Association for Information Science and Technology, 66*(12), 2399-2414.

Cover, R., Haw, A., & Thompson, J.D. (2022). *Fake news in digital cultures: Technology, populism and digital misinformation*. Emerald Publishing Limited.

Gieseke, A.P. (2020). "The new weapon of choice": Law's current inability to properly address deepfake pornography. *Vanderbilt Law Review, 73*(5), 1479-1515.

Hennink, M., Hutter, I., & Bailey, A. (2020). *Qualitative research methods* (2nd ed.). Sage Publications Ltd.

Hilligoss, B., & Rieh, S.Y. (2008). Developing a unifying framework of credibility assessment: Construct, heuristics, and interaction in context. *Information Processing & Management, 44*(4), 1467-1484.

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

Klawitter, E., & Hargittai, E. (2018). Shortcuts to well being? Evaluating the credibility of online health information through multiple complementary heuristics. *Journal of Broadcasting & Electronic Media, 62*(2), 251-268.

Kurohi, R. (2021, Jul 15). AI Singapore launches $700k competition to combat deepfakes. *The Straits Times*. Available at: https://www.straitstimes.com/tech/ai-singapore-launches-700k-competition-to-combat-deepfakes

Lavrakas. P.J. (2008). *Encyclopedia of survey research methods*. Sage Publications Inc.

Lyu, S. (2020). Deepfake detection: Current challenges and next steps. *Proceedings of the 2020 IEEE International Conference on Multimedia and Expo Workshops*, 1-6.

Lucassen, T., & Schraagen, J.M. (2011). Factual accuracy and trust in information: The role of expertise. *Journal of the American Society for Information Science and Technology, 62*(7), 1232-1242.

Lucassen, T., Muilwijk, R., Noordzij, M.L., & Schraagen, J.M. (2013). Topic familiarity and information skills in online credibility evaluation. *Journal of the American Society for Information Science and Technology, 64*(2), 254–264.

Metzger, M.J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology, 58*(13), 2078-2091.

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

Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y.K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. *Journal of Business Research, 154*, 113368.

Petty, R. E., & Cacioppo, J. T. (1981). *Attitude and persuasion: Classic and contemporary approaches*. Westview Press.

Petty, R. E., & Cacioppo, J. T. (1986). The Elaboration Likelihood Model of persuasion. *Advances in Experimental Social Psychology, 19*, 123–205.

Rasmussen, M. (2017). The role of theory in research. In D. Wyse, N. Selwyn, E. Smith, & L.E. Suter (Eds.), *The BERA/SAGE handbook of educational research* (pp. 53-71). SAGE Publications Ltd.

Taeb, M., & Chi, H. (2022). Comparison of deepfake detection techniques through deep learning. *Journal of Cybersecurity & Privacy, 2*, 89-106.

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., Zaffar, M.A., & Zaffar, M.F. (2021). Seeing is believing: Exploring perceptual differences in deepfake

videos. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, article 174.

Thaw, N.N., July, T., Wai, A.N., Goh, D.H., & Chua, A.Y.K. (2020). Is it real? A study on detecting deepfake videos. *Proceedings of the 83rd Annual Meeting of the Association for Information Science and Technology*, paper 366.

Thaw, N.N., July, T., Wai, A.N., Goh, D.H., & Chua, A.Y.K. (2021). How are deepfake videos detected? An initial user study. *Proceedings of the 23rd International Conference on Human-Computer Interaction*, 631-636.

Thomson, T. J., Angus, D., Dootson, P., Hurcombe, E., & Smith, A. (2022). Visual mis/disinformation in journalism and public communications: Current verification practices, challenges, and future opportunities. *Journalism Practice, 16*(5), 938-962.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. I*nformation Fusion, 64*, 131-148.

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), 1-13.

Yu, P., Xia, Z., Fei., J., & Lu, Y. (2021). A survey on deepfake video detection. *IET Biometrics, 10*, 607-624.

Wagner, T.L., & Blewer. A. (2019). "The word real is no longer real": Deepfakes, gender, and the challenges of AI-altered video. *Open Information Science, 3*, 32-46.

Wineburg, S., & McGrew, S. (2019). Lateral reading and the nature of expertise: reading less and learning more when evaluating digital information. *Teachers College Record, 121*, 1–40.