Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes

Article *in* Personality and Individual Differences · November 2021

DOI: 10.1016/j.paid.2021.111074

Saifuddin Ahmed
Nanyang Technological University

Personality and Individual Differences 182 (2021) 111074

Contents lists available at ScienceDirect

Personality and Individual Differences

journal homepage: www.elsevier.com/locate/paid

Short Communication

Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes

Saifuddin Ahmed [*]

Nanyang Technological University, Singapore

ARTICLE INFO

Keywords:
Deep fakes
Social media
Claim accuracy
Cognitive ability
Disinformation

ABSTRACT

We examine how individual differences influence perceived accuracy of deepfake claims and sharing intention. Rather than political deepfakes, we use a non-political deepfake of a social media influencer as the stimulus, with an educational and a deceptive condition. We find that individuals are more likely to perceive the deepfake claim to be true when informative cues are missing along with the deepfake (compared to when they are present). Also, individuals are more likely to share deepfakes when they consider the fabricated claim to be accurate. Moreover, we find that cognitive ability plays a moderating role such that when informative cues are present (educational condition), individuals with high cognitive ability are less trustful of deepfake claims. Unexpectedly, when the informative cues are missing (deceptive condition), these individuals are more likely to consider the claim to be true and share them. The findings suggest that adding corrective labels can help reduce inadvertent sharing of disinformation. Also, user biases should be considered in understanding public engagement with disinformation.

1. Introduction

In recent times, experts have highlighted concerns about the prevalence of deepfakes and their dangers to an unsuspecting citizenry. Deepfakes are anticipated to be more compelling because they leverage the ‘realism heuristic’ of visual compared to textual material (see Vaccari & Chadwick, 2020). In a post-truth era rife with disinformation, it is essential to understand how people engage with deepfakes to gauge their social implications. A growing body of work regarding disinformation and its consequences proves scholarly interest in the topic (Ahmed, 2021a). However, there is limited evidence to argue how social media users engage with deepfakes and why they share deepfakes. Furthermore, there are reasons to believe that the cognitive ability of online users (an indication of analytical reasoning and general intelligence) may affect how users perceive disinformation (Pennycook & Rand, 2019). Yet, there is a lack of evidence exploring the role of cognitive ability in deepfakes engagement.

In order to develop a more general understanding of how Americans engage with disinformation, there is a need to explore contexts where political attitudes have less of a role to play. There is also a need to understand the efficacy of commonly used informative cues, such as captions, warning labels, and interstitials, to communicate information about the integrity of social media posts, especially deepfakes. Therefore, the current study reports the findings of a survey experiment based on a deepfake of a social media influencer to understand how the public engages with non-political deepfakes. A conceptual model is presented in Fig. 1. The precise goals are discussed below.

At the first step, this study examines whether deepfakes can deceive people. Sophisticated deepfakes are often indistinguishable from reality (Güera & Delp, 2018). Thus, there is a likelihood that individuals who are deceived by deepfakes may perceive its claims to be true. Social media platforms have attempted to subvert the deceptive potential of malicious content by tagging them with “informative cues,” such as warnings, captions, and interstitials. These cues may also work to subvert the deception of deepfakes. We can anticipate this based on recent studies report that warning messages accompanying fake news can successfully reduce their perceived accuracy (Clayton et al., 2020). To test the efficacy of cues, we propose the following hypothesis:

H1. Individuals who watch a deepfake video that is not revealed as fake (deceptive deepfake) are more likely to perceive the presented claims to be true than those who watch a deepfake video where the fabrication is acknowledged (educational deepfake).

Recent findings suggest that those with higher cognitive ability are

[*] Corresponding author at: Wee Kim Wee School of Communication and Information, Nanyang Technological University, 31 Nanyang Link, Singapore, 637718, Singapore.

E-mail address: sahmed@ntu.edu.sg.

https://doi.org/10.1016/j.paid.2021.111074
Received 18 February 2021; Received in revised form 14 June 2021; Accepted 15 June 2021
Available online 23 June 2021

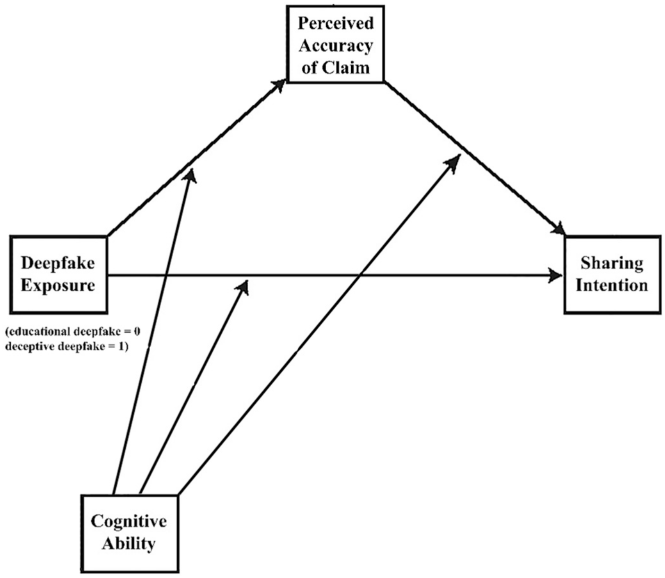


Fig. 1. Conceptual framework of the moderated mediation relationship.

less susceptible to fake news (Pennycook & Rand, 2019). However, it is not immediately evident whether the findings would replicate for deepfakes. Neural cognition literature argues that real faces or face-like stimuli are processed hastily in our minds and are not associated with deeper cognitive processing (Hadjikhani et al., 2009). Therefore, when sophisticated deepfakes manipulate real faces, it may be cognitively challenging to override the premature assessments supposed at the neural level (Smith, 2019). Given the uncertainty and a lack of evidence that higher cognitive ability will safeguard against deepfakes, we propose a research question:

RQ1. How does cognitive ability moderate the relationship between exposure to deepfakes (deceptive vs educational) and the perceived accuracy of the claims?

We are also interested in exploring how the perceived accuracy of deepfakes relate to participants' sharing intention. We speculate that in the absence of informative cues, participants would consider sharing deceptive deepfakes if they consider them to be credible. Our expectation is supported by recent findings reporting that the credibility assessment of fake news mediates the relationship between social media engagement and sharing of fake news (Halpern et al., 2019). Therefore, in line with existing research, we anticipate the following:

H2. Perceived accuracy of claims will mediate the relationship between exposure to deceptive deepfake (vs. educational deepfake) and social media sharing intention.

Given the earlier goal that cognitive ability can influence the perceived accuracy of false claims, we also explore if the mechanism of the sharing intention of deceptive deepfakes via perceived accuracy will be contingent upon individuals' cognitive ability. Recent research suggest that cognitive ability can not only influence perceived accuracy of claims after exposure to disinformation (Pennycook & Rand, 2019) but it can also impact sharing behavior (Ahmed, 2021b). Therefore, we believe that cognitive ability can act as a moderator (see Fig. 1). The following research question is proposed:

RQ2. How does cognitive ability conditionally influence the indirect effect of deepfake exposure on sharing intention through perceived accuracy of claim?

2. Method

2.1. Sample and procedure

Participants were recruited by using a US Qualtrics panel ($N = 440$; age and gender quota utilized). They were randomly assigned to either of the two variants of a deepfake video 'confession' featuring Kim Kardashian, where she appears to confess that she manipulated people online for money (see Appendix A for details). The first 'educational' condition included the variant of the deepfake with the original Instagram caption stating that the video is a conceptual artwork and is a part of a deepfake project. The second 'deceptive' condition included the same deepfake video but without any clarifying caption or hashtags. We use an existing deepfake since it increases the external validity of our study. See Fig. 2.

Participants answered demographic and cognitive ability questions before watching the deepfake video. Next, they answered questions on claim accuracy and sharing intention. In the end, the participants were debriefed.

2.2. Measures

Perceived accuracy of claims (1 = not at all accurate to 4 = very accurate) was measured by "To the best of your knowledge, how accurate is the claim that Kim Kardarshian said that she manipulates people online for money?" ($M = 2.88$; $SD = 0.99$) (see Clayton et al., 2020).

Sharing intention was measured (1 = not at all likely to 4 = very likely) by "How likely would you share this video on your social media? ($M = 1.87$; $SD = 1.12$).

Cognitive ability was measured by the three-item Cognitive Reflection Test (CRT; Pennycook & Rand, 2019). Sample item: "A bat and a ball cost $1.10 in total. The bat costs $1.00 more than the ball. How much does the ball cost? ____ cents". The correct responses were summed to create a scale of cognitive ability ($M = 0.63$; $SD = 0.85$).

Socio-demographics included age, sex, education, and income (see Appendix B). See Appendix C for response quality check.

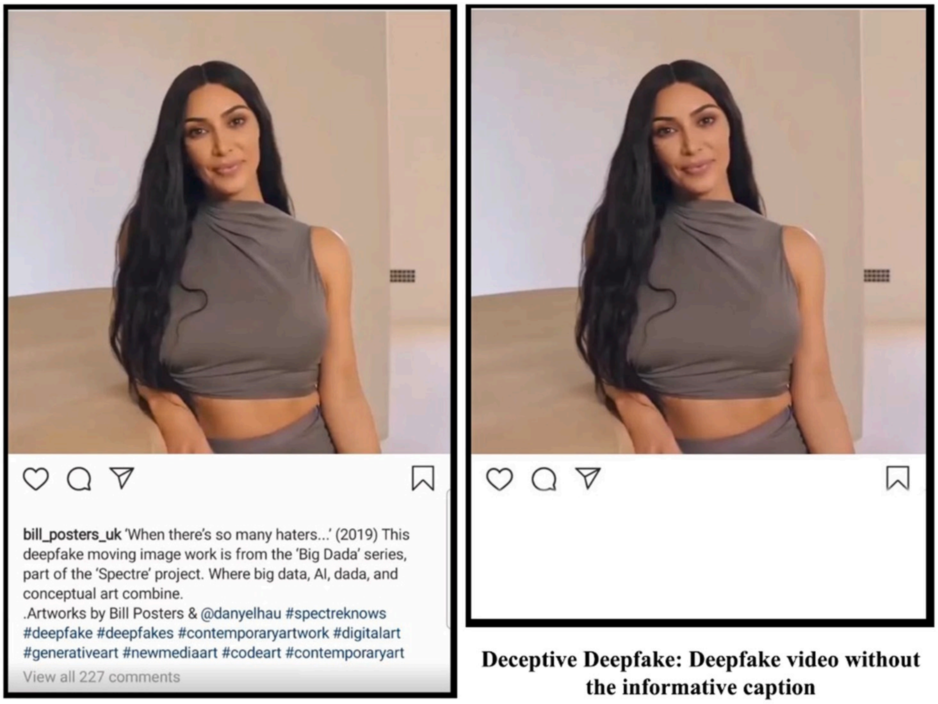
3. Results

Regression analysis suggests that exposure to the deceptive deepfake is positively associated with the fake claim's accuracy (H1: $\beta = 0.093$, $p < .05$). Therefore, those who watched the deepfake without a note of its veracity (deceptive condition) are more likely to perceive that the fake claim is accurate (H1 supported). See Appendix D for details.

The interaction term between the experimental conditions and cognitive ability is statistically significant (RQ1: $\beta = 0.311, p < .05$). The result plotted in Fig. 3 suggests that low cognitive ability individuals do not differ in their evaluation of claim accuracy across conditions. However, high cognitive ability individuals are less likely to trust the deepfake claim in the educational condition. Still, they are more likely to perceive the false claims to be accurate in the deceptive condition.

To test the mediation model (H2), we used the PROCESS macro (see Fig. 4 and Appendix E). An exposure to deceptive deepfake (compared to educational deepfake) was positively associated with perceived accuracy of claims ($B = 0.185$, $SE = 0.094$, $p < .05$), which in turn was positively associated with sharing intention ($B = 0.145$, $SE = 0.041$, $p < .001$). The indirect effect of deepfake exposure conditions on sharing intention through perceived accuracy of claims was statistically significant ($B = 0.027$, $SE = 0.016$, CI = 0.001 to 0.063).

Finally, a moderated-mediation analysis was employed (PROCESS Model 59) to examine the conditional influence of cognitive ability on the mediation process (RQ2). The results suggest that the indirect effects significantly increased as the level of cognitive ability increased (see Appendix F). As such, the indirect effect was stronger for high cognitive ability individuals ($B = 0.065$, $SE = 0.036$, CI = 0.012 to 0.152) than those with moderate cognitive levels ($B = 0.027$, $SE = 0.016$, CI = 0.001



bill_posters_uk 'When there's so many haters...' (2019) This deepfake moving image work is from the 'Big Dada' series, part of the 'Spectre' project. Where big data, AI, dada, and conceptual art combine.
.Artworks by Bill Posters & @danyelhau #spectreknows #deepfake #deepfakes #contemporaryartwork #digitalart #generativeart #newmediaart #codeart #contemporaryart
View all 227 comments

Educational Deepfake: Deepfake video with the informative caption

Deceptive Deepfake: Deepfake video without the informative caption

Fig. 2. Screenshot of educational vs. deceptive deepfake.

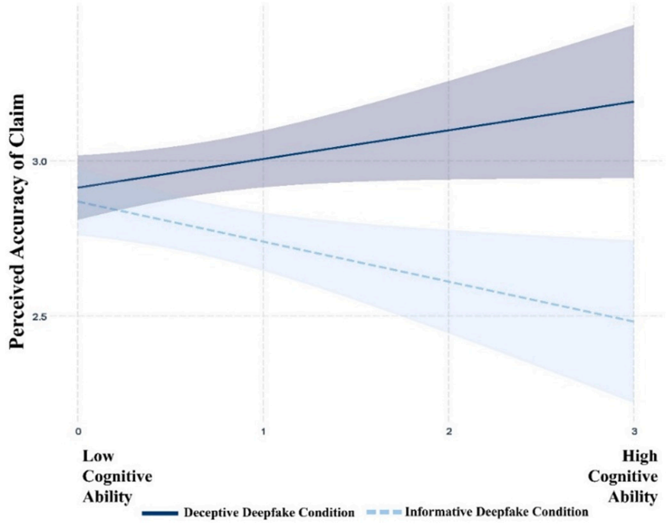


Fig. 3. Visualization of the interaction effect of cognitive ability and conditions.

to 0.063).

4. Discussion

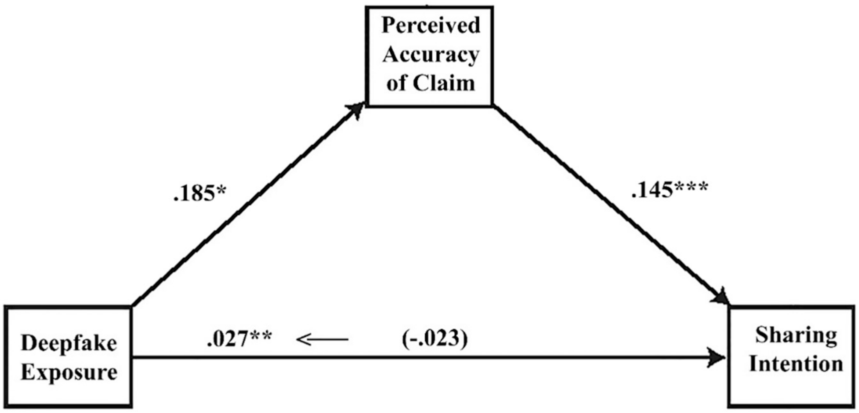
The current study aimed to examine the effects of individual differences on the perceived accuracy of claims and sharing intention of a non-political deepfake. First, the results suggest that individuals are more likely to perceive the deepfake claim to be true when informative cues are absent. This finding raises concern regarding the deceptive nature of well-made deepfakes. On the other hand, the result encourages recent efforts by social media platforms to combat disinformation on their platform and suggests that they may achieve their purpose by alerting users.

Next, when informed about the deceptive nature of deepfakes, high cognitive individuals utilize the available information and are less trustful of the fake claim. However, when the cues are lacking, they are more likely to perceive the claims to be more accurate. On the other hand, low cognitive individuals do not differ in their perceived accuracy of claims irrespective of the conditions. We believe the subject and actor in deepfakes may affect how cognitive ability influences the processing of the fabricated content.

Unexpectedly, our results suggest that individuals with higher cognitive ability are relatively vulnerable to the deceptive non-political deepfake used here. The moderation-mediation findings indicate that because these individuals perceive fake claims to be accurate, they are also more susceptible in their intention to share (fabricated) content. Why would these be? We can offer two possible explanations. First, individuals with higher cognitive ability have higher self-efficacy levels (Chen et al., 2001), which has been found to drive more protective online behavior (Milne et al., 2009). It is possible that these individuals are therefore more concerned by the topic of the deepfake, i.e., how celebrities manipulate audiences. Consequently, they may be persuaded by the video content rather than suspect its authenticity. Second, higher cognitive ability is also associated with greater social stereotyping (Lick et al., 2018). Therefore, these individuals may be more likely to stereotype the subject of the deepfakes, who has been controversial for several reasons, including the misleading endorsement on Instagram. Thus, they are more likely to trust the 'confession' that the subject manipulated her fans for money.

On the other hand, the absence of any substantial changes in perceived accuracy of the fake claim for the low cognitive individuals across the two conditions may suggest their attitude rigidity and an optimism bias where individuals with lower cognitive ability feel that they are less likely to be manipulated.

While a single study cannot disprove existing literature, it does raise a possibility that findings from past literature may not generalize to all



Note. *** $p < .001$;** $p < .01$;* $p < .05$. Estimates are calculated using the PROCESS macro (Model 4; Hayes, 2017). The reference category for deepfakes exposure is the educational condition with the revealing caption. The number in parenthesis is the total effect without adding the mediator. Bootstrap resample = 5,000. Statistical controls include *age, gender, education, income, prior exposure to video & cognitive ability.*

Fig. 4. Visualization of the direct and indirect paths in the mediation process.

domains. The findings contribute toward advancing the scholarship on public engagement with disinformation. As the technology behind deepfakes progresses, they can be utilized to erode the social fabric across societies. While we have seen cognitive ability to protect users against fake news, we observed the fallacies of higher cognitive ability transforming into biased judgment in the non-political deepfake example used here. Malicious actors could exploit such vulnerabilities. However, the results also present a hopeful perspective with regards to the findings. This study adds evidence that perception of accuracy and intention to sharing disinformation are associated. Therefore, digital deception is always worth tackling with flag corrections because it will reduce inadvertent sharing of false disinformation.

Before we conclude, it is essential to acknowledge that the lack of a true neutral condition restricts us from accounting for any broader effects of prior attitudes to the American socialite outside of exposure to the deepfake. It is also likely that not having a “don't know” option for the dependent variable question forced the participants to respond and might have influenced the results. While using real world post and captions add to the validity, another limitation is that this study did not have a measure of respondents' awareness of informative cues. Therefore, we recommend future work to test different variants of deepfakes and examine user vulnerabilities.

CRediT authorship contribution statement

Saifuddin Ahmed conceived and designed the analysis, collected the data, performed the analysis and wrote the paper.

Acknowledgements

The study was funded by Nanyang Technological University - start up grant

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.paid.2021.111074.

References

Chen, G., Casper, W. J., & Cortina, J. M. (2001). The roles of self-efficacy and task complexity in the relationships among cognitive ability, conscientiousness, and work-related performance: A meta-analytic examination. Human Performance, 14 (3), 209-230.

Ahmed, S (2021a). Navigating the maze: Deepfakes, cognitive ability, and social media news skepticism. *New Media & Society*. In press.

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

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., … Nyhan, B. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior, 42*(4), 1073–1095.

Güera, D., & Delp, E. J. (2018, November). Deepfake video detection using recurrent neural networks. In *2018 15th IEEE international conference on advanced video and signal based surveillance (AVSS)* (pp. 1–6). IEEE.

Hadjikhani, N., Kveraga, K., Naik, P., & Ahlfors, S. (2009). Early (N170) activation of face-specific cortex by face-like objects. *NeuroReport, 20*(4), 403.

Halpern, D., Valenzuela, S., Katz, J., & Miranda, J. P. (2019, July). From belief in conspiracy theories to trust in others: which factors influence exposure, believing and sharing fake news. In *ICHCI* (pp. 217–232). Cham: Springer.

Lick, D. J., Alter, A. L., & Freeman, J. B. (2018). Superior pattern detectors efficiently learn, activate, apply, and update social stereotypes. *Journal of Experimental Psychology: General, 147*(2), 209.

Milne, G. R., Labrecque, L. I., & Cromer, C. (2009). Toward an understanding of the online consumer’s risky behavior and protection practices. *Journal of Consumer Affairs, 43*, 449–473.

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

Smith, T. (2019). The neuroscience of deepfakes. Medium https://medium.com/swlh/its-easier-to-fake-a-face-than-a-cat-cfeeccdf0c0d.

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media+ Society, 6(1), 2056305120903408.