Lifelikeness is in the eye of the beholder: demographics of deepfake detection and their impacts on online social networks

Juniper Lovato[1], Laurent Hébert-Dufresne[1,2], Jonathan St-Onge[2], Gabriela Salazar Lopez[1], Sean P. Rogers[2], Randall Harp[1,3], Ijaz Ul Haq[2] and Jeremiah Onaolapo[1,2]

[1]Vermont Complex Systems Center, University of Vermont, Burlington, 05405, USA
[2]Department of Computer Science, University of Vermont, Burlington, 05405, USA
[3]Department of Philosophy, University of Vermont, Burlington, 05405, USA

Abstract

Deepfakes videos are becoming increasingly believable, and their pervasiveness poses critical ethical and technical concerns regarding the ability of humans to detect them. It is currently unknown to what extent our human preferences and prejudices impact human deepfake detection ability. The initial phase of our project presents a survey experiment (phase 1 of our study surveyed 1,000 participants) where people are exposed to short video clips, not knowing the content might be fake. Survey participants are sampled through a Qualtrics survey panel to match the demographics of U.S. social media users in 2021 [1]. Participants are subsequently asked to guess the demographics (e.g., age, gender) of the persona of the video and whether each video watched is real or a deepfake. We measure the accuracy rate at which survey participants of different demographic backgrounds are duped and by what types of self-similar or self-dissimilar deepfake personas. Our project explores four primary questions. (Q1) Are humans better at correctly classifying deepfake videos if primed about deepfake content before exposure? (Q2) self-similarity bias: Are there categories of humans better at detecting a deepfake video if the persona in the video matches their own identity? (Q3) self-dissimilarity bias: Are there categories of humans better at detecting a deepfake video if the persona in the video is different from their own identity? (Q4) Prior knowledge bias: Are human viewers better at detecting a deepfake video if they know more about deepfakes or use social media more frequently?

There is a growing body of work on the distributed threats in online social networks: from leaky data [2] and group privacy concerns [3] to hate speech [4], misinformation [5] and detection of computer-generated content such as deepfakes [6]. We tackle this last example by exposing human subjects to real and deepfake video clips to study the relationship between human demographics and the perceived demographics of personas depicted in video content.

To determine the variables that influence whether a participant's guess about the status of a video (real or fake) is correct, we ran a Matthew's Correlation Coefficient (MCC). MCC is typically used for classification models to test the classifier's performance. Here we are treating human participant subgroups as classifiers and are measuring their performance with MCC. Given the counts of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), the MCC is defined as

jlovato@uvm.edu (J. Lovato)

CEUR Workshop Proceedings (CEUR-WS.org)

$$\frac{(TP\times TN)-(FP\times FN)}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

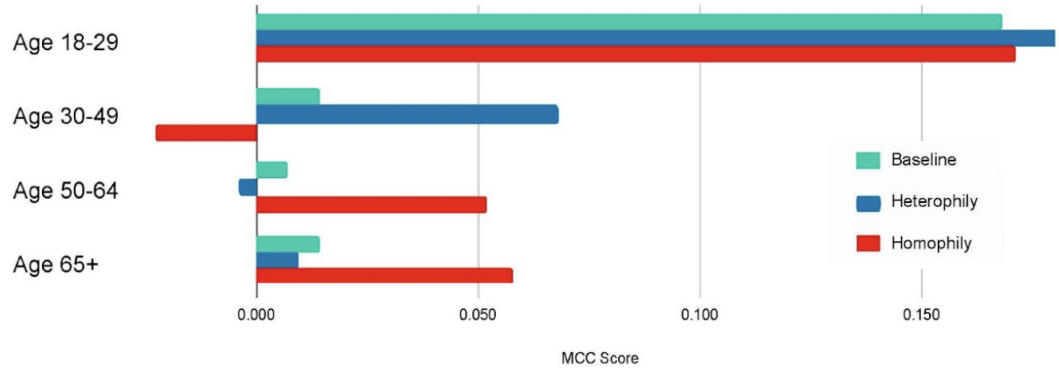


Figure 1: MCC per age category and perceived age of the video persona. Videos are separated as matching the participant's age (homophily) or not (heterophily). The baseline represents all answers from a given age group.

Out of the 1,866 total videos watched (half of the videos shown to our participants are deepfakes), 36% were deepfakes that duped our participants. The overall accuracy rate of our participants was 50% (where accuracy = (TP+TN)/(TP+FP+FN+TN) compared to 66% in previous work with primed subjects [6]. The MCC score for phase one of the study is 0.034. In Fig. 1, we break down the MCC scores based on potential homophily (self-similar) and heterophily (self-dissimilar) biases per age group.

The overarching takeaway of the survey can be summarised as follows. (Q1) If not primed, humans are not particularly accurate at detecting deepfakes. (Q2-Q3) Accuracy varies by demographics, and younger participants were much more accurate overall, but older demographics see a dramatic increase in accuracy when classifying videos that they identified as self-similar. (Q4) Accuracy increases with frequent social media usage, perhaps explaining our results from the previous question.

In essence, our results simply suggest that different age groups performed differently when classifying a video subject as either a real human or as a deepfake persona. This, in turn, implies that populations of social media users have a heterogeneous susceptibility to video misinformation. To explore the impacts of these results, 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 stochastic block model with modules like echo chambers and bridge nodes with diverse neighborhoods. We track individuals based on their demographics, here just abstract classes 1 or 2 that could represent younger and older people. We also track their state: duped (or infectious) and non-duped (or susceptible). We also track the demographics of their neighbor (tagged degrees k and ℓ) to know their role in the network. Individuals get duped by their duped neighbor at rate λ_i dependent on their demographic class i, and duped individuals can get corrected by their susceptible neighbors at rate γ. The dynamics

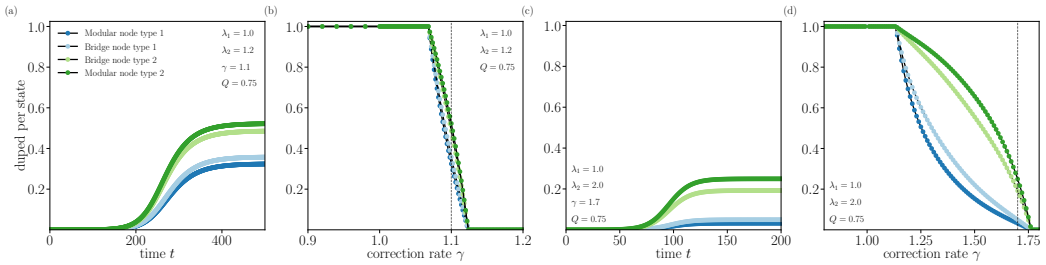
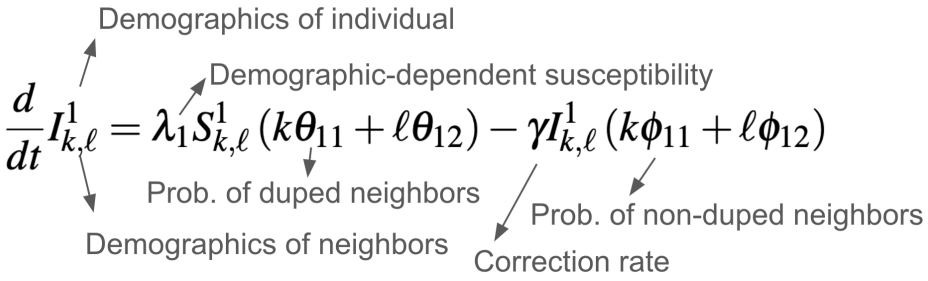


Figure 2: Spread of diverse deepfake on configurations of a degree-heterogeneous mixed membership stochastic block model with equal group size and densities (in-group density is set to Q and across the group to $1 - Q$). The degree distribution is a power-law with scale exponent -2.5, uncorrelated with demographic types. Other parameters are given in the figure. Panels (a) and (c) show the time evolution of misinformation in the system for two different parameter sets, respectively representing a medium level of misinformation ($\approx 30\%$) in a relatively homogeneous population (susceptibilities 1.0 and 1.2 respectively) and a lower level of misinformation ($\approx 10\%$). The transition from misinformation-free to misinformation-saturated populations as we vary the correction rate in these two populations is shown in panels (b) and (d), respectively.

can be tracked using a heterogeneous mean-field framework [7]:



where mean-field quantities like θ_{12} are calculated as $\sum_{k,l} kI^2_{k,l} / \sum_{k,l} k(I^2_{k,l} + S^2_{k,l})$ and represent the probability of an infected (θ) or susceptible (ϕ) neighbor given the demographics involved (indices $\{k, l\} \in \{1, 2\}$).

For a given specific set of parameters, we can use our model to ask who is affected by misinformation based on their place in the network (bridge or modular nodes) as well as their demographics (type). We show two different model runs in Fig. 2(a and c). Overall, we find that in populations with heterogeneity in susceptibility varying from small (20% variation in susceptibility) to large (100% variation), more susceptible nodes benefit greatly from being bridges between communities. In a nutshell, diverse friends can correct each other's blind spots.

Altogether, lifelikeness may not be an objective measure we can apply to all humans and lifelike personas in the same way. Humans hold a host of biases, and recognition of lifelikeness may depend on the biases and prior experiences of the viewer as well as on their social network.

Future work will incorporate the second phase of our study and model. We hope this study is a step towards understanding the impacts of social biases in an emerging societal problem with many multilevel interdependencies. We hope it will contribute to the timely literature

on the interplay between human biases and machine-generated content. As deepfakes begin to deceive viewers at greater rates, it becomes increasingly necessary to understand who gets duped by deepfakes and how our biases and those of our social circle impact our interaction with video content. Deepfakes also call into question numerous ethical issues such as the power of video evidence in legal frameworks [5]; consent of individuals featured in deepfakes [8]; bias in automated methods of deepfake detection software and deepfake training data [9]; degradation of our trust in news media and the epistemic climate, which includes issues such as misinformation [5]; and maybe even intrinsic wrongs of deepfakes themselves [10].

References

[1] Pew Research Center, Social media fact sheet, Pew Research Center: Washington, DC, USA (2021). URL: https://www.pewresearch.org/internet/fact-sheet/social-media/.

[2] J. P. Bagrow, X. Liu, L. Mitchell, Information flow reveals prediction limits in online social activity, Nature Human Behaviour 3 (2019) 122–128.

[3] J. L. Lovato, A. Allard, R. Harp, J. Onaolapo, L. Hébert-Dufresne, Limits of individual consent and models of distributed consent in online social networks, in: 2022 ACM Conference on Fairness, Accountability, and Transparency, 2022, pp. 2251–2262.

[4] J. Garland, K. Ghazi-Zahedi, J.-G. Young, L. Hébert-Dufresne, M. Galesic, Impact and dynamics of hate and counter speech online, EPJ Data Science 11 (2022) 3.

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

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

[7] R. Pastor-Satorras, A. Vespignani, Epidemic spreading in scale-free networks, Physical Review Letters 86 (2001) 3200.

[8] D. Harris, Deepfakes: False pornography is here and the law cannot protect you, Duke Law & Technology Review 17 (2018) 99.

[9] K. Haut, C. Wohn, V. Antony, A. Goldfarb, M. Welsh, D. Sumanthiran, J.-z. Jang, M. Ali, E. Hoque, et al., Could you become more credible by being white? assessing impact of race on credibility with deepfakes, arXiv (2021). arXiv:2102.08054.

[10] A. de Ruiter, The distinct wrong of deepfakes, Philosophy & Technology 34 (2021) 1311–1332.