Exercise 4: Design Recommendations for Combating Misinformation

Ian Habit

Department of Psychology, George Mason University

PSYC 530: Cognitive Engineering

Dr. William S. Helton

October 14, 2024

Misinformation Assessment for Generic Video First Social Media App:

We've completed an initial assessment of the app and have found a high network spread rate of misinformation transmission across a population of highly followed "influencer" accounts. These accounts are responsible for most of the spread of misinformation on the platform. Additionally, we tested the recommendation algorithm and found that, on average, it takes about 240 videos or around 35 minutes to become "addicted" to the personalized feed, figures reinforced by internal documents published during recent legal discovery by the State of Kentucky in its suit of your company (NPR, 2024).

Furthermore, as stated by the released Kentucky AG's report (NPR, 2024) it takes less than 20 minutes to train the app's feed to dip into echo chambers, some of which feature content encouraging self-harm, political violence, filters which reinforce unrealistic beauty standards, and sexually explicit material — content that that is psychologically damaging to adults let alone teenagers and children with relatively unrestricted access.

A recently released paper by researchers at NYU and the Norwegian School of Economics asserts that your platform increases negative perceptions of "out-groups" by virally spreading misinformation and keeping users in politically and socially charged echo chambers (Rathje et al. 2024). To combat the spread of information and protect the health of your users, and thus the compliance and continued existence of the platform in the United States, we recommend taking the following introductory actions:

1. **Clearly label misinformation and the accounts that spread it —** use written plain language descriptors, labels, and iconography to alert users that the information being presented is false or misleading or that the account itself has been flagged as a

frequent spreader or untrustworthy. This could be a transparent overlay or title header that usually accompanies the text at the bottom of the presented video. In account lists, label the flagged accounts with a warning icon and present interactions of intentional friction when someone tries to follow.

2. **Use community notes or tool tips with rational, calm tone to explain why the content is misinformation and avoid sensationalism —** Studies by Endsley (2018) have shown that providing rational, fact-based information with clear warnings can decrease the focus on sensational content and avoid biases that contribute to misinformation spread. Platforms like Threads and Twitter/X have also shown marked success with crowd-sourced community notes that provide corrections and factual sources to misleading content. Community notes also promote prosocial behavior and "shame" misinformation sources by "ratio-ing" the original poster via the number of likes or reposts.
3. **Present users the opportunity to unfollow untrustworthy accounts while suggesting neutral, fact-based and scientific "aggregator" accounts —** The study by Rathje et al. has shown that unfollowing harmful accounts acts like a "scalpel" on misinformation spread, out-group hate, and hyper-polarization. As an additional treatment, replacing the unfollowed accounts with non-political, fact and science based "aggregator" accounts decreased these network effects and user attitudes, primarily replaced by feelings of "awe" when engaging with scientific accounts. Self-reporting feeling remained 6+ months after the experimental treatment (2024). To that end, we recommend an intervention where the platform builds on our earlier recommendations of clear labeling while also presenting an opportunity for users to:

   a. unfollow harmful accounts at the point of label engagement and
   b. have the platform recommend following neutral, fact and science-based accounts

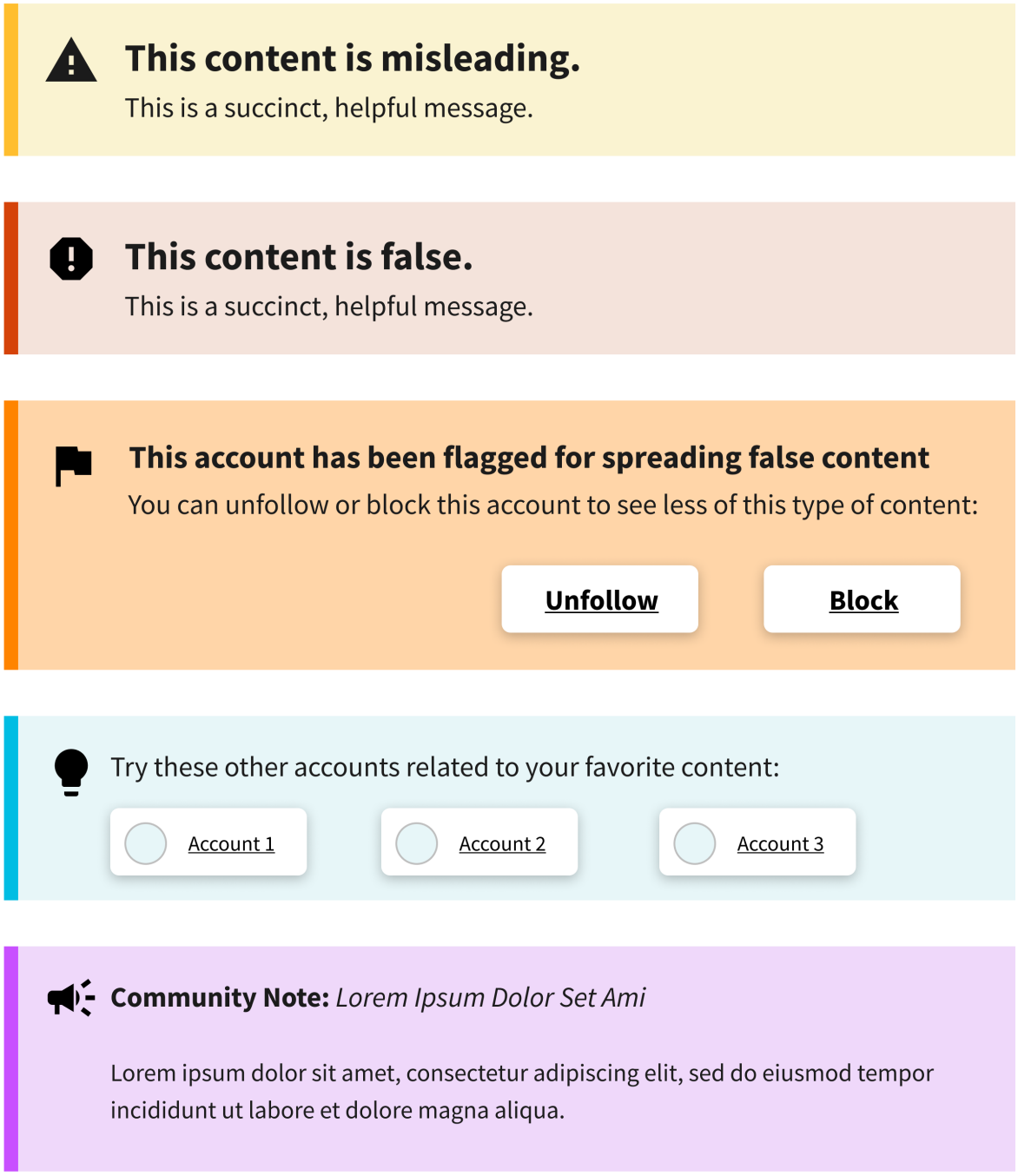
4. **Demonetize, deemphasize, and deplatform suspected misinformative accounts —** in an escalating order and with serious rigor, the platform must first demonetize suspected super-spreader accounts to remove the financial incentive for posting misinformation and seeking attention-based engagement. Next, we recommended de-emphasizing these accounts within the algorithm should the accounts continue to spread harmful and misleading content. Finally, after fair warnings and repeated engagement from your platform's Trust and Safety team, removal of the accounts and device logging at the root/device level to prevent further account creation.

We believe these are strong recommendations and represent "low-hanging fruit" development efforts that have a high return on value for minimal engineering time and cost.

Abbreviated Warning and Iconography Style Guide:

To supplement our recommendations, we have created an abbreviated warning and icon style guide that emphasizes simple forms and plain language statements that users can clearly assimilate and recognize, regardless of their engagement patterns. We want to reduce visual complexity and provide objects that are familiar (Wickens, 2022, p.230) to a wide variety of users of different cultural and socioeconomic backgrounds. Basically, the icons and associated warning messages must have broad and immediate recognition and comprehension from a majority of your users, which currently sits in the billions of MAU. Designing too many and too

complex of an icon set will confuse users and possibly lead to *more* misinformation transmission than less. *



* This style guide (warning banners and icons) was adapted from the United States Web Design System, and open-source design library for non-commercial website design or .gov domains. All written content and buttons are original and unique to this author.

References

Endsley, M.R. (2018). Combating information attacks in the age of the Internet: new challenges for cognitive engineering. *Human Factors, 60*(8), 1081 – 1094.

Rathje, S., Pretus, C., He, J, Harjani, T., Roozenbeck, J., Gray, K., van der Linden, S., & Van Bavel, J. J. (2024). Unfollowing hyperpartisan social medial influencers durably reduces out-party animosity. *Open Science Foundation*. https://osf.io/e3jfk/

States probed TikTok for years. Here are the documents the app tried to keep secret. (2024, October 11). NPR. https://www.npr.org/2024/10/11/g-s1-27676/tiktok-redacted-documents-in-teen-safety-lawsuit-revealed

Wickens, C. D., Helton, W. S., Hollands, J. G., & Banbury, S. (2022). *Engineering psychology and human performance* (5th edition). Routledge.