A Natural Form of Inoculation: Can Social Media Exposure Shape Future Credibility Assessments?

Short Paper

Introduction

The rapid advancement of artificial intelligence has enabled the creation of increasingly sophisticated deepfakes—synthetic multimedia that is manipulated to look realistic (Kumar & Taylor, 2024). While some applications of deepfakes are benign or even beneficial, many can be detrimental, such as when used to create false narratives or manipulate public opinion (Johnson & Diakopoulos, 2021). Because deepfakes may erode trust in authentic media, it is crucial to develop effective countermeasures against their misuse.

Current approaches to combating misinformation accompanying deepfakes largely focus on reactive measures, such as detection technologies and fact-checking systems (Tong et al., 2024). Unfortunately, debunking solutions may struggle to keep pace with advancing deepfake capabilities. Likewise, these techniques are often ineffective in countering misinformation that has already spread (Lewandowsky & van der Linden, 2021). To avoid playing a game of catch-up, scholars have suggested the use of proactive strategies which build resistance against deception before it occurs (Marx et al., 2023; Tong et al., 2024).

Inoculation theory offers a promising framework for developing preventative measures against misinformation (Lewandowsky & van der Linden, 2021). Just as how vaccines work by exposing individuals to weakened forms of a pathogen, inoculation theory suggests that individuals can build resistance to persuasion through exposure to a weakened variant (McGuire, 1961). Although passive inoculation techniques (such as through formal educational environments) are beneficial, the inoculation process may be more effective when people actively develop refutations on their own (Lewandowsky & van der Linden, 2021). However, this may be difficult, especially against sophisticated deepfakes which are virtually indistinguishable from reality.

Although it may be difficult for any one individual to recognize the deceptive elements of a piece of online content, users often do not make credibility assessments in a vacuum. Deepfakes are typically presented in social environments, accompanied by user-generated comments. When confronted with a deepfake, users may utilize the comments section as a form of crowdsourced wisdom to appraise the credibility of the video (Vogl et al., 2019). Furthermore, because users are heavily influenced by skeptical comments, interactions with deepfakes on social media may promote general skepticism (Kluck et al., 2019). As such, frequent exposure to deepfakes on social media may be beneficial in conferring resistance to future misinformation, acting as a form of natural inoculation. This leads to the following research questions:

RQ1: *Can mere exposure to deepfake videos on social media increase users' skepticism toward subsequent deepfake content?*
RQ2: *How do skeptical comments contribute to the above effect?*

To investigate these questions, we plan to conduct a between-subjects experiment which simulates the social media environment. Participants are randomly assigned to view either regular videos or deepfake videos, with or without skeptical comments, before evaluating the credibility of a target deepfake video. By comparing how different exposure conditions influence participants' credibility assessments, we can better understand if social media interactions serve as a form of inoculation against deepfake deception. Understanding how organic interactions on social media influence an individual's perceptions towards deepfakes may provide valuable insights into how to design more effective deepfake interventions, adding to the list of effective interventions which stem the flow of misinformation (Lewandowsky & van der Linden, 2021; Tong et al., 2024). In contrast to solely relying on formal interventions, we explore whether prevention strategies can leverage the natural learning that occurs through daily exposure (Compton, 2013). Such an extension seems natural, much like how the process of biological inoculation has its roots in coincidental exposure. Our findings will also contribute to our understanding of the effects of user comments. Though much research has been dedicated to how users utilize comments for singular interactions (e.g., Kluck et al., 2019; Gearhart et al., 2020), our results will add to the understanding of the long-term consequences of repeated interactions.

Background

Deepfakes and 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 (Tolosana et al., 2020). Unlike traditional fake content, deepfakes pose a unique challenge because they exploit the fundamental human cognitive tendency to believe what one sees (Barari et al., 2021; Sundar, 2008). This psychological vulnerability is particularly concerning since deepfakes are continually evolving, further enhancing the fidelity and believability of synthetic content.

In response to these threats, organizations such as social media platform operators have employed strategies to identify misinformation online (McPhedran et al., 2023). Subsequent debunking, such as through tagging content as AI-generated, is commonly used to inform users of potential misinformation. Although these implementations appear helpful, empirical studies investigating the effectiveness of such techniques are mixed (Clayton et al., 2020; Kreps & Kriner, 2022). A meta-analysis showed that the correction is most effective when it is accompanied by a coherent alternative explanation to displace the false information (Walter & Murphy, 2018).

Nevertheless, because fact-checking interventions are inherently delayed, they are unable to completely undo the damage done to those that have already been exposed to the misleading content (Chan et al., 2017). There is also evidence that individuals may still harbor traces that adhere to ideas present in misinformation, long after they have been debunked (Lewandowsky & van der Linden, 2021). Therefore, it is crucial to find ways to protect individuals prior to exposure to misinformation.

Inoculation Theory

The inoculation perspective, following the biological mechanism of vaccines, proposes that previous exposure to harmless versions of misleading information may be helpful in countering subsequent misinformation (Lewandowsky & van der Linden, 2021; McGuire, 1961). By exposing individuals to weakened versions of possible misinformation and refuting them, an individual's cognitive defense mechanism may be primed to act when a real threat occurs. Although there is not a one-to-one correspondence to the biological processes, the inoculation analogy serves as an instructive means to describe a range of defense-building processes (Compton, 2013).

Inoculation may occur through simple prebunking interventions, which may involve formal media literacy education or more active educational games (Hwang et al., 2021; Roozenbeek & van der Linden, 2019). Organizations can also find success through surprise interventions—one study investigated how deliberately sending out phishing emails to employees served as a much more potent intervention than often-ignored training emails (Caputo et al., 2013; Kumaraguru et al., 2007). Unfortunately, like its biological analog, the positive effects of inoculation interventions often fade with time, with effects diminishing in a few weeks (Banas & Rains, 2010). Thus, successful prebunking, like with biological vaccination, require "booster" interventions to maintain potency. Another challenge is implementing interventions at large scales—people often to not opt in to such interventions, making it difficult to achieve herd immunity (Roozenbeek et al., 2022).

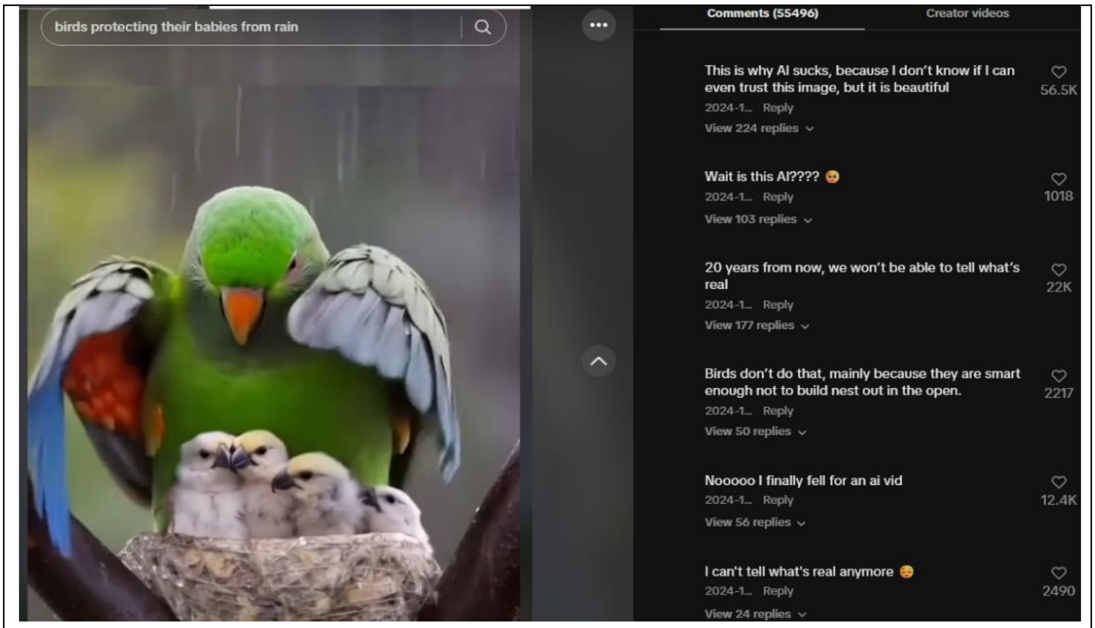
Despite our knowledge of the strengths and weaknesses of active prebunking interventions, research is generally agnostic about how passive exposure to misinformation affects user perceptions of online content (Compton, 2013). Just like how the human body passively acquires immunity through coincidental exposure to pathogens, a constant stream of benign deepfakes may subsequently strengthen an individual's defenses against malicious ones. Research from multiple areas supports the notion that exposure to discrepant events in the wild leads to significant changes in user perceptions. For instance, after falling prey to scams (e.g., online romance, phishing, technical support), people often show a significant and long-lasting reduction in trust towards people and technology (Buse et al., 2023; Coluccia et al., 2020; Kelley et al., 2012). Though such negative consequences are better avoided, the byproduct of these experiences is that users are less likely to be scammed again. Exposure to less harmful deepfakes on social media may work in a similar manner, serving a beneficial role in reducing one's susceptibility to future misinformation—a natural form of inoculation.

In contrast to active interventions, natural exposure to a variety of deepfake content is unlikely to protect against any singular narrative. Instead, the rise of synthetic media likely challenges the authenticity for all content, including authentic ones (Barari et al., 2021; Gregory, 2022). However, this tradeoff may be necessary; it is impractical nor feasible to produce a weakened strain of every form of misinformation (Lewandowsky & Van Der Linden, 2021). Generalized resistance—via a "broad spectrum" vaccine, may be better than the alternative.

Exposure to deepfakes on social media likely complement existing prebunking interventions. Whereas standard interventions may be one-time, frequent exposure to deepfakes on social media may serve as effective "boosters", prolonging the beneficial effects of inoculation. The ubiquity of social media use also implies that the positive effects of natural inoculation will compound to a large scale. Additionally, because users often self-select or may be informed of the goal of interventions, their ability to elicit threat may be weaker than more spontaneous forms of exposure (Compton, 2013; Roozenbeek et al., 2022). When users encounter counterattitudinal messages naturally, there may be a higher chance of eliciting perceived threat, which improves inoculation outcomes. For instance, although a user may believe a video to be genuine, this attitude may be at odds with the skeptical comments which say otherwise. The threat, acting through peer pressure, may be greater than those experienced through more structured settings.

User Comments

Exposure to deepfakes commonly occur on short-form video feeds, the most popular feature of social media (Ceci, 2024). Although more malicious forms of deepfake may be shown to users, the more typical deepfake video likely involve the utilization of beautification and face-swap filters (Barari et al., 2021). Users curious about the veracity of an ambiguous video may be motivated to look through the comments (Vogl et al., 2019). Comments, especially when paired with 'likes', provide signals related to the value or authenticity of online content (Ali et al., 2022; Jin et al., 2023; Kim & Dennis, 2019). In a variety of contexts, studies have shown that user's opinions often gravitate towards the majority opinion via the bandwagon heuristic (Sundar, 2008; Walther & Jang, 2012). Users are also heavily influenced by negative comments, such as those that express doubt (Kluck et al., 2019). Negative comments play an important role in influencing judgement because they signal potentially false information, which is often more valuable than positive or neutral comments (Graf, 2021; Kluck et al., 2019; Metzger et al., 2010). Figure 1 shows a popular video on TikTok and its corresponding comments. Note that skeptical comments are often highly liked or replied to, pushing them to the top of the comment feed.



This is why AI sucks, because I don't know if I can even trust this image, but it is beautiful
2024-1... Reply
View 224 replies
56.5K

Wait is this AI???? 😳
2024-1... Reply
View 103 replies
1018

20 years from now, we won't be able to tell what's real
2024-1... Reply
View 177 replies
22K

Birds don't do that, mainly because they are smart enough not to build nest out in the open.
2024-1... Reply
View 50 replies
2217

Nooooo I finally fell for an ai vid
2024-1... Reply
View 56 replies
12.4K

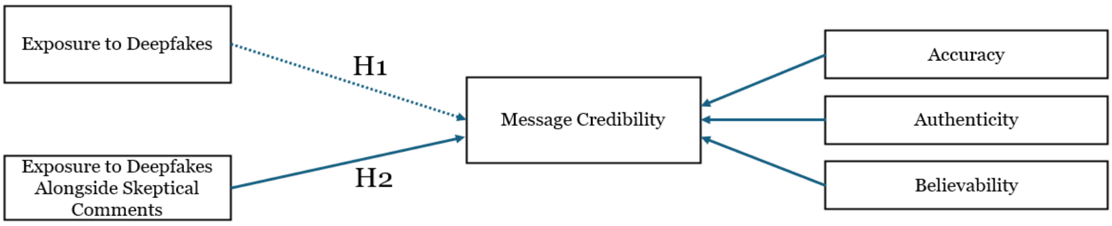
I can't tell what's real anymore 😔
2024-1... Reply
View 24 replies
2490

Figure 1. User Comments on Social Media

Extant studies have been useful in interpreting how one behaves within a singular interaction, but less is known about how such interactions may influence subsequent judgment. Because comments, especially negative ones, imply a refutation, they effectively serve as meaningful components in an inoculation intervention. In other words, comments provide the essential second step of inoculation: exposure to deepfakes is followed by a critical message which exposes some weakness in the content (e.g., "This is why AI sucks, because I don't know if I can even trust this image, but it is beautiful"). Frequent exposure to this two-step (i.e., exposure, refutation) is necessary for an effective inoculation intervention (McGuire, 1961).

Research Model and Hypotheses

In contrast to active interventions, we explore how natural social media activity may impact the credibility assessment of deepfakes. Our research model is summarized in Figure 2.



Note: The dashed line indicates a weaker effect. We hypothesize (H3) that the effect of deepfakes presented alongside skeptical comments will be greater than the effect of deepfakes lacking skeptical comments.

Figure 2. Research Model

Frequent exposure to a variety of deepfakes online may offer widespread protection against a range of narratives, providing a generalized resistance, much like a broad-spectrum vaccine (Lewandowsky & van der Linden, 2021). A skeptical public may be primed to doubt the authenticity of all online content (Barari et al., 2021; Gregory, 2022). We hypothesize:

H1: *Individuals that are exposed to deepfakes will perceive future deepfakes as less credible than individuals that are not exposed to deepfakes.*

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.*

Relative to exposure alone, utilizing skeptical comments is likely to increase the likelihood that users are successfully inoculated. Although users may be alerted after watching ambiguous videos, key components that trigger learning may be missing, particularly that of a threat and refutation (Compton, 2013). Negative comments, especially skeptical ones, may both elicit threat and provide associated refutations. Consider the following comments from popular YouTube deepfake videos (Lee et al., 2021) and how they may challenge the credibility of a video's message: "Man... you guys made this perfect looking." "Hey, This is fake." Although less directed than formal rebuttals, such comments inherently pinpoint a flaw in a message, the flaw being that the message itself came from a fictional source. Comments here can be instrumental, providing users with a way to develop a refutation against deception. We propose that:

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.*

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. This sample size was calculated by 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 webpage 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 (see Figure 3). 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. 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.

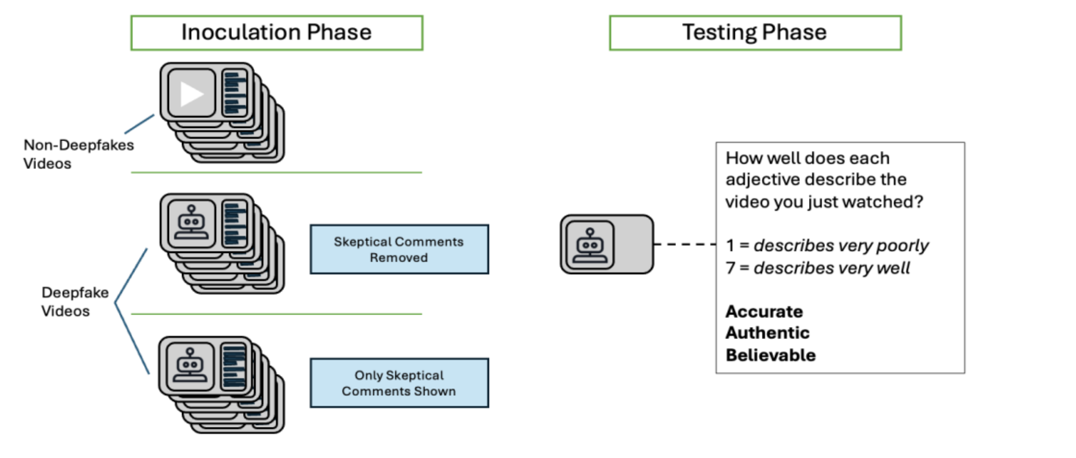

Inoculation Phase
Testing Phase
Non-Deepfakes Videos
Deepfake Videos
Skeptical Comments Removed
Only Skeptical Comments Shown
How well does each adjective describe the video you just watched?
1 = describes very poorly
7 = describes very well
Accurate
Authentic
Believable

Figure 3. Experimental Procedure

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. Our decision for the comment selection is intended to strike a balance between realistic conditions and the ability to test our hypotheses. Because it is unlikely to see videos online that are unaccompanied by comments, we include comments in each condition. However, because we want to see the lasting effect of previous exposure, we omit comments in the video that appears in the testing phase. This lets us set aside the potential effects of the comments which appear in the target video. Similarly, although it may be unlikely to see unanimous threads (i.e., have all skeptical comments), we display comments of only one type in the experimental conditions so that we can focus our analysis on the variable of interest. These design choices limit the generalization of our results but ensure that outcomes are due primarily to our manipulation and not some confounding variable.

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 plan to obtain deepfake videos from a curated research database (Cho et al., 2023). This database includes popular deepfake videos found on YouTube. We also plan to select non-deepfake videos of similar length and content. In total, we will identify 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 appear on each video's comment feed will be categorized by each of the three authors. The categorization is binary: we will categorize a comment as skeptical if it pertains to believability or perceived realism (Lee et al., 2021). All conflicts will be discussed and resolved. 10 comments from each category will be selected for each video, prioritizing comments that had unanimous agreement. These comments will be added alongside the corresponding video during the inoculation phase. To select the video that would appear in the testing phase, we will pilot test several deepfake videos. To ensure that there is variance in user responses, we want 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, which refers to an individual's judgment of the veracity of the content of communication (Appelman & Sundar, 2016). We embedded three key formative measures (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 manipulation experienced in our experiment.

Discussion and Next Steps

This research proposes an innovative perspective on formulating a deepfake intervention by examining how natural social media interactions may serve as an informal inoculation mechanism. We add to existing literature in several ways. First, 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. We supplement the growing list of interventions used to prevent the spread of misinformation through deepfakes (Tong et al., 2024). Second, we contribute to the understanding of how social cues influence online user perceptions. Specifically, we show how skeptical comments may exert a lasting influence on how users perceive related online content. Although this effect may be useful when users encounter misinformation, it may also contribute to the erosion of trust of *all* media (Barari et al., 2021; Gregory, 2022). Such a tradeoff may be implicit when credibility assessments depend on crowdsourced wisdom, and future research is needed to understand such ramifications.

Building on these experimental insights, our next study will employ qualitative interviews to explore users' cognitive and perceptual processes when encountering deepfakes. Rather than focusing on one singular cognitive variable (e.g., message credibility), we can explore how ambiguous videos and skeptical comments work together to elicit threat, motivating individuals to develop refutations. This multi-method research design promises to advance theoretical understanding of how users develop skepticism towards online content, providing insight into how we can embed natural interactions into practical interventions in the fight against misinformation.

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. https://doi.org/10.1108/INTR-10-2019-0442

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, *93*(1), 59-79. https://doi.org/10.1177/1077699015606057

Banas, J. A., & Rains, S. A. (2010). A meta-analysis of research on inoculation theory. *Communication Monographs*, *77*(3), 281-311. https://doi.org/10.1080/03637751003758193

Barari, S., Munger, K., & Lucas, C. (2021). Political deepfakes are as credible as other fake media and (sometimes) real media, *0*(0), 1-57. https://doi.org/10.1086/732990

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. https://doi.org/10.1109/MSP.2013.106

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. https://doi.org/10.1177/0956797617714579

Cho, B., Le, B. M., Kim, J., Woo, S., Tariq, S., Abuadbba, A., & Moore, K. (2023). Towards understanding of deepfake videos in the wild. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, 4530-4537. https://doi.org/10.1145/3583780.3614729

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., & Morgan, E. (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*, 1073-1095. https://doi.org/10.1007/s11109-019-09533-0

Coluccia, A., Pozza, A., Ferretti, F., Carabellese, F., Masti, A., & Gualtieri, G. (2020). Online Romance Scams: Relational Dynamics and Psychological Characteristics of the Victims and Scammers. A Scoping Review. *Clinical practice and epidemiology in mental health : CP & EMH*, *16*, 24–35. https://doi.org/10.2174/1745017902016010024

Compton, J. (2013). Inoculation theory. *The SAGE handbook of persuasion: Developments in theory and practice*, *2*, 220-237.

Gearhart, S., Moe, A., & Zhang, B. (2020). Hostile media bias on social media: Testing the effect of user comments on perceptions of news bias and credibility. Human Behavior and Emerging Technologies, 2(2). https://doi.org/10.1002/hbe2.185

Graf, J. (2021). *The Effects of Uncivil and Skeptical Online Comments On News Credibility and Believability*. (Publication No. 28865450) [Doctoral disseration, George Mason University]. ProQuest Dissertations & Theses.

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. https://doi.org/10.1177/14648849211060644

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

H. M. Buse, J. , Ee, J. and Tripathi, S. (2023) Unveiling the Unseen Wounds—A Qualitative Exploration of the Psychological Impact and Effects of Cyber Scams in Singapore. *Psychology*, *14*(11), 1728-1742. https://doi.org/10.4236/psych.2023.1411101.

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*, 27(3), 1651-1672. https://doi.org/10.1177/14614448231199664

Johnson, D. G., & Diakopoulos, N. (2021). What to do about deepfakes. *Communications of the ACM*, *64*(3), 33-35. 10.1145/3447255

Kelley, C. M., Hong, K. W., Mayhorn, C. B., & Murphy-Hill, E. (2012). Something Smells Phishy: Exploring Definitions, Consequences, and Reactions to Phishing. Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 56(1), 2108-2112. https://doi.org/10.1177/1071181312561447

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. https://www.jstor.org/stable/26848066

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. https://doi.org/10.5771/2192-4007-2019-4-446

Kreps, S. E., & Kriner, D. L. (2022). The COVID-19 infodemic and the efficacy of interventions intended to reduce misinformation. *Public Opinion Quarterly*, *86*(1), 162-175. https://doi.org/10.1093/poq/nfab075

Kumar, A., & Taylor, J. W. (2024). 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*, *317*(2), 401-413. https://doi.org/10.1016/j.ejor.2023.10.003

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*, 905-914. https://doi.org/10.1145/1240624.1240760

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. https://doi.org/10.1089/cyber.2020.0176

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. https://doi.org/10.1080/10463283.2021.1876983

Marx, J., Blanco, B., Amaral, A., Stieglitz, S., & Aquino, M. C. (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

McGuire, W. J. (1961). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *The Journal of Abnormal and Social Psychology*, *63*(2), 326. https://psycnet.apa.org/doi/10.1037/h0048344

McPhedran, R., Ratajczak, M., Mawby, M., King, E., Yang, Y., & Gold, N. (2023). Psychological inoculation protects against the social media infodemic. *Scientific Reports*, *13*(1), 5780. https://doi.org/10.1038/s41598-023-32962-1

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. https://doi.org/10.1111/j.1460-2466.2010.01488.x

Palan, S., & Schitter, C. (2018). Prolific. ac—A subject pool for online experiments. *Journal of behavioral and experimental finance*, *17*, 22-27. https://doi.org/10.1016/j.jbef.2017.12.004

Roozenbeek, J., Traberg, C. S., & van der Linden, S. (2022). Technique-based inoculation against real-world misinformation. *Royal Society Open Science*, *9*(5), 211719. https://doi.org/10.1098/rsos.211719

Roozenbeek, J., & van der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. *Journal of risk research*, *22*(5), 570-580. https://doi.org/10.1080/13669877.2018.1443491

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. https://doi.org/10.1177/1461444817731756

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

Tong, J., Marx, J., Turel, O., & Cui, T. (2024). Combatting Deepfake Misinformation on Social Media: A Scoping Review and Research Agenda. https://doi.org/10.1016/j.inffus.2020.06.014

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. https://doi.org/10.3389/fpsyg.2019.02474

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication monographs*, *85*(3), 423-441. https://doi.org/10.1080/03637751.2018.1467564

Walther, J. B., & Jang, J.-w. (2012). Communication processes in participatory websites. *Journal of Computer-Mediated Communication*, *18*(1), 2-15. https://doi.org/10.1111/j.1083-6101.2012.01592.x