Diverse Misinformation: Impacts of Human Biases on Detection of Deepfakes on Networks

Keywords: Privacy, Misinformation, Deepfakes, Social Networks, Survey study, Survey study

Extended Abstract

Social media users are not equally susceptible to all misinformation. We call “diverse misinformation” the complex relationships between human biases and demographics represented in misinformation. To investigate how users’ biases impact their susceptibility to misinformation and their ability to correct each other, we analyze human classification of computer-generated videos (deepfakes) as a type of diverse misinformation. We chose deepfakes as a case study for three reasons: 1) their classification as misinformation is more objective; 2) we can control the demographics of the personas presented; 3) deepfakes are a real-world concern with associated harms that need to be better understood. Our project presents a survey (N=2,016) where U.S.-based participants are exposed to videos and asked questions about their attributes, not knowing some might be deepfakes. Our analysis measures the extent to which different users are duped and which perceived demographics of deepfake personas tend to mislead. Importantly, we find that accuracy varies significantly by demographics, and participants are generally better at classifying videos that match them (especially for white participants). We extrapolate from these results to understand the population-level impacts of these biases using an idealized mathematical model of the interplay between diverse misinformation and crowd correction. Our model suggests that a diverse set of contacts might provide “herd correction” where friends can protect each other’s blind spots. Altogether, human biases and the attributes of misinformation matter greatly, but having a diverse social group may help reduce susceptibility to misinformation.

This project adopts a multidisciplinary approach to answer these questions and understand their impacts. In an effort to avoid assumptions about any demographic group, we chose four specific biases to explore vis-à-vis deepfakes: **(Question 1) Priming bias:** How much does classification accuracy depend on participants being primed about the potential of a video being fake? Our participants are not primed on the meaning of deepfakes and are not explicitly looking for them. **(Question 2) Homophily bias:** Are humans better classifiers of video content if the perceived demographic of the video persona matches their own identity? **(Question 3) Heterophily bias:** Inversely, are humans more accurate if the perceived demographic of the video persona does not match their own?

We present an empirical survey (N=2,016) testing what shapes participants’ ability to detect deepfake videos. Survey participants entered the study under the pretense that they would judge the communication styles of video clips. Our study is careful not to prime participants so we could gauge their ability to view and judge deepfakes when they were not expecting them. Our survey also investigates the relationship between human demographics and the video person(a)’s features and, ultimately, how this relationship impacts the participant’s ability to detect deepfake content. We use results from the survey to develop an idealized mathematical model to theoretically explore population-level dynamics of diverse misinformation on online social networks. Altogether, this allows us to hypothesize the *mechanisms* and *impacts* of diverse misinformation, as illustrated in Figure 1.

The overarching takeaway of our results can be summarized as follows: If not primed, our survey participants are not particularly accurate at detecting deepfakes (accuracy = 51%, essentially a coin toss) as seen in Table 2. Accuracy varies by some participants' demographics and perceived demographics of video persona (especially for white participants), as seen in Table 1. In general, participants were better at classifying videos that they perceived as matching their own demographic.

Our results show that of the 4,032 videos watched, 49% were deepfakes, and 1,429 successfully duped our survey participants. We also note that the overall accuracy rate (where accuracy = (TP+TN)/(TP+FP+FN+TN)) of our participants was 51%. This translates to an overall Matthew's Correlation Coefficient (MCC) score of 0.334 for all participant's guesses vs. actual states of the videos (35% of the participants were duped). MCC [Matthews, 1975, Boughorbel et al., 2017] is a simple binary correlation between the ground truth and the participant's guess. We use a bootstrap approach to then test the credibility of a superior accuracy (frequency of bootstrap pairs that produce a superior accuracy).

We integrate some of our findings into a mathematical model to understand how deepfakes spread on social networks with a diverse population. The model uses a network with a heterogeneous degree distribution and a structure inspired by the mixed-membership SBM with modules like echo chambers and bridge nodes with diverse neighborhoods. We find that more susceptible nodes benefit greatly from being bridges between communities in populations with heterogeneity in susceptibility. In a nutshell, diverse friends can correct each other's blind spots.

Understanding the structure and dynamics of misinformation is important as it can bring a great amount of societal harm [Chesney and Citron, 2019, Lazer et al., 2018]. Misinformation has negatively impacted the ability to disseminate important information during critical elections, humanitarian crises, global unrest, and global pandemics. More importantly, misinformation degrades our epistemic environment, particularly regarding distrust of truths. It is necessary to understand who is susceptible to misinformation and how it spreads on social networks to mitigate its harm and propose meaningful interventions. Further, as deepfakes deceive viewers at greater rates, it becomes increasingly critical to understand who gets duped by this form of misinformation and how our biases and social circle impact our interaction with video content at scale. We hope this work will contribute to the critical literature on human biases and help to better understand their interplay with machine-generated content.

References

[Boughorbel et al., 2017] Boughorbel, S., Jarray, F., and El-Anbari, M. (2017). Optimal classifier for imbalanced data using matthews correlation coefficient metric. *PLOS One*, 12(6):e0177678.

[Chesney and Citron, 2019] Chesney, B. and Citron, D. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107:1753.

[Groh et al., 2022] Groh, M., Epstein, Z., Firestone, C., and Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1).

[Lazer et al., 2018] Lazer, D. M., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., et al. (2018). The science of fake news. *Science*, 359(6380):1094–1096.

[Matthews, 1975] Matthews, B. W. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451.

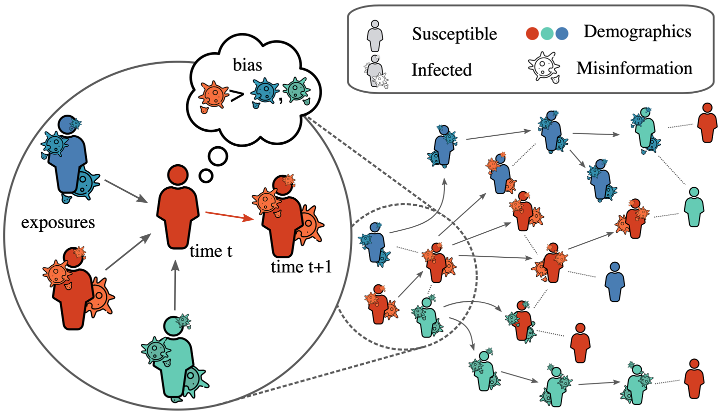

Figure 1: Illustration of the problem considered in this work. Populations are made of individuals with diverse demographic features (e.g., age, gender, race; here represented by colors), and misinformation is likewise made of different elements based on the topics they represent (here shown as pathogens). Through their biases,certain individuals are more susceptible to certain kinds of misinformation. The cartoon represents a situation where misinformation is more successful when it matches an individual's demographic. Red pathogens spread more readily around red users with red neighbors. In reality, the nature of these biases is still unclear, and so are their impacts on online social networks.

Video/User Demographics	MCC of User	N	Credibility
White Viewer/White Videos	0.0518	1372	0.99
White Viewer/Non-white Videos	-0.0498	1224	
Male Persona/Male Viewer	0.0827	918	0.97
Male Persona/Female Viewer	0.0567	1188	
POC Persona/POC Viewer	0.0858	708	0.99
POC Persona/White Viewer	-0.0544	1143	
Age 18-29 Persona/Age 18-29 Viewer	0.1475	303	0.99
Age 18-29 Persona/Age 30-49 Viewer	0.0354	264	
Age 18-29 Persona/Age 50+ Viewer	-0.0198	694	
Age 30-49 Persona/Age 18-29 Viewer	0.1168	282	0.96
Age 30-49 Persona/Age 30-49 Viewer	-0.0037	607	

Table 1: Significant (above a 95% credibility) categories of interest. Matthew's Correlation Coefficient (MCC) is a correlation measure between a participant's guess about the video being real or fake (0,1) versus the actual state of the video (real 0, fake 1). We use a bootstrap approach to then test the credibility of a superior accuracy (frequency of bootstrap pairs that produce a superior accuracy).

Type	Accuracy
Non-Primed Human	**51%**
Primed Human [Groh et al., 2022]	66%
Machine Only [Groh et al., 2022]	65%
Primed Human with Machine Helper [Groh et al., 2022]	73%

Table 2: Accuracy scores of machine deepfake detectors versus primed human deepfake detectors versus non-primed human deepfake detectors. We compare primed and non-primed survey participants and their abilities to detect deepfakes. Our results show that humans who are not primed to find deepfakes reach an accuracy of 51% (MCC Score 0.334, 35% participants duped). The accuracy scores of our survey participants are 15% points below those of primed human deepfake detectors from previous work. [Groh et al., 2022]