Examining Natural Inoculation Against Deepfakes: Can Social Media Exposure Shape Future Credibility Assessments?

# Introduction

The rapid advancement of artificial intelligence has enabled the creation of increasingly sophisticated deepfakes—synthetic media that can realistically manipulate or generate audio-visual content (Kumar & Taylor, 2023). Deepfakes can be used to create false narratives, manipulate public opinion, and erode trust in authentic media, making it crucial to develop effective countermeasures against their misuse. While some applications of deepfakes are benign or even beneficial, most of them are malicious and causes a series of detrimental outcomes (Johnson & Diakopoulos, 2021; Kaushal et al., 2022; Yu et al., 2023).

Current approaches to combating deepfake misinformation largely focus on reactive measures, such as detection technologies and fact-checking systems (Tong et al., 2024). However, these solutions often struggle to keep pace with advancing deepfake capabilities and may come too late after misinformation has already spread. Extant literature highlights the need for preventative strategies that can build resistance against deepfake deception before exposure to harmful content (Marx et al., 2023; Tong et al., 2024).

Inoculation theory offers a promising framework for developing such preventative measures. Just as vaccines work by exposing individuals to weakened forms of a virus to build immunity, inoculation theory suggests that exposure to weakened forms of misinformation can help build cognitive resistance against future deceptive content (Lewandowsky and van der Linden, 2021; Marx et al., 2023). In the context of deepfakes, this traditionally involves structured educational interventions that deliberately expose people to example deepfakes while explaining their deceptive nature.

However, we propose that a form of “natural inoculation” may already be occurring through regular social media use. As users encounter clearly labelled or discussed deepfake content in their social media feeds, this exposure alone can potentially serve as a continuous, low-intensity form of inoculation that builds cognitive resistance against deepfake deception. Beyond mere exposure, the social context of these encounters may further enhance this inoculating effect. Specifically, when users are unsure about the authenticity of a video, they may be motivated to explore the comments for clarifying or confirming consensus (Vogl et al., 2019). Consequently, viewing sceptical comments from others on social media may also influence a user’s perceptions about online content (Kluck et al., 2019; Lee et al., 2021). Applied into our context, they may alert or remind participants of the possibility of deepfake deception.

Understanding how this natural inoculation process works could provide valuable insights for designing more effective deepfake interventions. Rather than relying solely on formal training programs, prevention strategies could potentially leverage and enhance the natural learning that occurs through social media exposure. This leads us to investigate two key questions:

**RQ1:** *Can mere exposure to deepfake videos on social media increase users’ scepticism toward subsequent deepfake content?*

**RQ2:** *How do user-generated sceptical comments enhance the inoculation effect of deepfake exposure?*

To investigate these questions, we plan to conduct a between-subject experiment to stimulate and compare typical social media browsing experiences. Specifically, participants will be randomly assigned to view either regular videos or deepfake videos, with or without sceptical user comments, before evaluating a target deepfake video. By comparing how these different exposure conditions influence participants’ credibility assessments, we can better understand how natural social media interactions might serve as an inoculation mechanism against deepfake deception. Our findings can therefore inform the development of more effective and sustainable approaches to deepfake intervention that work in harmony with users’ natural media consumption patterns.

In 1938, a radio show broadcasted alarming messages to listeners. Among the messages were reports of strange gas explosions on Mars and a meteorite landing on Earth. Later, the broadcast described a horrific monster that fired a heat ray into a public crowd. Listeners were alarmed; many quickly called newspapers or the police.

Today, few are likely to react in the same manner to *The War of the Worlds*, a classic—but fictional—story about a Martian invasion. People today are more familiar with artificial content, such as computer-generated images and audio. Such content appears frequently on social media, especially short-form video feeds (e.g., TikTok, Instagram Reels, YouTube Shorts). Consistently watching these videos likely imparts a healthy dose of skepticism, habituating users and preventing the sort of panic we might see for first-time viewers of synthetic media.

Such skepticism may be more important than ever before due to the advent of deepfakes, synthetic media that utilizes advanced technology to generate realistic content. Although deepfakes can be quite docile, consisting of silly face swaps, they may at times be more malicious, such as when used to generate deceiving audio. Because whole scenes can be conjured up from scratch, having a vigilant populace is crucial in preventing the spread of misinformation.

Preventative measures against misinformation consist of early interventions, such as education (e.g., media literacy) and prebunking, which consists of falsifying potentially misleading messages before they arise. Such methods can reduce the spread of misinformation, which reduces the need for reactive interventions, such as debunking false claims after they arise. Though we know about how active mitigation strategies are useful in proactively deterring misinformation, less is known about how passive and frequent exposure to artificial content on social media affects perceptions of artificial content. It is possible that exposure to deepfakes on social media acts as a form of natural inoculation, or preventative exposure, against future misinformation. Consistent exposure to deepfakes may remind users of their existence, leading them to be more vigilant when appraising online content. In this article, we address the following research questions:

RQ1: How does exposure to deepfakes influence how individuals judge the credibility of future deepfakes?

Does mere exposure to deepfake videos trigger viewers' awareness of synthetic media, leading to increased skepticism toward future deepfakes?

To explore whether exposure to deepfakes plays a role in reducing susceptibility to misinformation, we deploy a controlled experiment that mimic typical social media use. We direct participants to scroll through a short-video feed that may or may not consist of deepfake content. We then measure participants’ perceived credibility toward a deepfake video. We hypothesize that participants exposed to deepfake content would view a subsequent deepfake as less credible.

Though exposure alone may be beneficial, we hypothesize that user-generated comments on social media are crucial in shaping user perceptions. When users are unsure about the authenticity of a video, they may be motivated to explore the comments for clarifying or confirming consensus (Vogl et al., 2019). Ambiguous videos such as deepfakes are likely to motivate exploration for additional information. Comments that are related to the believability of a video on social media (e.g., it’s crazy that this looks so real) may influence a user’s perceptions about online content (Kluck et al., 2019; Lee et al., 2021). Particularly, they may alert or remind participants of the possibility of deception. It is possible that exposure serves as effective learning events, promoting skepticism towards future online content. Our study probes this possibility:

RQ2: How does exposure to deepfakes alongside skeptical comments influence how individuals judge the credibility of future deepfakes?

# Background

## Understanding Deepfakes: An Advanced Form of Misinformation

The proliferation of deepfake technology represents a significant evolution in the landscape of digital misinformation. Deepfakes, which leverage AI to generate highly realistic synthetic audiovisual content, exemplify how technological advancement has fundamentally transformed the creation and dissemination of false information. Unlike traditional fake content, deepfake pose unique challenges by exploiting fundamental human cognitive tendencies—i.e., the realism heuristic, because people are naturally inclined to trust content that closely mimics real-world sensory experiences (Barari et al., 2021; Sundar, 2008). This psychological vulnerability becomes particularly concerning as deepfake technology continues to enhance the fidelity and believability of synthetic content.

The concept of misinformation has garnered increasing attention due to the development of advanced technology. The use of artificial intelligence to produce realistic audiovisual content has allowed for much more complex forgery of lifelike scenes and subsequently, the reversal of the adage: seeing is believing. Deepfakes exemplify the realism heuristic: people are more likely to trust what they can see, because audiovisual content resembles the real world (Barari et al., 2021; Sundar, 2008). Ultimately, people may be more inclined to believe or share deepfakes, increasing transmission rates of false information.

In response to this threat, there have been calls for better education, specifically related to media literacy (Hwang et al., 2021). Organizations have also employed ways to quickly identify misinformation online. Correcting misinformation, or debunking, is then employed to handle extant cases of misinformation. However, its effects are mixed. Studies show that once misinformation is spread, the damage is often permanent (Chan et al., 2017). Even after correction, individuals may still harbor traces that adhere to ideas present in misinformation (Lewandowsky & Van Der Linden, 2021). Perhaps more severely, as GenAI technology continues to advance, it is very likely that it can reach up to a point where deepfakes are technically undetectable.

The Natural “Inoculation” Against Deepfakes

Because reactive strategies are often not fully adequate, many scholars have suggested that proactive measures are crucial in alleviating the consequences of misinformation. The inoculation perspective, following the biological mechanism of vaccines, proposes that exposure to harmless versions of misleading information may be helpful in countering misinformation. By exposing individuals to weakened versions of possible misinformation, an individual’s cognitive defense mechanism may be primed to act when the time comes (Lewandowsky & Van Der Linden, 2021). For example, companies may embed cybersecurity training by sending out their own phishing links, allowing users to fall prey to a fraudulent email without possible harm to the company itself (Kumaraguru et al., 2007). Active interactions like these have been shown to more much more impactful than often-ignored training modules (Caputo et al., 2013).

Unfortunately, like its biological analog, inoculation often fades with time, with effects disappearing in a few weeks (Banas & Rains, 2010). Thus, they require “booster” interventions to maintain potency. Furthermore, outside of organizations such interventions are not easily implemented—people often do not opt in to such interventions and therefore such strategies may not be effective at large scale, affecting only those that are targeted or are self-selected (Roozenbeek et al., 2022). This is especially relevant since misinformation spreads rapidly through social networks, which is often free of major organizational influence.

Furthermore, because it is impractical nor feasible to produce a weakened strain of every form of misinformation, inoculation treatments often offer widespread protection against a range of narratives, providing generalized resistance – a “broad-spectrum” vaccine (Lewandowsky & Van Der Linden, 2021). This combined effect of misinformation and accompanying educational interventions lead to a reduced perceived credibility of all media (Ternovski et al., 2022; Weikmann et al., 2024). The rise of synthetic media further challenges the concept of authenticity for both artificial and authentic video content (Barari et al., 2021; Gregory, 2022). Online, a skeptical public may be primed to doubt the authenticity of all content – as reflected by the phrase ‘fake news’ (Chesney & Citron, 2019).

## Impact of Skeptical Comments on Message Credibility

Short-form video feeds, which consist of infinite scrolling line of short videos, are currently the most popular function on social media (Ceci, 2024). There are often few restrictions for the content placed on such feeds, including that of deepfakes. For example, the popular social media platform TikTok allows users to easily access beautification and face-swap filters (Barari et al., 2021). Consequently, social media users are constantly exposed to deepfake content.

Although numerical metrics are usually present in the user interface, they may not attest to the credibility of a video. Users curious about the validity of an ambiguous video may be motivated to look through the comments (Berlyne, 1954; Vogl et al., 2019). The way users judge online content is heavily influenced by the comment section (Kümpel & Springer, 2016). Comments, especially when paired with ‘likes’, may provide signals related to the value or authenticity of online content (Ali et al., 2022; Jin et al., 2023; Kim & Dennis, 2019). For ambiguous videos, user’s opinions may gravitate towards the majority opinion via the bandwagon heuristic (Sundar, 2008; Walther & Jang, 2012). This may be crucial when users need to evaluate the authenticity of deepfakes. Here, skeptical comments can play an important role in influencing judgement because they signal potentially false information (Graf, 2021; Kluck et al., 2019; Metzger et al., 2010).

Pls use roughly 50 words to explain what message credibility is and introduce the sub-metrics [accurate, authentic, and believable]. Cite the paper that mentioned these items.

# Research Model

In contrast to active interventions, we explore how natural social media activity may impact credibility assessment of deepfakes. At a minimum, exposure to deepfakes in the wild provides an opportunity for individuals to question the authenticity of online content in general. We hypothesize:

H1: Participants that are exposed to deepfakes will perceive future deepfakes as less credible than participants that are not exposed to deepfakes.

H1: Exposure to deepfake videos (vs. non-deepfake videos) will decrease participants’ perceived message credibility of a subsequent deepfake video.

Inoculation theory proposes that resistance to misinformation benefits from both exposure to a message and also its refutational preemption (or prebunking) (Lewandowsky & Van Der Linden, 2021). In the context of deepfakes, pairing commentary alongside deepfakes may be more influential than watching deepfakes alone. Specifically, it is unlikely that users attempt to validate deepfakes by using their own wisdom, instinct, or insight (Tandoc Jr et al., 2018). It is also unlikely they seek out external sources of authentication, such as authority figures or news. On social media, users generally rely greatly on aggregated metrics, using heuristics in order to judge the validity of content (Jin et al., 2023; Tandoc Jr et al., 2018). Skeptical comments, which highlight that something may be wrong, are especially potent in influencing user perceptions of online content (Lee et al., 2021). Accordingly, we hypothesize the following:

H2: Participants that are exposed to deepfakes alongside skeptical comments will perceive future deepfakes as less credible than participants that are not exposed to deepfakes.   
H3: The effect on credibility for future deepfakes is stronger for participants who are exposed to deepfakes alongside skeptical comments than for participants who are exposed to deepfakes lacking skeptical comments.

H2: Exposure to deepfake videos with skeptical comments will lead to lower perceived message credibility of a subsequent deepfake video compared to exposure to deepfake videos without comments.

# Experimental Design

## Participants

We plan to recruit 250 participants from the online crowdsourcing platform Prolific, which allows for the recruitment of a diverse subject pool (Palan & Schitter, 2018). Participants will be paid $2 for the study for approximately 10 minutes of their time. We calculated the required sample size from G\*Power, assuming a medium effect size (.25), high power (.9), and possible dropout/exclusion (~20%).

## Procedure

We utilize a controlled experiment to determine how individuals judge the credibility of a video’s message after 1) deepfake exposure and 2) presence of skeptical comments. To test our hypotheses, we developed a page which emulates the design of typical short-video feeds, similar to those on social media applications (e.g., TikTok, YouTube Shorts, Instagram Reels). Participants are expected to scroll through the videos on the feed to complete the study.

After completing informed consent, users are randomly assigned into one of three conditions. All conditions comprise two stages, the inoculation phase and the testing phase. In the inoculation phase, participants are exposed to five videos. Participants in the control group are presented with non-deepfake videos, whereas participants in the two experimental groups are presented with deepfake videos. The comments are varied within each group. There are no modifications to the comments in videos appearing in the control group. In one of the experimental groups, no skeptical comments appear alongside the video. In the other experimental group, only skeptical comments are displayed. In the testing phase, all participants will watch one deepfake video and be asked to judge the credibility of its message. No comments of any type are provided alongside this target video.

A diagram of a program

AI-generated content may be incorrect.

A white rectangular object with black text

AI-generated content may be incorrect.

Pairwise comparisons of each treatment group with the control will allow us to evaluate the effect of deepfake exposure or the combined effect of deepfake exposure and skeptical comments (H1/H2). Comparison of the two treatment groups allow us to infer if the combined effect of deepfake exposure and skeptical comments is stronger than exposure without skeptical comments (H3).

To account for the possibility that participants may be simply affected by the act of watching five non-deepfake videos, we introduce a baseline condition, in which the inoculation and testing phases are reversed. We expect that credibility assessments in the baseline and control conditions will be similar.

## Stimuli

We obtain deepfake videos from a curated research database. This database includes popular deepfake videos found on YouTube. We also selected non-deepfake videos of similar length and content. In total, we identified a pool of 15 deepfake and 15 non-deepfake videos. For each participant, 5 videos are randomly selected for the inoculation phase, making it unlikely that any two participants watch the same videos in the same order.

The top 50 displayed comments that appeared on each video’s comment feed were categorized by each of the three authors. The categorization was binary: we categorized a comment as skeptical if it pertained to believability or perceived realism (Lee et al., 2021). Initial inter-rater reliability was []. All conflicts were discussed and resolved. 10 comments from each category were selected for each video, prioritizing comments that had unanimous agreement. These comments were added alongside the corresponding video during the inoculation phase.

To select the video that would appear in the testing phase, we pilot tested several deepfake videos. To ensure that there is variance in user responses, we wanted to avoid selecting an extreme video. That is, we wanted to avoid a video where users were likely to give extreme ratings in either direction (i.e., rating as very credible or not credible at all).

## Measures

To explore how users respond to misinformation, we utilize items validated for measuring message credibility (Appelman & Sundar, 2016). We embedded three key items (accurate, authentic, and believable) among distractor items (enjoyable, funny, engaging, entertaining, and useful). Flanking items were included to minimize demand bias by obfuscating the primary dependent variable. Users were asked to indicate how well each adjective represented the video they just watched, from 1 = *describes very poorly* to 7 = *describes very well*. Our key dependent variable, message credibility, is calculated by averaging the scores of accuracy, authenticity, and believability.

We included basic demographic questions in a final survey. Furthermore, we asked participants about their familiarity with deepfakes. It is possible that frequent exposure, outside of our study, would likely lead to diminishing returns for the exposure experienced in our experiment. Participants that are very familiar with deepfakes would likely be less impacted by our manipulation.

## Discussion and Next Steps

This research proposes an innovative perspective on formulating deepfake intervention by examining how natural social media interactions may serve as an informal inoculation mechanism. By investigating both the direct effects of deepfake exposure and the potential enhancement through skeptical comments, we aim to bridge existing gaps between technical detection approaches and human-centered prevention strategies. Building on these experimental insights, our next study will employ qualitative interviews to explore users’ cognitive and perceptual processes when encountering deepfake content. This multi-method research design promises to advance both theoretical understanding of deepfake resistance and practical validations for synthetic media literacy.

# References

Ali, K., Li, C., Zain-ul-abdin, K., & Zaffar, M. A. (2022). Fake news on Facebook: examining the impact of heuristic cues on perceived credibility and sharing intention. *Internet Research*, *32*(1), 379-397.

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, *93*(1), 59-79.

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281-311.

Barari, S., Munger, K., & Lucas, C. (2021). Political deepfakes are as credible as other fake media and (sometimes) real media.

Berlyne, D. E. (1954). A theory of human curiosity.

Caputo, D. D., Pfleeger, S. L., Freeman, J. D., & Johnson, M. E. (2013). Going spear phishing: Exploring embedded training and awareness. *IEEE security & privacy*, *12*(1), 28-38.

Ceci, L. (2024). *Mobile app usage - Statistics & Facts*. Statista. <https://www.statista.com/topics/1002/mobile-app-usage/>

Chan, M.-p. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological science*, *28*(11), 1531-1546.

Chesney, B., & Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, *107*, 1753.

Graf, J. (2021). *The Effects of Uncivil and Skeptical Online Comments On News Credibility and Believability*. George Mason University.

Gregory, S. (2022). Deepfakes, misinformation and disinformation and authenticity infrastructure responses: Impacts on frontline witnessing, distant witnessing, and civic journalism. *Journalism*, *23*(3), 708-729.

Jin, X., Zhang, Z., Gao, B., Gao, S., Zhou, W., Yu, N., & Wang, G. (2023). Assessing the perceived credibility of deepfakes: The impact of system-generated cues and video characteristics. *New Media & Society*, 14614448231199664.

Johnson, D. G., & Diakopoulos, N. (2021). What to do about deepfakes. Communications of the ACM, 64(3), 33–35. https://doi.org/10.1145/3447255

Kaushal, A., Singh, S., Negi, S., & Chhaukar, S. (2022). A Comparative Study on Deepfake Detection Algorithms. IEEE Xplore. https://doi.org/10.1109/ICAC3N56670.2022.10074593

Kim, A., & Dennis, A. R. (2019). Says who? The effects of presentation format and source rating on fake news in social media. *MIS quarterly*, *43*(3), 1025-1039.

Kluck, J. P., Schaewitz, L., & Krämer, N. C. (2019). Doubters are more convincing than advocates. The impact of user comments and ratings on credibility perceptions of false news stories on social media. *SCM Studies in Communication and Media*, *8*(4), 446-470.

Kumaraguru, P., Rhee, Y., Acquisti, A., Cranor, L. F., Hong, J., & Nunge, E. (2007). Protecting people from phishing: the design and evaluation of an embedded training email system. Proceedings of the SIGCHI conference on Human factors in computing systems.

Kumar, A., & Taylor, J. W. (2023). Feature importance in the age of explainable AI: Case study of detecting fake news & misinformation via a multi-modal framework. European Journal of Operational Research. https://doi.org/10.1016/j.ejor.2023.10.003

Kümpel, A. S., & Springer, N. (2016). Commenting quality. *SCM Studies in Communication and Media*, *5*(3), 353-366.

Lee, Y., Huang, K.-T., Blom, R., Schriner, R., & Ciccarelli, C. A. (2021). To believe or not to believe: framing analysis of content and audience response of top 10 deepfake videos on youtube. *Cyberpsychology, Behavior, and Social Networking*, *24*(3), 153-158.

Lewandowsky, S., & Van Der Linden, S. (2021). Countering misinformation and fake news through inoculation and prebunking. *European review of social psychology*, *32*(2), 348-384.

Marx, J., Blanco, B., Amaral, A., Stieglitz, S., & Maria Clara Aquino. (2023). Combating misinformation with internet culture: the case of Brazilian public health organizations and their COVID-19 vaccination campaigns. Internet Research, 33(5), 1990–2012. https://doi.org/10.1108/intr-07-2022-0573

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.

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of behavioral and experimental finance*, *17*, 22-27.

Roozenbeek, J., Traberg, C. S., & van der Linden, S. (2022). Technique-based inoculation against real-world misinformation. *Royal Society Open Science*, *9*(5), 211719.

Sundar, S. S. (2008). *The MAIN model: A heuristic approach to understanding technology effects on credibility*. MacArthur Foundation Digital Media and Learning Initiative Cambridge, MA.

Tandoc Jr, E. C., Ling, R., Westlund, O., Duffy, A., Goh, D., & Zheng Wei, L. (2018). Audiences’ acts of authentication in the age of fake news: A conceptual framework. *New Media & Society*, *20*(8), 2745-2763.

Ternovski, J., Kalla, J., & Aronow, P. (2022). The negative consequences of informing voters about deepfakes: evidence from two survey experiments. *Journal of Online Trust and Safety*, *1*(2).

Tong, J., Marx, J., Turel, O., & Cui, T. (2024). Combatting Deepfake Misinformation on Social Media: A Scoping Review and Research Agenda. AIS Electronic Library (AISeL). https://aisel.aisnet.org/acis2024/25

Vogl, E., Pekrun, R., Murayama, K., Loderer, K., & Schubert, S. (2019). Surprise, curiosity, and confusion promote knowledge exploration: Evidence for robust effects of epistemic emotions. *Frontiers in psychology*, *10*, 2474.

Walther, J. B., & Jang, J.-w. (2012). Communication processes in participatory websites. *Journal of Computer-Mediated Communication*, *18*(1), 2-15.

Weikmann, T., Greber, H., & Nikolaou, A. (2024). After Deception: How Falling for a Deepfake Affects the Way We See, Hear, and Experience Media. *The International Journal of Press/Politics*, 19401612241233539.

Yu, Z., Zhai, S., & Zhang, N. (2023). AntiFake: Using Adversarial Audio to Prevent Unauthorized Speech Synthesis. In Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS ’23), 460–474. Association for Computing Machinery. https://doi.org/10.1145/3576915.3623209