A natural form of inoculation: interacting with computer-generated content on social media reduces believability of future misinformation

Introduction

In 1938, panic swept listeners during a scheduled radio broadcast. An alarming news report mentioned strange gas explosions on Mars and a meteorite landing on Earth. Later, the broadcast described a horrific "monster" which fired a heat ray into a public crowd. Some people shared this sensational report as genuine, calling newspapers or the police to inquire.

Today, few are likely to react strongly to *The War of the Worlds*, a classic – but fictional – story about a Martian invasion. People are now more familiar with artificial content, such as computer-generated images and audio. Such content appears frequently on social media news feeds, especially short-form video feeds (e.g., TikTok, Instragram Reels, YouTube Shorts). Constant exposure to artificial content imparts a healthy amount of skepticism, habituating users and preventing the sort of panic we might see for first-time viewers.

Our study attempts to explore how passive exposure and interaction with computer-generated content likely serves to reduce the believability of such content. Specifically, we explore whether frequent, passive encounters with artificial content on social media serve to inoculate users against future encounters, therefore possibly increasing skepticism towards misinformation. We add to a well-established stream of research which has suggested many active interventions to promote information skepticism, such as media literacy, inoculation, and debunking.

We argue that constant passive engagement with artificial content online may serve as a natural form of inoculation, or preventative exposure, which gradually imparts skeptical thinking. Exposure to videos on social media is often low-stakes – funny or emotional videos that use computer generated images are unlikely to cause much harm compared to partisan political content. Because inoculation interventions often fade with time, the prolific use of social media also serves as effective "boosters", reminding users that the content they see online may often be artificially generated.

We further explore how user generated comments are influential in shaping behavior. When users see conflicting or ambiguous information, they may be motivated to look through the comments for clarifying or confirming consensus. Importantly, user comments on social media may emphasize skepticism, often at exaggerated levels. Viewers may use these comments as salient cues which inadvertently influence one's mental model about the believability of video content.

To test these hypotheses, we employ a controlled laboratory experiment to study the effects of exposure to artificially generated content. We expose participants to a mock short-form video

feed that either contains or does not contain computer generated content. We then measure participants' credibility ratings for a target video that contains subtle misinformation. We also explore the effects of comments by manipulating the presence and types of comments present on each video.

Background

Misinformation

The concept of fraud and misinformation has garnered increasing attention due to the development of new technologies. The use of artificial intelligence (AI) to mimic realistic scenarios has allowed for much more complex forgery of lifelike scenes and subsequently, the reversal of the adage: seeing is believing. Both deepfakes, content created with AI, and cheapfakes, which comprise of non-AI computer-generated images, take advantage of 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).

A collaborative effort has been taken to tackle possible consequences of these new technologies. In general, there are calls for better education, specifically that related to media literacy (Hwang et al., 2021). Furthermore, correcting misinformation, or debunking, is often employed to handle extant cases of misinformation, though its effects are mixed (Chan et al., 2017). Studies show that once misinformation is spread, the damage is often permanent. Even after correction, individuals may still harbor traces that adhere to ideas present in misinformation (Lewandowsky & Van Der Linden, 2021). Therefore, scholars have suggested that preventative measures are crucial in alleviating the consequences of misinformation.

Many studies have used the inoculation perspective, which borrows from biological mechanism of vaccines. 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). 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, such interventions are by nature optional – people often do not opt in to such interventions and therefore such strategies may not be effective at large scale and only affect the people who are targeted or are self-selected (Roozenbeek et al., 2022).

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). However, 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 challenges 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).

Short-Form Video Applications

Short-form video applications are currently the most popular form of entertainment on social media (Ceci, 2024). Such applications incorporate feeds which incorporate infinite scrolling through short, engaging videos. There are often few restrictions for the content placed on such feeds, including that of computer-generated content. Furthermore, certain affordances (e.g., filters) actively encourage the use of AI generated content (Barari et al., 2021). Therefore, in any given instance of use, users may be exposed to several computer-generated videos.

Most short-form video applications include in their user interface a metrics section which contains the number of likes/favorites, the number of shares, and a button with an option to see comments. Past research has consistently found that the user metrics, such as ratings or likes, are influential in affecting believability (Ali et al., 2022; Jin et al., 2023; Kim & Dennis, 2019).

One of the key modes of interaction on media posts is the comments section. Comments contribute additional information that may be useful for addressing the quality of content (Kümpel & Springer, 2016). Users' opinions may also be swayed by other comments, following the majority opinion, the so called bandwagon heuristic (Sundar, 2008; Walther & Jang, 2012). For believability, skeptical comments are much more influential than ratings in influencing judgement because they signal potentially false information (Graf, 2021; Kluck et al., 2019; Metzger et al., 2010).

Learning

To reduce the damaging effects of misinformation, typical interventions are designed to foster change in a person's mental model – their forms of mental representation (Vandenbosch & Higgins, 1996). Mental models describe a person's thoughts towards specific stimuli, such as artificially generated content. Ideally, the goal of interventions is to activate skepticism when an individual encounters potential misinformation.

Learning, which can be considered a change in mental models, is proposed to occur either through assimilation or accommodation (Piaget, 2013). Assimilation refers to the process by which elements of the environment are incorporated into an individual's mental model. That is, information is incorporated into existing cognitive structures. Common interventions act through this mode: debunking adds to existing mental models by correcting a previous source of misinformation; inoculation serves to cue individuals by presenting a harmless piece of misinformation to guard against similar cases in the future.

Inoculation, education, and debunking interventions trigger gradual changes to mental models through assimilation. They present individuals with new information in an attempt to update one's understanding. However, these methods often fall short because they do not present a genuine situation that leads to mental conflict, a situation which is a precursor to dramatically altering mental models. Furthermore, these methods often involve self-selection: people do not need to look at or believe debunking claims, may not be randomly shown inoculation messages, and may not have access or motivation to seek relevant media literacy education.

Significant changes in mental models may occur through assimilation, but deeper learning is more likely to occur through the process of accommodation (Piaget, 2013). Accommodation occurs when mental models cannot simply be modified by addition. Instead, the mental model needs to be redefined to process a new source of information. Accommodation is often rare; most forms of learning are often incremental. However, because accommodation involves

significant changes to mental models, it is often associated with greater and permanent learning.

In terms of misinformation, accommodation may occur through extreme cases of conflict which contradict existing mental structures. One case that may be relevant to most is the situation of falling prey to an online scam. First-time victims may experience disbelief or confusion after realizing they fell prey to fraud. However, the event acts as an important learning moment which is likely to lead to strong skepticism of similar situations in the future. Though not as extreme, consistent encounters with potentially artificial content online may also trigger curiosity and a drive to seek out confirmatory information (Vogl et al., 2019). Interaction with lifelike, computer-generated images potentially requires mental model updates. Specifically, there is cognitive dissonance that needs to be resolved: *is seeing believing*?

Research Model

The process of accommodation, or changing mental models to fit with disconfirming information, is expected to occur whenever an individual views a discordant video that is potentially artificial. We expect that when a user sees a video that is likely to be artificial, they are more likely to seek out information to confirm the validity of the content. Encounters with deepfakes that appear genuine are likely to spur curiosity and subsequent exploration to verify authenticity (Vogl et al., 2019).

Users may validate content by using their own wisdom, instinct, or insight (Tandoc Jr et al., 2018). If motivated, they may seek out external sources of authentication, such as authority figures or news sites. However, we posit that it is unlikely that users are always motivated to exert great effort to check each piece of content on social media. Instead, users often rely greatly on aggregated metrics and use heuristics in order to judge the validity of an item (Jin et al., 2023; Tandoc Jr et al., 2018). Although numerical metrics are usually present in the user interface, users may only explore the comments section when viewing ambiguous videos that trigger curiosity (Berlyne, 1954; Vogl et al., 2019).

H1: Users are more likely to check comments when viewing an ambiguous video.

Past research into misinformation has typically used the construct of credibility, which is typically comprised of accuracy, authenticity, and believability (Appelman & Sundar, 2016). There are also different types of credibility, such as those specific to the source, the message, or the specific media (Flanagin & Metzger, 2008). However, on social media users are not often exploring credible news articles, which makes traditional concepts of credibility potentially outdated (Graf, 2021; Metzger et al., 2010). Users may scroll through hundreds of videos, some of which are obviously computer generated while others may be more ambiguous. It is unlikely users assess the credibility of anonymous individuals that are not affiliated with a news organization. Yet, users may implicitly make a believability judgment when they confront any content that appears artificial. Importantly, comments related to the perceived realism or believability are likely to be influential for ambiguous videos (Lee et al., 2021). Viewing comments related to skepticism is likely to lead to a change in mental models associated with media content. We expect that this change will be reflected in how users judge the believability of future audiovisual content.

H2: After viewing videos using CGI, individuals will perceive future videos as less believable.

H2a: The effect of H2 is stronger when individuals view the comments section of videos using CGI.

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.

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.

Flanagin, A. J., & Metzger, M. J. (2008). *Digital media and youth: Unparalleled opportunity and unprecedented responsibility*. MacArthur Foundation Digital Media and Learning Initiative Cambridge, MA, USA.

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.

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.

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.

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.

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.

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.

Piaget, J. (2013). *The construction of reality in the child*. Routledge.

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).

Vandenbosch, B., & Higgins, C. (1996). Information acquisition and mental models: An investigation into the relationship between behaviour and learning. *Information Systems Research*, *7*(2), 198-214.

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.