The Harvard Kennedy School Misinformation Review[1]
July 2020, Volume 1, Issue 5
Attribution 4.0 International (CC BY 4.0)
Reprints and permissions: misinforeview@hks.harvard.edu
DOI: https://doi.org/10.37016/mr-2020-033
Website: misinforeview.hks.harvard.edu

Research Article

Exposure to social engagement metrics increases vulnerability to misinformation

News feeds in virtually all social media platforms include engagement metrics, such as the number of times each post is liked and shared. We find that exposure to these signals increases the vulnerability of users to low-credibility information in a simulated social media feed. This finding has important implications for the design of social media interactions in the post-truth age. To reduce the spread of misinformation, we call for technology platforms to rethink the display of social engagement metrics. Further research is needed to investigate how engagement metrics can be presented without amplifying the spread of low-credibility information.

Authors: Mihai Avram (1,3), Nicholas Micallef (2), Sameer Patil (3, 4), Filippo Menczer (1,3)
Affiliations: (1) Observatory on Social Media, Indiana University, Bloomington, IN, USA, (2) Center for Cybersecurity, New York University, Abu Dhabi, United Arab Emirates, (3) Luddy School of Informatics, Computing, and Engineering, Indiana University Bloomington, IN, USA, (4) Tandon School of Engineering, New York University, Brooklyn, NY, USA
How to cite: Avram, M.; Micallef, N.; Patil, S.; Menczer, F., (2020). Exposure to social engagement metrics increases vulnerability to misinformation, *The Harvard Kennedy School (HKS) Misinformation Review*, 1, 5
Received: April 1st, 2020, Accepted: June 30th, 2020, Published: July 29th, 2020

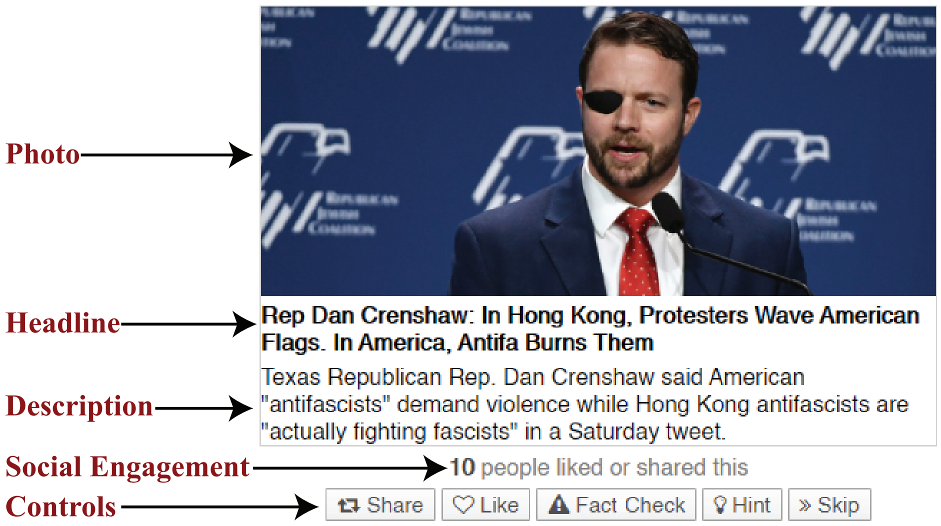
Research questions

- What effect does exposure to social engagement metrics have on people's propensity to share content?
- Does exposure to high numbers for social engagement metrics increase the chances that people will like and share questionable content and/or make it less likely that people will engage in fact checking of low-credibility sources?

Essay summary

- We investigated the effect of social engagement metrics on the spread of information from low-credibility sources using Fakey, a news literacy game that simulates a social media feed (see Figure 1). The game presents players with actual current news articles from mainstream and low-credibility media sources. A randomly generated social engagement metric is displayed with each presented article. Players are instructed to *Share*, *Like*, *Fact Check*, or *Skip* articles.
- From a 19-month deployment of the game, we analyzed game-play sessions of over 8,500 unique players, mostly from the United States, involving approximately 120,000 articles, half from sources flagged as low-credibility by news and fact-checking organizations.
- Our findings suggest that the display of social engagement metrics can strongly influence interaction with low-credibility information. The higher the engagement, the more prone people are to

[1] *A publication of the Shorenstein Center for Media, Politics, and Public Policy, at Harvard University, John F. Kennedy School of Government.*



Rep Dan Crenshaw: In Hong Kong, Protesters Wave American Flags. In America, Antifa Burns Them

Texas Republican Rep. Dan Crenshaw said American "antifascists" demand violence while Hong Kong antifascists are "actually fighting fascists" in a Saturday tweet.

10 people liked or shared this

Figure 1: A news post in the social media feed simulated by the Fakey game.

sharing questionable content and less to fact checking it.

- These findings imply that social media platforms must rethink how engagement metrics should be displayed such that they do not facilitate the spread of misinformation or hinder the spread of legitimate information. Further research is needed to guard against malicious tampering with engagement metrics at an early stage and to design educational interventions that teach users to prioritize trustworthiness of news sources over engagement metrics.

Implications

Online misinformation is a critical societal threat in the digital age, and social media platforms are a major vehicle used to spread it (Guess et al., 2019; Hameleers et al., 2020; Lazer et al., 2018). As an illustration, the International Fact-Checking Network found more than 3,500 false claims related to the coronavirus in less than 3.5 months.[2] Misinformation can cause serious societal harm in multiple ways: affecting public health (Sharma et al., 2020), influencing public policy (Lazer et al., 2018), instigating violence (Arif et al., 2018; Starbird et al., 2014), spreading conspiracies (Samory & Mitra, 2018), reducing trust in authorities (Gupta et al., 2014; Shin & Thorson, 2017; Vosoughi et al., 2018), and increasing polarization and conflict (Stewart et al., 2018).

The growing societal impact of misinformation has driven research on technical solutions to detect and stop actors that generate and spread such content. The detection techniques have leveraged network analysis (Jin et al., 2013; Ratkiewicz et al., 2011), supervised models of automated behavior (Ferrara et al., 2016; Hui et al., 2019; Varol, Ferrara, Davis, et al., 2017; K.-C. Yang et al., 2019; K.-C. Yang et al., 2020), time series analysis to detect campaigns (Varol, Ferrara, Menczer, et al., 2017), and natural language processing for flagging factually incorrect content (Kumar et al., 2016; Pérez-Rosas et al., 2018). On the user interface side, researchers have explored the use of credibility indicators to flag misinformation and alert users (Clayton et al., 2019). Such credibility indicators can lead to a reduction in sharing the flagged content (Nyhan et al., 2019; Pennycook, Bear, et al., 2020; Pennycook, McPhetres, et al., 2020; Yaqub et al., 2020).

Studies have explored the role of environmental, emotional, and individual factors that impact online misinformation spread (Coviello et al., 2014; Ferrara & Yang, 2015; Grinberg et al., 2019; Kramer et al., 2014; Yaqub et al., 2020). However, there has been little empirical research on the effects of interface ele-

[2] https://poynter.org/coronavirusfactsalliance

ments of social media feeds on the spread of misinformation (Hameleers et al., 2020; Shen et al., 2019). To address this gap, we empirically investigated how the spread of low-credibility content is affected by exposure to typical social engagement metrics, i.e., the numbers of Likes and Shares shown for a news article. Player behavior in the Fakey game shows near-perfect correlations between displayed social engagement metrics and player actions related to information from low-credibility sources. We interpret these results as suggesting that social engagement metrics amplify people's vulnerability to low-credibility content by making it less likely that people will scrutinize potential misinformation while making it more likely that they like or share it. For example, consider the recent disinformation campaign video "Plandemic" related to the COVID-19 pandemic.[3] Our results suggest that people may be more likely to endorse the video without verifying the content simply because they see that many other people liked or shared it.

To interpret these findings, consider that the probability of sharing a piece of information grows with the number of times one is exposed to it, a phenomenon called *complex contagion* (Mønsted et al., 2017; Romero et al., 2011). Social engagement metrics are proxies for multiple exposures; therefore, they are intended to provide signals about the importance, relevance, and reliability of information — all of which contribute to people's decisions to consume and share the information. In other words, users are likely to interpret high numbers for engagement metrics for an article as suggesting that it must be worthy of attention because many independent sources have validated it by liking or sharing it.

A key weakness in the cognitive processing of social engagement metrics is the assumption of independence; a malicious entity can trick people by falsely boosting engagement metrics to create the *perception* that many users endorsed an article. In fact, most disinformation campaigns rely on inauthentic social media accounts to tamper with engagement metrics, creating an initial appearance of virality that becomes reality once enough humans are deceived (Shao, Ciampaglia, et al., 2018). To prevent misinformation amplified by fake accounts from going viral, we need sophisticated algorithms capable of early-stage detection of coordinated behaviors that tamper with social engagement metrics (Hui et al., 2019; Pacheco et al., 2020; K.-C. Yang et al., 2020).

Our findings hold important implications for the design of social media platforms. Further research is needed to investigate how alternative designs of social engagement metrics could reduce their impact on misinformation sharing (e.g., by hiding engagement metrics or making them less visible for certain posts), without negatively impacting the sharing of legitimate and reliable content. A good trade-off between these two conflicting needs requires a systematic investigation of news properties that can help determine differential display of social engagement metrics. Such properties may include types of sources (e.g., unknown/distrusted accounts) and topics (e.g., highly sensitive or polarizing matters with potential for significant impact on society).

Further research is also needed to design media literacy campaigns, such as Fakey, that teach users to prioritize trustworthiness of sources over engagement metrics when consuming content on social media. Studies could explore the possibility of introducing deliberate pauses when consuming news through a social media feed (Fazio, 2020) and limiting automated or high-speed sharing. A comprehensive digital literacy approach to reduce the vulnerability of social media users to misinformation may require a combination of these interventions with additional ones, such as inoculation (Basol et al., 2020; Roozenbeek & van der Linden, 2019a, 2019b; Roozenbeek et al., 2020), civic online reasoning (McGrew, 2020), critical thinking (Lutzke et al., 2019), and examination of news feeds (Nygren et al., 2019).

[3] https://www.snopes.com/collections/plandemic/

Findings

Finding 1: High levels of social engagement result in lower fact checking and higher liking/sharing, especially for low-credibility content.

For each article shown in the game, the player is presented with a photo, a headline, a description, and a randomly generated social engagement number. Based on this information, the player can *Share*, *Like*, or *Fact Check* the article (see Figure 1). The player must *Share* or *Like* articles from mainstream sources and/or *Fact Check* articles from low-credibility sources to earn points in the game. The Methods section provides details on source selection.

We measured the correlation between the social engagement number η displayed to players and the rates at which the corresponding articles from low-credibility sources were liked/shared or fact checked by the players. Given the realistically skewed distribution of η values, we sorted the data into logarithmic bins based on the shown social engagement numbers. For each bin $\lfloor \log_{10}(\eta + 1) \rfloor$, we calculated the liking/sharing and fact-checking rates across articles and players. We measured correlations using the non-parametric Spearman test as the data is not normally distributed. For articles from low-credibility sources, we found a significant positive correlation between social engagement level and liking/sharing (Spearman $\rho = 0.97$, $p < 0.001$) and a significant negative correlation between social engagement level and fact checking (Spearman $\rho = -0.97, p < 0.001$). We found similar statistically significant relationships between social engagement level and player behavior for mainstream news article as well, however the correlations are less strong: $\rho = 0.66$ ($p < 0.001$) for liking/sharing and $\rho = -0.62$ ($p < 0.001$) for fact checking.

Finding 2: People are more vulnerable to low-credibility content that shows high levels of social engagement.

The previous finding is for the whole player population, with measures aggregated across all players. To delve further into the effect of social engagement exposure on individual players, we analyzed whether the displayed social engagement number influenced each player's liking/sharing and fact-checking rates for articles from low-credibility sources. We treated each player as an independent entity and categorized social engagement numbers into three bins: low ($0 \leq \eta < 10^2$), medium ($10^2 \leq \eta < 10^5$), and high ($10^5 \leq \eta \leq 10^6$). Within each social engagement bin, we counted the number of low-credibility articles to which each player was exposed. We then calculated the corresponding proportions of these articles that each player liked/shared or fact checked. Figure 2 plots the mean liking/sharing and fact-checking rates for low-credibility articles. Although players were more likely to fact check than like or share low-credibility content, Figure 2 shows that the trends observed at the population level held at the individual player level as well.

Since the data is not normally distributed ($p < 0.05$ using the Shapiro-Wilk test for normality), we used the Kruskal-Wallis test to compare differences among the three bins of social engagement. Liking/sharing ($\chi^2(2) = 417.14$, $p < 0.001$) and fact checking ($\chi^2(2) = 214.26$, $p < 0.001$) rates for low-credibility articles differed across bins. To determine which levels of social engagement impacted the rates at which low-credibility articles were liked/shared or fact checked, we conducted post-hoc Mann-Whitney tests across all pairs of social engagement bins, with Bonferroni correction for multiple testing. We found that liking/sharing as well as fact-checking rates were statistically significantly different across all bin pairings ($p < 0.001$).

We employed the same approach to examine liking/sharing and fact-checking rates for mainstream articles across the three bins of social engagement. Similar to low-credibility articles, Kruskal-Wallis tests revealed a statistically significant effect of social engagement level on liking/sharing ($\chi^2(2) = 161.80$, $p < 0.001$) and fact checking ($\chi^2(2) = 576.37$, $p < 0.001$) rates for mainstream articles.

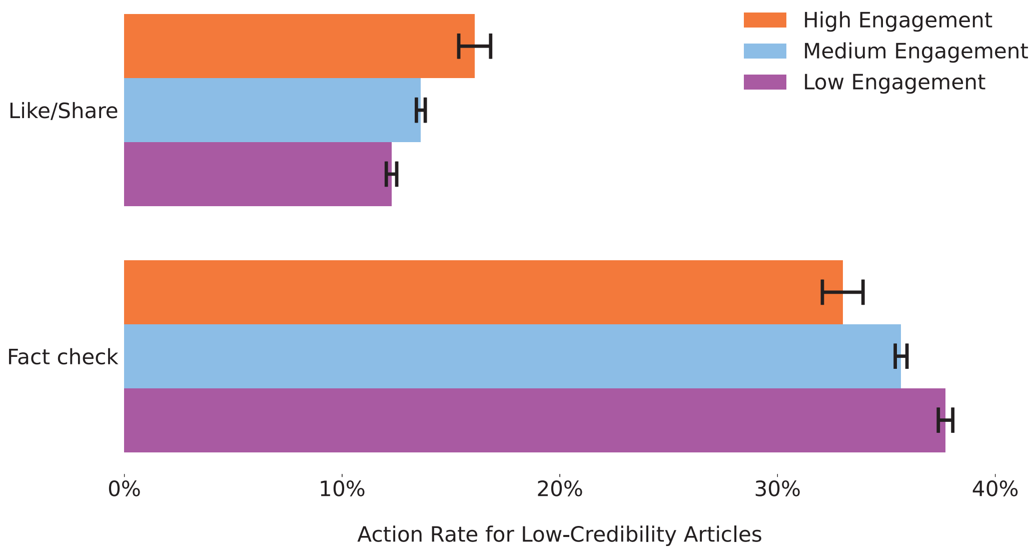


Figure 2: Mean rates of liking/sharing and fact checking low-credibility articles for low, medium, and high social engagement levels. Error bars indicate the standard error.

Methods

To investigate the effect of exposure to social engagement metrics on susceptibility to questionable content, we developed and deployed Fakey,[4] an online news literacy game.

Social media simulation

Fakey simulates fact checking on a social media feed. The user interface of Fakey mimics the appearance of either Facebook or Twitter feeds for players who log into the game through the respective platforms. The game provides players with batches of ten news articles in the form of a news feed, as shown in Figure 1. Each article consists of elements typically displayed by popular social media platforms: photo, headline, description, and social engagement.

For each article from mainstream as well as low-credibility sources, the game displays a single social engagement metric indicating the combined number of Likes and Shares. Having a single metric decreases the cognitive workload for players and simplifies the experimental design. The metric uses social engagement values drawn randomly from an approximately log-normal distribution with a maximum possible value (cutoff) of $\eta = 10^6$. The distribution is such that roughly 69% of the articles display engagement values $\eta > 10^2$ and roughly 3% display values $\eta > 10^5$. Although the engagement metric simulated in the game is not drawn from empirical data, the randomly generated metric numbers have a heavy tail similar to those typically observed on social media platforms (Vosoughi et al., 2018).

Below each article is a set of buttons to *Share*, *Like*, *Fact Check*, or *Skip* the article or use a *Hint* (see Figure 1). Before playing the game, players are instructed that *Share* is equivalent to endorsing an article and sharing it with the world, *Like* is equivalent to endorsing the article, and *Fact Check* is a signal that the article is not trusted. After playing each round of ten articles, players have the option to play another round or check a leaderboard to compare their skill points with those of other players.

[4] https://fakey.iuni.iu.edu/

Content selection

We followed the practice of analyzing content credibility at the domain (website) level rather than the article level (Bovet & Makse, 2019; Grinberg et al., 2019; Lazer et al., 2018; Pennycook & Rand, 2019; Shao, Ciampaglia, et al., 2018; Shao, Hui, et al., 2018). Each article in the game is selected from one of two types of news sources: mainstream and low-credibility.

For mainstream news, we manually selected 32 sources with a balance of moderate liberal, centrist, and moderate conservative views: *ABC News Australia, Al Jazeera English, Ars Technica, Associated Press, BBC News, Bloomberg, Business Insider, Buzzfeed, CNBC, CNN, Engadget, Financial Times, Fortune, Mashable, National Geographic, New Scientist, Newsweek, New York Magazine, Recode, Reuters, Techcrunch, The Economist, The Guardian, The Independent, The New York Times, The Next Web, The Telegraph, The Verge, The Wall Street Journal, The Washington Post, Time, and USA Today.* The game obtains mainstream news articles via the News API.[5]

We selected the set of low-credibility sources based on flagging by various reputable news and fact-checking organizations (Shao, Ciampaglia, et al., 2018; Shao, Hui, et al., 2018). The selected low-credibility sources tend to publish fake news, conspiracy theories, clickbait, rumors, junk science, and other types of misinformation. The game uses the Hoaxy API[6] to retrieve articles from these low-credibility sources.

For each round, the game randomly selects five articles each from mainstream and low-credibility sources. For a given source, any article returned by the News or Hoaxy API is shown to the player regardless of topic, without further selection or filtering except for ensuring that the same story is not shown to the same player multiple times across rounds.

Data collection

The game is available online through a standard web interface and as a mobile app via the Google Play Store and the Apple App Store. The mobile app is available in the following English-speaking countries: Australia, Canada, United Kingdom, and United States. People from other countries can play the game on the web.

Our analysis uses data from a 19-month deployment of the game, between May 2018 and November 2019. During this period, we advertised the game through several channels, including social media (Twitter and Facebook), press releases, conferences, keynote presentations, and word of mouth. We recorded game sessions involving approximately 8,606 unique players[7] and 120,000 news articles, approximately half of which were from low-credibility sources. We did not collect demographic information, but we collected anonymous data from Google Analytics embedded by the game's hosting service. Players originated from the United States (78%), Australia (8%), UK (4%), Canada (3%), Germany (3%), and Bulgaria (2%).

Limitations

Our news literacy game emulates relevant interface elements of popular social media platforms, such as Facebook and Twitter, without raising ethical concerns of real-world content manipulation (Kramer et al., 2014). Yet, conducting the study in a simulated game environment rather than an actual platform presents clear limitations as the experience and context are not identical. For example, we limited cognitive burden on players by capturing only *Like* and *Share* actions; these were the earliest ones deployed on social media platforms and, as such, are the most common across platforms and the most familiar to users.

[5] https://newsapi.org

[6] http://rapidapi.com/truthy/api/hoaxy

[7] We used recorded analytics to aggregate anonymous sessions by the same person. However, the aggregation approach cannot ascribe anonymous sessions to a single person with complete certainty. Therefore, we cannot provide a precise number of unique players.

The even mix of articles from mainstream and low-credibility sources is not necessarily representative of the proportion of misinformation to which social media users are exposed in the wild. Further, the fact-checking framing of the game primes players to *expect* misinformation, potentially making it more likely to be spotted. These factors might make players more suspicious within the game compared to the real world, correspondingly increasing fact-checking rates. However, there is no reason to believe that these factors impact our results regarding the influence of social engagement metrics.

While this study is focused on user interaction elements, other factors related to users and content can affect the spread of questionable content. To respect privacy, we chose not to collect any player information apart from game analytics. However, knowledge about the background of the players (e.g., education, demographics, political affiliation) might provide further insight into vulnerability to misinformation. Similar refinements in insight would be provided by examining whether certain types of content are more likely to be influenced by social engagement metrics. These are important avenues for future research.

Acknowledgments

We are grateful to Alessandro Flammini for inspiring the social engagement exposure experiment, Chengcheng Shao for the development of the Hoaxy API, and Clayton A. Davis for technical support during the implementation of the Fakey game.

Bibliography

Arif, A., Stewart, L. G., & Starbird, K. (2018). Acting the part: Examining information operations within #BlackLivesMatter discourse. *Proc. ACM Hum.-Comput. Interact.*, *2*(CSCW). https://doi.org/10.1145/3274289

Basol, M., Roozenbeek, J., & van der Linden, S. (2020). Good news about bad news: Gamified inoculation boosts confidence and cognitive immunity against fake news. *Journal of cognition*, *3*(1), 2–2. https://doi.org/10.5334/joc.91

Bovet, A., & Makse, H. A. (2019). Influence of fake news in Twitter during the 2016 US presidential election. *Nature Communications*, *10*(1), 7. https://doi.org/10.1038/s41467-018-07761-2

Clayton, K., Blair, S., Busam, J. A., Forstner, S., Glance, J., Green, G., Kawata, A., Kovvuri, A., Martin, J., Morgan, E., Sandhu, M., Sang, R., Scholz-Bright, R., Welch, A. T., Wolff, A. G., Zhou, A., & Nyhan, B. (2019). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior*. https://doi.org/10.1007/s11109-019-09533-0

Coviello, L., Sohn, Y., Kramer, A. D. I., Marlow, C., Franceschetti, M., Christakis, N. A., & Fowler, J. H. (2014). Detecting emotional contagion in massive social networks. *PLOS ONE*, *9*(3), 1–6. https://doi.org/10.1371/journal.pone.0090315

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*, *1*(2). https://doi.org/10.37016/mr-2020-009

Ferrara, E., Varol, O., Davis, C., Menczer, F., & Flammini, A. (2016). The rise of social bots. *Commun. ACM*, *59*(7), 96–104. https://doi.org/10.1145/2818717

Ferrara, E., & Yang, Z. (2015). Measuring emotional contagion in social media. *PLOS ONE*, *10*(11), 1–14. https://doi.org/10.1371/journal.pone.0142390

Grinberg, N., Joseph, K., Friedland, L., Swire-Thompson, B., & Lazer, D. (2019). Fake news on Twitter during the 2016 U.S. presidential election. *Science*, *363*(6425), 374–378. https://doi.org/10.1126/science.aau2706

Guess, A., Nagler, J., & Tucker, J. (2019). Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Science Advances*, *5*(1). https://doi.org/10.1126/sciadv.aau4586

Gupta, A., Kumaraguru, P., Castillo, C., & Meier, P. (2014). TweetCred: Real-time credibility assessment of content on Twitter. In L. M. Aiello & D. McFarland (Eds.), *Social informatics: 6th international conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings* (pp. 228–243). Cham, Springer International Publishing. https://doi.org/10.1007/978-3-319-13734-6_16

Hameleers, M., Powell, T. E., Meer, T. G. V. D., & Bos, L. (2020). A picture paints a thousand lies? The effects and mechanisms of multimodal disinformation and rebuttals disseminated via social media. *Political Communication*, *37*(2), 281–301. https://doi.org/10.1080/10584609.2019.1674979

Hui, P.-M., Yang, K.-C., Torres-Lugo, C., Monroe, Z., McCarty, M., Serrette, B. D., Pentchev, V., & Menczer, F. (2019). Botslayer: Real-time detection of bot amplification on Twitter. *Journal of Open Source Software*, *4*(42), 1706. https://doi.org/10.21105/joss.01706

Jin, F., Dougherty, E., Saraf, P., Cao, Y., & Ramakrishnan, N. (2013). Epidemiological modeling of news and rumors on Twitter, In *Proceedings of the 7th workshop on social network mining and analysis*, Chicago, Illinois, Association for Computing Machinery. https://doi.org/10.1145/2501025.2501027

Kramer, A. D. I., Guillory, J. E., & Hancock, J. T. (2014). Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, *111*(24), 8788–8790. https://doi.org/10.1073/pnas.1320040111

Kumar, S., West, R., & Leskovec, J. (2016). Disinformation on the web: Impact, characteristics, and detection of Wikipedia hoaxes, In *Proceedings of the 25th international conference on World Wide Web*, Montréal, Québec, Canada, International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/2872427.2883085

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., Metzger, M. J., Nyhan, B., Pennycook, G., Rothschild, D., Schudson, M., Sloman, S. A., Sunstein, C. R., Thorson, E. A., Watts, D. J., & Zittrain, J. L. (2018). The science of fake news. *Science*, *359*(6380), 1094–1096. https://doi.org/10.1126/science.aao2998

Lutzke, L., Drummond, C., Slovic, P., & Árvai, J. (2019). Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. *Global Environmental Change*, *58*, 101964. https://doi.org/10.1016/j.gloenvcha.2019.101964

McGrew, S. (2020). Learning to evaluate: An intervention in civic online reasoning. *Computers & Education*, *145*, 103711. https://doi.org/10.1016/j.compedu.2019.103711

Mønsted, B., Sapieżyński, P., Ferrara, E., & Lehmann, S. (2017). Evidence of complex contagion of information in social media: An experiment using Twitter bots. *PLOS ONE*, *12*(9), 1–12. https://doi.org/10.1371/journal.pone.0184148

Nygren, T., Brounéus, F., & Svensson, G. (2019). Diversity and credibility in young people's news feeds: A foundation for teaching and learning citizenship in a digital era. *Journal of Social Science Education*, *18*(2), 87–109.

Nyhan, B., Porter, E., Reifler, J., & Wood, T. J. (2019). Taking fact-checks literally but not seriously? The effects of journalistic fact-checking on factual beliefs and candidate favorability. *Political Behavior*. https://doi.org/10.1007/s11109-019-09528-x

Pacheco, D., Hui, P.-M., Torres-Lugo, C., Truong, B. T., Flammini, A., & Menczer, F. (2020). Uncovering coordinated networks on social media. *CoRR*, *abs/2001.05658*. https://arxiv.org/abs/2001.05658

Pennycook, G., Bear, A., Collins, E. T., & Rand, D. G. (2020). The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science*. https://doi.org/10.1287/mnsc.2019.3478

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention [PMID: 32603243]. *Psychological Science*, 0956797620939054. https://doi.org/10.1177/0956797620939054

Pennycook, G., & Rand, D. G. (2019). Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, *116*(7), 2521–2526. https://doi.org/10.1073/pnas.1806781116

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., & Mihalcea, R. (2018). Automatic detection of fake news, In *Proceedings of the 27th international conference on computational linguistics*, Santa Fe, New Mexico, USA, Association for Computational Linguistics. https://www.aclweb.org/anthology/C18-1287

Ratkiewicz, J., Conover, M. D., Meiss, M., Gonçalves, B., Flammini, A., & Menczer, F. M. (2011). Detecting and tracking political abuse in social media, In *Fifth international AAAI conference on weblogs and social media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM11/paper/view/2850/3274

Romero, D. M., Meeder, B., & Kleinberg, J. (2011). Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on Twitter, In *Proceedings of the 20th international conference on World Wide Web*, Hyderabad, India, Association for Computing Machinery. https://doi.org/10.1145/1963405.1963503

Roozenbeek, J., & van der Linden, S. (2019a). Fake news game confers psychological resistance against online misinformation. *Palgrave Communications*, *5*(1), 65. https://doi.org/10.1057/s41599-019-0279-9

Roozenbeek, J., & van der Linden, S. (2019b). The fake news game: Actively inoculating against the risk of misinformation. *Journal of Risk Research*, *22*(5), 570–580. https://doi.org/10.1080/13669877.2018.1443491

Roozenbeek, J., van der Linden, S., & Nygren, T. (2020). Prebunking interventions based on "inoculation" theory can reduce susceptibility to misinformation across cultures. *Harvard Kennedy School Misinformation Review*, *1*(2). https://doi.org/10.37016/mr-2020-00

Samory, M., & Mitra, T. (2018). Conspiracies online: User discussions in a conspiracy community following dramatic events, In *Twelfth international AAAI conference on web and social media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17907/17025

Shao, C., Ciampaglia, G. L., Varol, O., Yang, K.-C., Flammini, A., & Menczer, F. (2018). The spread of low-credibility content by social bots. *Nature Communications*, *9*(1), 4787. https://doi.org/10.1038/s41467-018-06930-7

Shao, C., Hui, P.-M., Wang, L., Jiang, X., Flammini, A., Menczer, F., & Ciampaglia, G. L. (2018). Anatomy of an online misinformation network. *PLOS ONE*, *13*(4), 1–23. https://doi.org/10.1371/journal.pone.0196087

Sharma, K., Seo, S., Meng, C., Rambhatla, S., Dua, A., & Liu, Y. (2020). COVID-19 on social media: Analyzing misinformation in Twitter conversations. *CoRR*, *abs/2003.12309*. https://arxiv.org/abs/2003.12309

Shen, C., Kasra, M., Pan, W., Bassett, G. A., Malloch, Y., & O'Brien, J. F. (2019). Fake images: The effects of source, intermediary, and digital media literacy on contextual assessment of image credibility online. *New Media & Society*, *21*(2), 438–463. https://doi.org/10.1177/1461444818799526

Shin, J., & Thorson, K. (2017). Partisan selective sharing: The biased diffusion of fact-checking messages on social media. *Journal of Communication*, *67*(2), 233–255. https://doi.org/10.1111/jcom.12284

Starbird, K., Maddock, J., Orand, M., Achterman, P., & Mason, R. M. (2014). Rumors, false flags, and digital vigilantes: Misinformation on Twitter after the 2013 Boston marathon bombing. *iConference 2014 Proceedings*. https://doi.org/10.9776/14308

Stewart, L. G., Arif, A., & Starbird, K. (2018). Examining trolls and polarization with a retweet network, In *Proc. ACM WSDM, Workshop on misinformation and misbehavior mining on the web*.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., & Flammini, A. (2017). Online human-bot interactions: Detection, estimation, and characterization, In *Eleventh international AAAI conference on web and social media*. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15587/14817

Varol, O., Ferrara, E., Menczer, F., & Flammini, A. (2017). Early detection of promoted campaigns on social media. *EPJ Data Science*, *6*(1), 13. https://doi.org/10.1140/epjds/s13688-017-0111-y

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science*, *359*(6380), 1146–1151. https://doi.org/10.1126/science.aap9559

Yang, K.-C., Varol, O., Davis, C. A., Ferrara, E., Flammini, A., & Menczer, F. (2019). Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, *1*(1), 48–61. https://doi.org/10.1002/hbe2.115

Yang, K.-C., Varol, O., Hui, P.-M., & Menczer, F. (2020). Scalable and generalizable social bot detection through data selection, In *Proceedings of the AAAI conference on artificial intelligence*. https://doi.org/10.1609/aaai.v34i01.5460

Yaqub, W., Kakhidze, O., Brockman, M. L., Memon, N., & Patil, S. (2020). Effects of credibility indicators on social media news sharing intent, In *Proceedings of the 2020 CHI conference on human factors in computing systems*, Honolulu, HI, USA, Association for Computing Machinery. https://doi.org/10.1145/3313831.3376213

Funding

This research is supported in part by the Democracy Fund, Craig Newmark Philanthropies, and the Knight Foundation. The sponsors had no role in study design, data collection and analysis, and article preparation and publication. The contents of the article are solely the work of the authors and do not reflect the views of the sponsors.

Competing interests

The authors have no competing interests.

Ethics

The study mechanisms and procedures reported in this article were reviewed and approved by the Institutional Review Board (IRB) of Indiana University Bloomington.

Copyright

Data Availability

All materials needed to replicate this study are available via the Harvard Dataverse https://doi.org/10.7910/DVN/OPMIS4.