Combating Inaccurate Information on Social Media*
Invited Talk - Extended Abstract

Mohsen Mosleh

University of Exeter Business School
Department of Science, Innovation, Technology, and Entrepreneurship
Sloan School of Management
Massachusetts Institute of Technology, United States
mmosleh@mit.edu

There has been a great deal of concern currently about negative societal impacts of social media and the potential threats social media poses to society and democracy [5, 2]. One main area of concern is in terms of the prevalence of fundamentally low quality information and the potential of social media to facilitate the spread of misinformation and fake news. In my talk, I discuss a series of our studies that provide a potential solution that can be implemented by social media platforms at large scale to combat spread of misinformation.

First, I discuss studies examining the spread of misinformation on Twitter. I begin by describing a hybrid lab-field study in which I investigate the relationship between individual differences in cognitive reflection and behavior on Twitter in a sample of N = 1,901 users [3]. In doing so, I use the lens of cognitive science considering people decision making arising from two different modes of information processing: *i*) they may stop and carefully think about the piece of information they receive or *ii*) they just rely on their intuition and guts responses. We expect people who rely more versus less on analytical thinking demonstrate different behavior on social media platforms. To measure the extent to which one replies on intuitive gut responses versus careful thinking, I used the Cognitive Reflection Test (CRT) which is a set of questions with intuitively compelling but wrong answers. For example, "if you are running a race and you pass the person in second place what place are you in?" The intuitive answer that comes to mind for many people is first place, however, this is not the correct answer. If you pass the person in second place, you will end up being in second place. Questions of this type, captures the extent to which one says the first thing comes to mind versus stopping to think carefully before saying something.

To investigate the relationship between cognitive style and online behavior, I devised a hybrid lab-field study. In a survey study, I asked subjects to do the Cognitive Reflection Test – and also asked them to provide their Twitter handles. I used the subjects' Twitter handles to retrieve information from their public

profile on Twitter including general account information, accounts followed, and the content of their tweets. Analyzing this hybrid data set of users' CRTs (cognitive style) and their digital fingerprints using natural language processing and network science methods, I show people who give more wrong answers to the CRT are less discerning in their social media use: they follow more questionable accounts, share lower quality content from less reliable sources, and tweet about less weighty subjects (e.g., less politics). Together, these results paint a fairly consistent picture: People who engage in less cognitive reflection are more likely to consume and share low quality content.

Building off the above observation, I discuss a subtle behavioral intervention we developed to make users think before they make sharing decisions on social media [4]. We direct messaged N=5,379 Twitter users who had previously shared links to misinformation websites (in particular they shared content from hyper-partisan and low quality websites Breitbart and Infowars), and asked them to rate the accuracy of a single non-political headline - therefore making the concept of accuracy more top of mind for them, such that they would be more likely to think about accuracy when they went back to their news feed. To allow for causal inference, we used a stepped-wedge (randomized roll-out) design in which users were randomly assigned to a date on which to receive the treatment message. Within each 24-hour time-window, we then compared the links shared by users who received the treatment message at the beginning of that time window to the links shared by all the users who had not yet been messaged (who thereby represented the control condition). To quantify the quality of content shared by the users, we used a list of 60 domains (20 mainstreams, 20 hyper-partisan, and 20 fake news websites) where for each domain we had a quality score between 0 and 1 provided by 8 professional fact-checkers. As predicted, we find that the intervention leads to significant increase in the average quality of news sites shared. After receiving the message, users share proportionally more links to high-quality mainstream news outlets and proportionally fewer links to hyper-partisan low-quality news outlets as rated by professional fact-checkers. Given the complexity of the experimental design and tweet data, there are a multitude of reasonable approaches for assessing whether our intervention successfully increased the quality of news sharing. Thus, we computed effect size estimates using 198 different analysis approaches. Considering the analyses in aggregate provides strong evidence that, indeed, the accuracy message significantly increased the average quality of news sources subsequently shared by the users in our experiment. For the large majority of analytic approaches, the increase is statistically significant.

Finally, I talk about a follow-up study where instead of a subtle accuracy nudge through a private message to users, we publicly corrected those who shared misinformation on Twitter [1]. We identified N=2,000 users who shared false political news on Twitter, and replied to their false tweets with links to fact-checking websites. Unlike our subtle accuracy nudge intervention, we find causal evidence that being corrected decreases the quality, and increases the partisan slant and language toxicity, of the users' subsequent retweets (but has

no significant effect on primary tweets). This suggests that being publicly corrected by another user shifts one's attention away from accuracy - presenting an important challenge for social correction approaches.

Our experimental designs translates directly into an intervention that social media companies could deploy at scale to fight misinformation online.

References

1. Mosleh, M., Martel, C., Eckles, D., Rand, D.G.: Perverse downstream consequences of debunking: Being corrected by another user for posting false political news increases subsequent sharing of low quality, partisan, and toxic content in a twitter field experiment. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. pp. 1–11 (2021)
2. Mosleh, M., Martel, C., Eckles, D., Rand, D.G.: Shared partisanship dramatically increases social tie formation in a twitter field experiment. Proceedings of the National Academy of Sciences **118**(7) (2021)
3. Mosleh, M., Pennycook, G., Arechar, A.A., Rand, D.G.: Cognitive reflection correlates with behavior on twitter. Nature communications **12**(1), 1–10 (2021)
4. Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A.A., Eckles, D., Rand, D.G.: Shifting attention to accuracy can reduce misinformation online. Nature (2021)
5. Stewart, A.J., Mosleh, M., Diakonova, M., Arechar, A.A., Rand, D.G., Plotkin, J.B.: Information gerrymandering and undemocratic decisions. Nature **573**(7772), 117–121 (2019)