Social Network Analysis and Mining (2021) 11:32
https://doi.org/10.1007/s13278-021-00739-x

ORIGINAL ARTICLE

Use of bot and content flags to limit the spread of misinformation among social networks: a behavior and attitude survey

Candice Lanius[1] · Ryan Weber[1] · William I. MacKenzie Jr.[1]

Received: 16 October 2020 / Revised: 6 February 2021 / Accepted: 15 February 2021 / Published online: 12 March 2021

Abstract

The COVID-19 infodemic is driven partially by Twitter bots. Flagging bot accounts and the misinformation they share could provide one strategy for preventing the spread of false information online. This article reports on an experiment ($N = 299$) conducted with participants in the USA to see whether flagging tweets as coming from bot accounts and as containing misinformation can lower participants' self-reported engagement and attitudes about the tweets. This experiment also showed participants tweets that aligned with their previously held beliefs to determine how flags affect their overall opinions. Results showed that flagging tweets lowered participants' attitudes about them, though this effect was less pronounced in participants who frequently used social media or consumed more news, especially from Facebook or Fox News. Some participants also changed their opinions after seeing the flagged tweets. The results suggest that social media companies can flag suspicious or inaccurate content as a way to fight misinformation. Flagging could be built into future automated fact-checking systems and other misinformation abatement strategies of the social network analysis and mining community.

Keywords Twitter · Misinformation · COVID-19 · Fact-checking · Survey study

1 Introduction

The COVID-19 crisis, which led to much of social life migrating online, has contributed to an infodemic, where information of varying quality quickly spreads in social media networks around the world. While ideally high-quality health information would be shared from credible sources and saturate social networks, misinformation about COVID-19 poses a significant public health risk during a global pandemic (O'Connor and Murphy 2020; Barau et al. 2020). As World Health Organization Director General Tedros Adhanom Ghebreyesus remarked, "We're not just fighting a pandemic; we're fighting an infodemic" (The Lancet Infectious Diseases 2020). Twitter provides a significant source of COVID-19 misinformation (Yang et al. 2020). In one analysis, almost 25% of COVID-19-related tweets contained some misinformation (Kouzy et al. 2020). Much of this misinformation spreads through bots, automated accounts that often share false or conspiracy-based information in order to amplify a political message. According to one analysis of COVID-19 information on Twitter, bot accounts share a high volume of tweets linking to low-credibility sources (Yang et al. 2020). Analysis has also revealed that "*high bot score accounts* are used to promote political conspiracies and divisive hashtags alongside with COVID-19 content" (Ferrara 2020 p. 17), while accounts likely run by humans focus more on health and public welfare.

The spread of misinformation on Twitter has been noted with alarm by scholars even before COVID-19 (Zubiaga and Ji 2014; Waszak et al. 2018; Sommariva et al. 2018). When reviewing the state of medical information on social media, Wang et al. (2019) conclude that "misinformation is abundant on the internet and is often more popular than accurate information" (p. 7), while Chen and colleagues (2018) found that medical misinformation spread more broadly on Twitter than accurate information. For instance, misleading information about Zika on Twitter was more popular than accurate posts (Sharma et al. 2017). With the rise in AI, bots spread much of this disinformation, often

✉ William I. MacKenzie Jr.
william.mackenzie@uah.edu

Candice Lanius
candice.lanius@uah.edu

Ryan Weber
ryan.weber@uah.edu

[1] University of Alabama in Huntsville, Huntsville, AL, USA

contributing significantly to the spread of low-credibility content (Shao et al. 2018). Twitter bots played "a disproportionate role in spreading and repeating misinformation" about the U.S. presidential election 2016 (Shao et al. 2017, p. 1), hold a "small but strategic role in Venezuelan political conversations" (Forelle et al. 2015 p. 1), and retweet anti-vaccination information (Broniatowski et al. 2018), especially to receptive users (Yuan et al. 2019).

Correcting Twitter misinformation remains a huge public health challenge, both in and beyond the COVID-19 crisis. This challenge exists because people are notoriously difficult to persuade when they hold false or conspiratorial beliefs (Gruzd and Mai 2020; Rice 2020) and because some analyses suggests that on Twitter, "COVID-19 misinformed communities are denser, and more organized than informed communities, with a possibility of a high volume of the misinformation being part of disinformation campaigns" (Memon and Carley 2020 p. 1). Scholars have increased calls for research into combating misinformation online. Chou et al. (2018) called for research that develops and tests interventions in response to online misinformation. According to Pagato et al. (2019), research must address the following questions: "How does health (mis)information spread, how does it shape attitudes, beliefs and behavior, and what policies or public health strategies are effective in disseminating legitimate health information while curbing the spread of health misinformation?" (2019, p. 1). Wei et al. (2016) describe the challenges that "undesirable users" create for using Twitter as a medium for understanding the "cultural landscape" and helping the response to important events and crises (p. 51).

Misinformation is often defined in a way that allows for its automatic detection. Dhar et al. (2016) describe misinformation as a rumor; pushing that definition further, Tsugawa and Ohsaki (2017) identify misinformation with the concept of "flaming" where falsehoods become viral when expressed in negative terms; by using a sentiment analysis, Tsugawa and Ohsaki then identified possible misinformation. Dewan and Kumaraguru (2017), on the other hand, focused on the motives of those who shared the misinformation, describing it as a tool of cybercriminals perpetuating a scam or hoax. Another approach to identifying misinformation uses automated fact-checking, which focuses on a direct comparison of the message to a known, credible outside source. Thorne and Vlachos (2018), using this definition, look at the state of natural language processing and journalistic sources to see where there are gaps in the automated fact-checking process. Each definition provides benefits for the automated identification and tracking of misinformation to monitor the health of social networks. While the work of social network analysis and mining scholars is of great importance for addressing the COVID-19 infodemic, the second step in addressing misinformation in social networks is what is done once a message (tweet, FB post, etc.) is identified as a problem.

Many social network platforms like Facebook, particularly those located in societies which emphasize the importance of freedom of expression, may feel uncomfortable outright banning or censoring posts (Kang and Isaac 2019). Instead, flagging posts from a questionable source or flagging information that is known to *miss* what credible sources are saying is a common approach. Yet, does flagging misinformation or a questionable source sway social media users if they already believe the information being flagged?

In response to these calls and the special theme of this issue, which asks for strategies to mitigate and fact check COVID-19 misinformation, this article reports on a novel, branching survey experiment ($N = 299$) that tested how participants responded to tweets featuring conspiracy theories about the official count of COVID-19 deaths in the United States. Participants first viewed a tweet that aligned with their existing opinion about the COVID-19 death tallies and then saw the same tweet with a flag indicating that the tweet was generated by a bot and then saw a flag warning that the tweet contained false information. The results suggest that both flags significantly decrease participants' willingness to engage with tweets and may change some participants' minds about COVID-19 misinformation. Social media platforms can use this information in their approaches to help combat a COVID-19 infodemic. This finding is an important contribution to social network analysis and mining so that the warnings from automated detection techniques can be crafted into persuasive messages that will motivate users to be cautious during the COVID-19 infodemic.

2 Literature review

2.1 Human perception of messages shared by bots

People tend to trust content attributed to AI authors less than they trust content attributed to humans (Waddell 2018). This makes sense, as users often rely on the authority of a Twitter account to separate reliable and unreliable information (Zubiaga and Ji 2014). However, studies tend to find that people only mistrust AI-generated content under certain conditions. Readers did not assign higher credibility scores to human-written vs. bot-written news articles when they did not know who wrote the story, but they considered stories labeled as written by humans more credible and readable (Graefe and Bohlken 2020). Adding low-confidence indicators to AI-generated content decreases participant trust, but high-confidence indicators do not increase trust (Bruzzese et al. 2020). Research of participants who viewed tweets labeled as coming from either a CDC Twitterbot or a human working at the CDC found that "a Twitterbot is perceived as

a credible source of information" (Edwards et al. 2014, p. 374). Participants gave similar credibility scores for a set of 10 Airbnb profiles regardless of whether they thought they were human or computer generated; however, when participants engaged with a set of 10 profiles and received information that some of the profiles were human generated and some were AI generated, they gave lower trustworthiness scores to profiles they assumed were AI generated (Jakesch et al. 2019).

2.2 Correcting misinformation on social media

Many studies find that interventions to correct misinformation on Twitter work to reduce misperceptions. Giving people accuracy nudges before they consider sharing COVID-19-related information significantly improves their truth discernment, suggesting that "nudging people to think about accuracy is a simple way to improve choices about what to share on social media" (Pennycook et al. 2020). Labeling information as rumor caused participants to consider it less important than information labeled as news (Oh and Lee 2019). Correcting misinformation about the Zika virus on Twitter by providing a source lowered misperceptions in participants (Vraga and Bode 2017a, b), as did correcting conspiracy theories about Zika (Lyons et al. 2019). Corrections can be effective coming from either algorithms or other platform users and can even affect individuals with high levels of conspiracy beliefs (Bode and Vraga 2017). WhatsApp messages from civil society organizations in Zimbabwe can correct COVID-19 misperceptions and affect positive changes in social distancing behavior (Bowles et al. 2020). Corrections from government agencies were more effective than corrections from other users (Vraga and Bode 2017a, b; van der Meer and Jin 2020), though other research has found that comments about Twitter content being fake news were more effective coming from other users than as a disclaimer from a social media platform (Colliander 2019). In an experimental situation where participants saw a fake news story on Facebook about a nonprofit organization along with a refutation from the nonprofit, denial created higher credibility for the nonprofit than comments attacking the source of the fake news (Vafeiadis et al. 2019).

However, attempts to correct misinformation can sometimes work against their intended effect (Lewandosky et al. 2012), especially in individuals who accept conspiracy theories (Miller et al. 2016). In two experiments designed to combat Zika and yellow fever misinformation in Brazil, Carey et al. (2020) found partial success for interventions to correct health myths, but also concluded that "current approaches to combating misinformation and conspiracy theories about disease epidemics and outbreaks may be ineffective or even counterproductive" (p. 9). A meta-analysis of attempts to correct misinformation online (Wang et al. 2019) finds that "although interventions to correct misperceptions are proven effective at times, efforts to retract misinformation need to be carried out with caution in order to prevent backfiring" (p. 7).

2.3 Research gap

The experiment reported here contributes to this ongoing investigation of methods for best countering and correcting the spread of misinformation on social media. Specifically, we make two unique contributions to this effort. First, while most research randomly assigns participants into experimental groups, this study assigned participants to conditions based on their previous beliefs about COVID-19 misinformation. Participants who believed COVID-19 death tallies were over- or undercounted saw tweets confirming their beliefs. (Those with no opinion or who felt the counts were accurate were randomly sorted into the over- or undercounted groups.) This approach allowed us to test the impact of flags on audiences sympathetic to the misinformation in the tweets and also allowed us to more directly test for backfire effects sometimes associated with message correction. This methodology specifically responds to calls by Lewandosky et al. (2012) to test whether retractions "fail to reduce reliance on misinformation specifically among people for whom the retraction violates personal belief" (p. 118) and Wang (2019) to understand "the role of belief systems on the intention to spread misinformation" (p. 1). Second, the experiment employs a sequence of two flags, telling users that the tweet is a suspected bot and then informing users that the tweet contains misinformation. This approach allows us to test the influence of these flags individually and together and represents a more sustained fact-checking approach.

2.4 Research questions

This study answers the following research questions:

RQ1 Does a flag that the tweet is written by a bot change participants' engagement with and attitudes about the tweet?
RQ2 Do flags that the tweet is both written by a bot and contains misinformation change participant's engagement with and attitudes about the tweet?
RQ3 Are participants capable of changing their opinion after viewing flags that a tweet was shared by a bot and contained misinformation?
RQ4 What personal experiences and attitudes are associated with the respondent's willingness to change their

opinion about coronavirus numbers after viewing the flagged tweets?

3 Method

3.1 Research context

Data were collected over a three-day period from September 8, 2020, through September 10, 2020. On September 10, 2020, the USA had a total of 6,366,986 cases of COVID-19, including 183,950 reported deaths from the virus (US Historical Data). At the time data were collected for this study, there was no scientific evidence indicating these numbers were incorrect. However, posts sharing false and misleading information, including suggestions that official COVID-19 numbers from the CDC were being either over- or underreported, were abundant on social media platforms such as Facebook and Twitter during this time period (Ebrahimji 2020; Kouzy et al. 2020). In the month prior to the current study, the then current President of the United States shared a tweet that was removed by Twitter for reporting false coronavirus statistics (Quinn 2020). The removed tweet claimed the CDC had quietly updated coronavirus numbers and suggested prior COVID-19 deaths were being overreported.

3.2 Participants

We collected a total of 332 initial responses from participants using Amazon's Mechanical Turk (MTurk) between September 8, 2020 and September 10, 2020. We choose MTurk for recruitment because its participants have been found representative of the general US population (Levay et al. 2016; McCredie and Morey 2019; Redmiles et al. 2019), especially the general Internet-using population (Keith et al. 2017). Further, MTurk participants tend to accept and take seriously experimental conditions at roughly the same rate as lab experiment participants (Thomas and Clifford 2017). According to Ford (2017), one major problem with MTurk results comes from "speeders," participants who rush through answers to get paid as quickly as possible. To help combat this issue, our survey included three attention check questions embedded throughout the survey where participants were asked to select a specific response option. We dropped 33 subjects from our final dataset for failing to correctly respond to all three attention check questions, resulting in a final sample of 299 individuals.

The final sample was on average 35 years old ($M = 35.49$, $SD = 10.03$), primarily male (59.9% male, 39.5% female, 0.6% other or prefer not to say), and White (White = 75.95%, Asian = 10.0%, Black/African-American = 7.0%, Hispanic/Latinx = 2.3%, Native American = 1.7%, biracial = 2.1%, and other = 0.7%). Our sample is a good reflection of Twitter's users: 30.9% are between the age of 25 and 34, and the majority of Twitter's users are male (Clement 2020a, 2020b).

3.3 Survey instrument

Participants were first presented with four statements and asked to select the one that best described their view of coronavirus case reporting data from the U.S. federal government. The four statements were: (1) there is underreporting—actual numbers are higher than reported numbers; (2) there is overreporting—actual numbers are lower than reported numbers; (3) there is accurate reporting—actual numbers are consistent with reported numbers; and (4) I do not have an opinion regarding coronavirus numbers. Participant responses to this question were used to assign participants to either the overreporting tweet condition or the underreporting tweet condition. One novel aspect of this study is that we presented respondents with the tweets that were aligned with respondents' current beliefs. For respondents who did not have an opinion on reporting or believed the numbers were accurate, they were randomly assigned to either the overreporting or underreporting tweet condition. Participants were then asked questions to assess their attitudes and behaviors before being presented with one of two fabricated tweets claiming coronavirus deaths are being misreported, either overreported (See Fig. 1) or underreported (See Fig. 2). After viewing the tweet, respondents were asked to respond to items to assess their attitudes regarding the tweet's credibility. After the initial tweet was presented to the respondent, they were then shown the tweet again, this time flagged with a statement which read "Caution: Suspected Bot Account. Learn More." Respondents were then asked to assess the flagged tweet's credibility. Respondents were ultimately shown the tweet a third time with a second flag added which read "Caution: Tweet contains misinformation about the novel coronavirus. Learn more." After being presented with the tweet with two flags, respondents were asked to assess the tweet's credibility. Tweets were created using the Tweetgen.com service (beta-0.3.2 2020) and used identical share and like numbers for all versions. The numbers were chosen to neither appear very low nor very high to keep the participants focused on the content flags. Our methodology follows other recent studies (Borah and Xiao 2018; Wasike 2017; Lim and Lee-Won 2017; Oeldorf-Hirsch et al. 2020; Scott et al. 2020; Solnick et al. 2020) that test the effects of static representations of Twitter and Facebook posts.

Fig. 1 Tweets for overcounted coronavirus numbers condition

Fig. 2 Tweets for undercounted coronavirus numbers condition

3.4 Measures

Participant preventative behaviors were assessed by asking respondents how frequently (never, sometimes, often, or always) they engaged in seventeen different behaviors designed to reduce the risk of catching the coronavirus. Behaviors included avoiding nonessential shopping, frequently washing hands for 20 seconds, cleaning regularly touched surfaces with disinfectant, and limiting gatherings to fewer than 10 people.

To assess respondents' fears related to the coronavirus, we asked respondents to indicate their level of agreement using a 5-point Likert-type scale for three statements designed to capture their coronavirus health-related concerns. These statements included: "I am scared that I might contract coronavirus," "I am scared that someone in my family will contract coronavirus," and "I fear that if I or someone in my family gets coronavirus, we will face complications that require hospitalization."

Respondents were asked to report the number of hours they spent on social and news media both before the coronavirus and in the 30 days prior to completing the survey. To determine time spent on social media, respondents were asked to report the number of hours spent per day on social media, while time spent on news media was collected by asking respondents to report the number of hours they spent watching/reading news for each time period.

Respondents were presented with 22 different news sources (including a write-in "other" option) and asked to indicate where they received their news. Respondents were allowed to select multiple options from the list. Major television and social media sites were listed separately (e.g., CNN, Fox News, Facebook, Twitter), while less frequently consumed sources or regional media were presented using categories such as "Local Television News" or "Liberal News Websites (Mother Jones, the Nation)." Respondents were also given the option of selecting news information directly from President Trump through either "President Trump Tweets" or "President Trump White House briefings."

Anomie, or the breakdown in belief in social bonds, was assessed using the nine item GSS Anomie Scale (Smith et al. 2019). Sample items include "Most people don't care what happens to others" and "A person must live pretty much for today." Respondents responded using a 5-point Likert-type scale. The Cronbach's alpha of this scale was 0.845.

To assess the extent to which respondents generally believe and engage in conspiratorial thinking we used the Conspiracy Mentality Questionnaire (CMQ; Bruder et al. 2013). The scale consisted of five items (e.g., there are secret organizations that greatly influence political decisions) and used a 5-point Likert-type scale with strongly agree and strongly disagree anchors. Cronbach's alpha for the CMQ was 0.832.

Government trust was measured using the Citizen Trust in Government Organizations' scale (Grimmelikhuijsen and Knies 2017). Respondents were presented with nine statements and asked the extent to which they agree or disagree with each statement using a 5-point Likert-type scale. Sample items include "the federal government is capable" and "the federal government is honest." The overall Cronbach's alpha for this scale was 0.959.

To assess respondent attitudes regarding tweet credibility, we adapted items used to evaluate Twitter posts first used by Vraga and Bode (2017a, b). After viewing each tweet, respondents were asked to evaluate the tweet as being useful, interesting, trustworthy, credible, biased, accurate, or relevant using a 5-point Likert-type scale. Additionally, respondents were asked to indicate how they would interact with the tweet by responding to four questions to gauge likely behaviors in regard to the tweet. These behaviors included following the Twitter account, retweeting the tweet, liking the tweet, and searching for additional information related to the tweet.

After reading the tweets, respondents were presented with a cognitive dissonance measure developed by Metzger et al. (2020) to determine the impact of viewing attitude-challenging information, such as flagging a tweet that shares respondents' beliefs, on feelings of dissonance. This was measured using a nine-item 5-point Likert-type scale. Cronbach's alpha for the scale was 0.638.

Respondent religiosity was captured using a 3-item measure developed by Barnett et al. (1996). Respondents were presented with three statements (e.g., "my religion is very important to me") and asked to respond to each statement using a 5-point Likert-type scale. The Cronbach's alpha for this scale was 0.930.

Given the nature of COVID-19, basic respondent health information related to the virus was also collected for this study. Respondents were asked to report whether they, someone in their household, a family member, close friend or acquaintance, or coworker had been diagnosed with COVID-19. Respondents were also asked to indicate whether they suffered from any of the preexisting conditions that increased the risk of severe illness from COVID-19 (Centers for Disease Control and Prevention 2020).

Demographic data including age, gender, ethnicity, and highest degree completed were collected from all respondents.

3.5 Analysis

Data were analyzed using Kruskal–Wallis, ANOVA, Chi-squared test for independence, independent t-tests, and Pearson correlations using IBM SPSS 26 (2020). Graphs were created in Microsoft Excel (2016).

4 Results

4.1 Association between belief in COVID-19 numbers and preventative behaviors

To identify whether there is an association between the participants' belief about the accuracy of COVID-19 mortality figures and preventive behaviors, we performed a Kruskal–Wallis test and found that there was a statistically significant difference in hand washing ($H_3 = 15.653$, $p = 0.001$), avoiding touching the face ($H_3 = 15.407$, $p = 0.002$), avoiding using cash when making purchases ($H_3 = 13.725$, $p = 0.003$), limiting gatherings to fewer than 10 people ($H_3 = 33.311$, $p < 0.001$), working from home ($H_3 = 16.313$, $p = 0.001$), avoiding nonessential shopping ($H_3 = 22.595$, $p < 0.001$), monitoring news about coronavirus ($H_3 = 11.326$, $p = 0.01$, practicing social distancing ($H_3 = 24.511$, $p < 0.001$), using electronic communication to avoid meeting with people in person ($H_3 = 27.78$, $p < 0.001$), wearing a mask ($H_3 = 43.923$, $p < 0.001$), staying at home unless shopping for core needs ($H_3 = 16.195$, $p = 0.001$), and quarantining from others if symptoms appear ($H_3 = 18.879$, $p < 0.001$). The average preventative behavior score for each coronavirus case count group is shown in Fig. 3.

For those who have personally experienced COVID-19 themselves or had someone they know, either a friend or family member, contract the disease, they were more likely to take a clear position on the COVID-19 mortality count; the "unsure" respondents were those without personal experience ($H_3 = 13.998$, $p = 0.003$). A Kruskal–Wallis test also showed that those who fear contracting the coronavirus, fear that their family will contract the virus, and that they or their family may face complications were more likely to believe the COVID-19 numbers were accurate or undercounted compared to those who believe the numbers are overestimated ($H_3 = 20.063$, $p < 0.001$; $H_3 = 18.732$, $p < 0.001$; $H_3 = 15.649$, $p = 0.001$).

4.2 Impact of Twitter flags on change in belief about COVID-19 numbers

A series of paired t-tests show how individual respondents changed their opinions about the Twitter accounts after a flag was placed on the tweet. The first flag warned participants that the tweet was shared by a suspected bot account. After being flagged as a potential bot, participant's perspective on the Tweet's credibility changed for every measure except bias, which remained consistent regardless of the Twitter flags. In responding with their desire to follow the Twitter account, the version with no flag was higher than the bot flag ($t_{298} = 8.638$, $p < 0.001$) and the bot–misinformation flag ($t_{298} = 9.443$, $p < 0.001$). For willingness to retweet this tweet, the unflagged tweet was rated more highly than the bot flag ($t_{298} = 5.165$, $p < 0.001$) and the bot–misinformation flag ($t_{298} = 5.819$, $p < 0.001$). Willingness to like the tweet followed the same pattern from unflagged to a reduced willingness after the tweet was flagged as a bot ($t_{298} = 5.862$, $p < 0.001$) and then as misinformation ($t_{298} = 8.581$, $p < 0.001$). The next set of questions dealt with the perception of the tweet. The unflagged tweet was rated more highly for willingness to seek more information compared to the bot flag ($t_{298} = 6.177$, $p < 0.001$) and the bot–misinformation flag ($t_{298} = 8.793$, $p < 0.001$), respectively; the same was true for usefulness ($t_{298} = 6.113$, $p < 0.001$; $t_{298} = 9.43$, $p < 0.001$), interest ($t_{298} = 6.199$, $p < 0.001$, $t_{298} = 8.318$, $p < 0.001$), trustworthiness ($t_{298} = 6.304$, $p < 0.001$; $t_{298} = 9.349$, $p < 0.001$), credibility ($t_{298} = 7.977$, $p < 0.001$; $t_{298} = 10.439$, $p < 0.001$), accuracy ($t_{298} = 6.264$, $p < 0.001$; $t_{298} = 11.581$, $p < 0.001$), and relevance ($t_{298} = 6.942$, $p < 0.001$; $t_{298} = 9.412$,

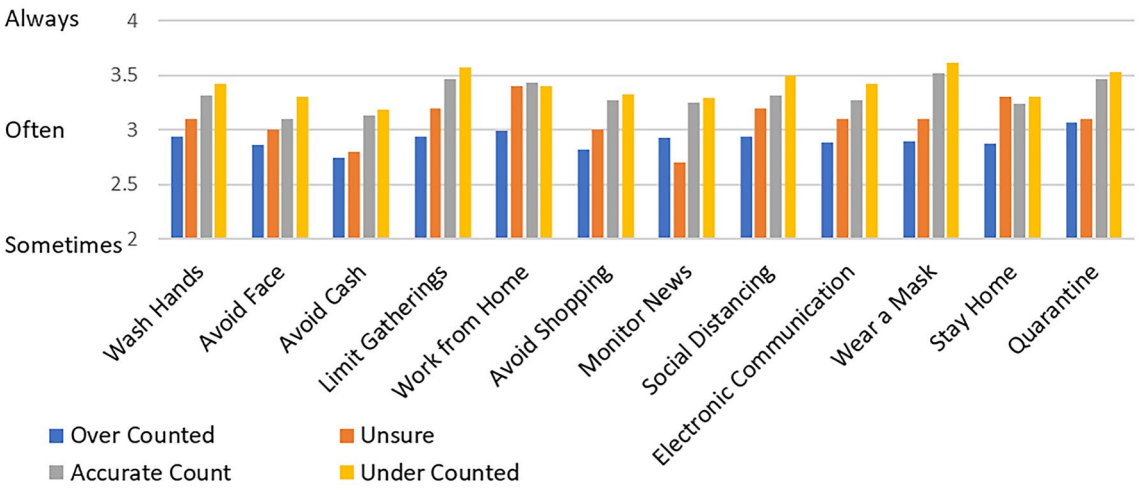

Fig. 3 Differences in preventative behaviors based on belief in COVID-19 Count

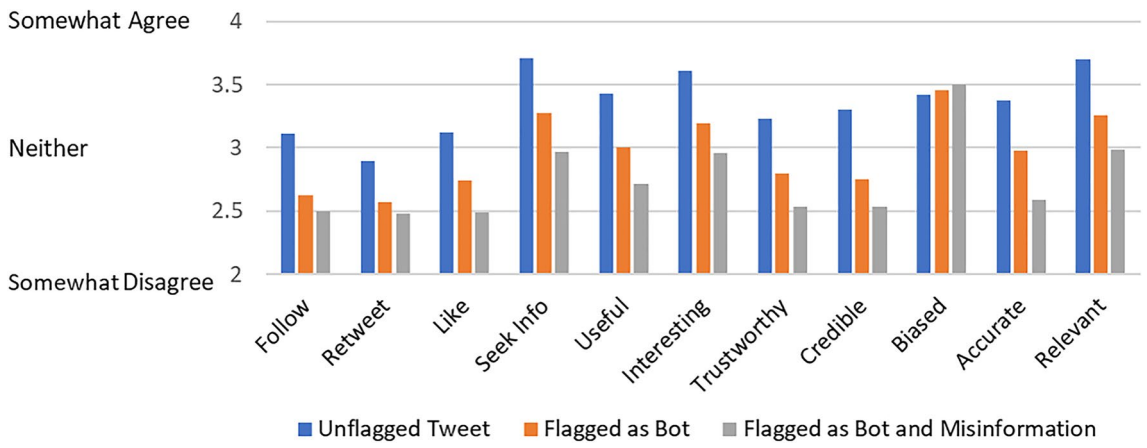

Fig. 4 Change in rating after tweets flagged

$p < 0.001$). The one aspect that remained the same despite the flags as a bot or as a bot–misinformation was bias: $t_{298} = -0.452$, $p = 0.652$; $t_{298} = -0.951$, $p = 0.342$. The average rating for each flag condition is shown in Fig. 4.

A Chi-squared test of independence shows that exposure to the series of tweets with flags changed the opinions of certain participants at a statistically significant rate: $\chi^2_9 = 462.360$, $p < 0.001$. Those who were unsure or believed the count is accurate were more likely to switch their perspective to the tweet they viewed and agree with it even after it was flagged as a bot and cautioned it contained misinformation. For those who were initially unsure of the count's accuracy, 80% remained unsure and 20% switched to saying the numbers are overcounted after seeing the overcount tweet. For those who believe the numbers are accurate, 78% held that position, but 12% changed their opinion to match the tweet they saw, while 9% adjusted their opinion against the tweet they saw. Those who believed the numbers are overcounted were most susceptible to changing their opinions after seeing the cautionary flags; 73% continued to believe numbers were overcounted, with 5% becoming unsure, 11% saying the numbers are accurate, and 12% saying the numbers are undercounted. Those who believe the numbers are undercounted were the most dependable in their belief with 88% stating the count is underreported, 4% saying it is accurate, and 8% saying it is overcounted.

4.3 Characteristics based on belief in COVID-19 count accuracy

Using an ANOVA test, we found that there was a difference between respondents' view of the COVID-19 mortality count and their score on the cognitive dissonance scale ($F_{3,295} = 3.437$, $p = 0.017$). Those who believed the count is overstated averaged 0.33 less than those who believe the count is accurate and averaged 0.232 less than those who believe the count is undercounted. There were also differences in the conspiracy scale ($F_{3,295} = 3.21$, $p = 0.023$) and trust in the government ($F_{3,295} = 11.068$, $p < 0.001$). Those who believe the count is overstated had a higher average conspiracy score by 0.308 compared to those who think the number is accurate and by 0.33 for those who believe the count is undercounted. The participants who believe the number is undercounted did not trust the government compared to the other groups with an average difference of 0.608 for those who think the number is overstated and an average difference of 0.77 for those who think the number is accurate.

An additional ANOVA test revealed that religiosity and political affiliation were also associated with differences in belief. On a seven-point scale, the overcount participants were 1.1 points more conservative on average than the undercount participants; the accurate count participants were 1.14 points more conservative than the undercount participants ($F_{3,295} = 8.227$, $p < 0.001$). Those who believe the numbers are accurate or overcounted also agreed more strongly to the statement "I am very religious," "My religion is very important to me," and "I believe in God" when compared to the undercount position ($F_{3,295} = 9.610$, $p < 0.001$).

4.4 Characteristics for changed opinion

Anomie, or the breakdown in belief in social bonds, was higher in those who changed their mind (3.38) than in those who didn't (3.12) ($t_{297} = -2.147$, $p = 0.033$).

Trust in government was higher in those who changed their mind (3.57) than those who did not (2.99) ($t_{100.028} = -4.18$, $p < 0.001$).

Those who changed their mind had more formal education (5.21) than those who did not change their mind (4.68) ($t_{118.965} = -4.111$, $p < 0.001$).

Those willing to change their mind also had more pre-existing conditions (2.54) compared to those unwilling to change their mind (1.7) ($t_{297} = -2.197$, $p = 0.029$).

Lastly, those willing to change their mind were less concerned about the economy being negatively impacted from the coronavirus (3.16) than those who did not change their mind who had higher levels of concern (3.8) ($t_{297} = 3.9$, $p < 0.001$).

4.5 News media consumption

A Chi-squared test of independence showed that certain news media outlets had an impact on what the participant believed. The number of participants from each group who read or view the following news resources are shown in Fig. 5. Fox news ($\chi^2 = 12.191$, $p = 0.007$), One America News Network ($\chi^2 = 13.379$, $p = 0.004$), National Newspapers ($\chi^2 = 11.495$, $p = 0.009$), Liberal News Websites ($\chi^2 = 8.641$, $p = 0.034$), Conservative News Websites ($\chi^2 = 13.863$, $p = 0.003$), and Facebook ($\chi^2 = 14.977$, $p = 0.002$) all had differences. The White House Briefings, President Trump's Twitter Feed, Instagram, Twitter, Reddit, Satire, general news websites, news magazines, local and national radio, local television news, BBC, MSNBC, or CNN consumption was not associated with a difference in view on COVID-19 mortality figures.

To further understand the effect of news media consumption on our participants' response to the flagged tweets, we performed Mann–Whitney tests to identify whether there were differences for regular consumers of Fox News, National Newspapers, and news sourced from Facebook. Fox News consumers were statistically significantly different from those who do not consumer Fox News, showing a willingness to continue to engage with the tweet by following the account (2.97), retweeting (2.92), and liking the tweet (2.85) compared to those who do not view Fox News who were less likely to follow (2.08), retweet (2.09), and like (2.17); $U = 7103$, $p < 0.001$, $U = 7370$, $p < 0.001$, $U = 8129$, $p < 0.001$. The same pattern held for attitudes toward the tweets, with Fox News viewers rating the tweet's usefulness, interest level, trustworthiness, credibility, accuracy, and relevance more highly, even with bot and misinformation flags, than those who do not watch Fox News. Seeking news from Facebook also led to a resistance to the flags, with those participants continuing to engage and keep their attitude ratings higher than those who do not use Facebook for news. The National Newspaper readers, on the other hand, were distinct from those who do not read national newspapers because they decreased the amount of engagement they had with the tweets after seeing the flags. National newspaper readers were less likely to follow (2.14) or like the tweet (2.22) compared to those who do not read national newspapers (follow: 2.67, like: 2.59); $U = 7945$, $p = 0.004$, $U = 8501$, $p = 0.038$.

4.6 Hours on social media and news media

A Pearson correlation identified a trend for participants in how their rating changed after viewing the Tweets with a flag for a suspected bot account and when there was a

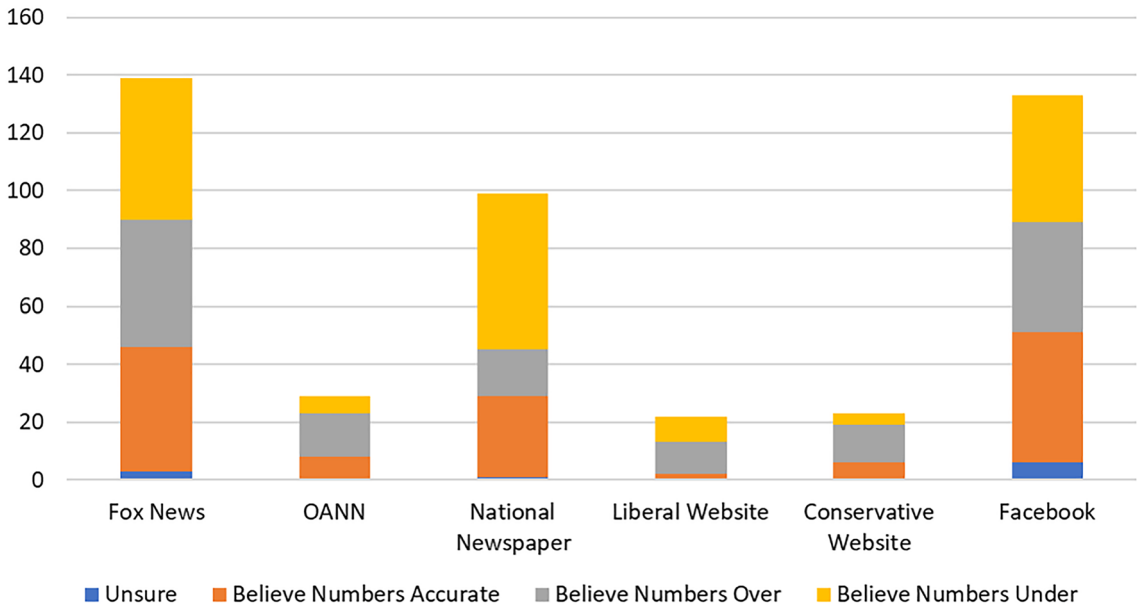

Fig. 5 Differences in news media consumption on opinions about COVID-19 death count

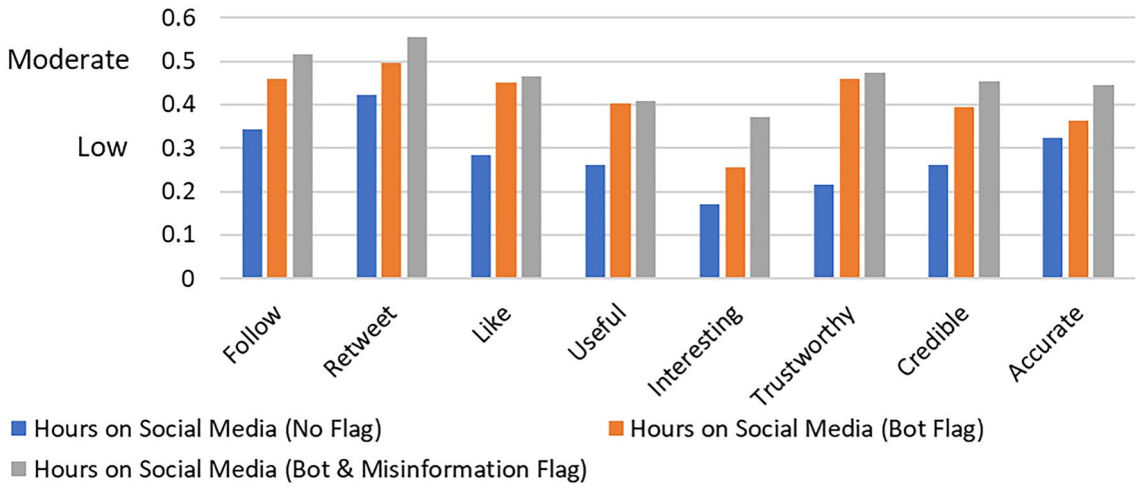

Fig. 6 Hours spent on social media correlated with higher tweet rating despite flags

flag for both the bot account and misinformation ($n = 299$, $p < 0.001$). With a greater number of hours spent on social media both before and during the pandemic, participants were more likely to continue engaging with the tweet through follows, retweets, or likes, and rate the Tweet as having higher usefulness, interest, trust, credibility, and accuracy. While those who spent fewer hours on social media lowered their rating of the Tweets on engagement and perception *after* seeing the cautionary flags, the high-volume users kept their scores higher (See Fig. 6).

The trend was the same for hours spent consuming news media both before and during the pandemic. A Pearson correlation ($n = 299$, $p < 0.001$) showed a trend where those who spent more hours consuming news media kept their rating of the tweet's usefulness, interest, trustworthiness, credibility, and accuracy high; they also were more likely to continue to engage through follows, retweets, and likes. The cautionary flags had less impact on regular news media consumers who continued to rate and engage at higher rates (See Fig. 7).

5 Discussion

5.1 Bot flags change participants' engagement and attitudes about tweets

Our results strongly suggest that Twitter message flags negatively affect participants' perceptions of unreliable tweets. After seeing a warning that the tweet comes from a suspected bot account, participants in both groups decreased their willingness to engage with the tweet and lowered their opinion regarding how useful, interesting, trustworthy, credible, helpful, accurate, and relevant they found the tweet. (Participants rated the unlabeled tweet as

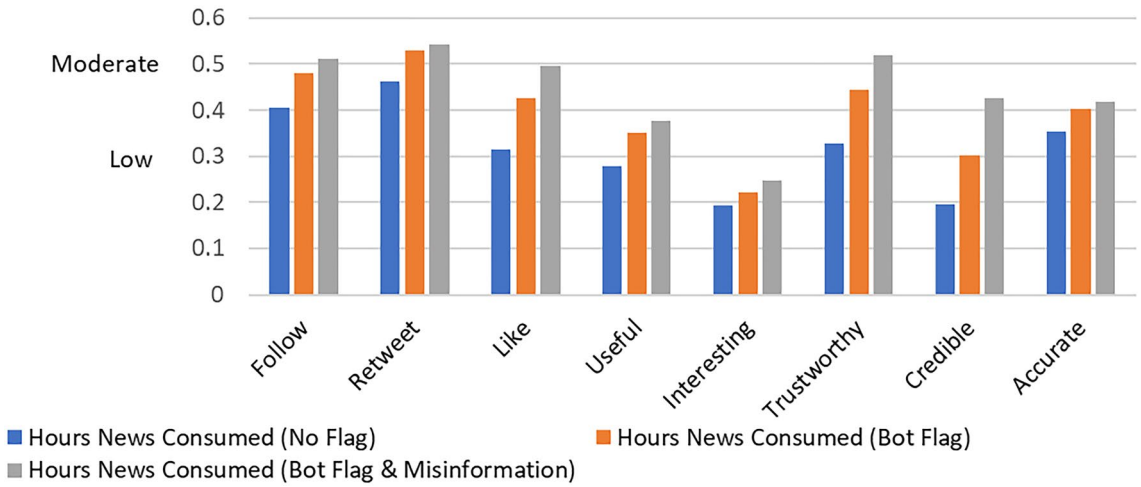

Fig. 7 Higher news consumption in hours correlated with higher tweet rating despite flags

highly biased and did not increase their bias ratings after seeing the flags.) These responses suggest that efforts to flag bot content may mitigate their ability to spread misinformation on Twitter. These results align with previous studies finding that people consider identified or suspected bots as less trustworthy than human-generated content (Waddell 2018; Bruzzese et al. 2020; Graefe and Bohlken 2020; Jakesch et al. 2019). Additionally, because "the author is the feature that led the users to the most accurate perceptions" about the veracity of information (Zubiaga and Ji 2014), it is not surprising that finding out that Twitter posters are bots reduces user attitudes about the tweet.

5.2 Misinformation flags change participants' engagement and attitudes about tweets

Results similarly showed that flagging the tweet as containing misinformation further lowered participants' willingness to engage it and negatively affected participant perceptions, especially in terms of the trustworthiness, accuracy, and credibility of the tweet. These responses show encouraging indications that fact-checking can help combat the COVID-19 infodemic. This experiment reinforces similar findings from other studies about the effectiveness of identifying online health misinformation (Oh and Lee 2019; Kim and Dennis 2019; Bode and Vraga 2018), and our results show that merely identifying misinformation, without offering corrections, can decrease participants' opinions of the misinformation source. Unlike Colliander (2019), these results demonstrate that misinformation flags from social media companies can lower participants' attitudes about inaccurate tweets. Twitter reported the impact of labeling tweets that contain false information during the 2020 Presidential election (Gadde and Beykpour 2020). During a 16-day period around the 2020 election, Twitter labeled 300,000 tweets as containing disputed or misleading information. A subset of these tweets (n = 456) were flagged and locked so only limited user engagement was allowed. For these 456 tweets, Twitter would only allow them to be "quote tweeted," where tweets are retweeted but must include an original comment from the user retweeting. Twitter reports tweets with a misleading information flag experienced a 29% decrease in user quote tweets. Further, our research suggests that multiple flags work better than one; flagging tweets for both suspected bot authorship and misinformation decreased positive attitudes more strongly than a bot flag alone, indicating that social media platforms may want to provide multiple flags identifying various issues with unreliable content.

However, misinformation flags did not affect all participants equally. People who reported spending more time on social media showed more resistance to both flags, suggesting that perhaps these participants who spend more time on social media have greater trust in online information or have grown immune to warnings from social media companies about information veracity. These results connect with Allington et al. (2020), who found a link between the use of social media for COVID-19 information and a propensity toward COVID-19 conspiracy beliefs. Additionally, research has connected trust in social media with the likelihood of sharing fake COVID-19 news online (Laato et al. 2020). Participants who spent more time watching news also showed more resistance to the flags overall, though differences also emerged based on the news source. Participants who watched network news were more responsive to the flags, a result that evokes a correlation Allington et al. (2020) found between watching network news and participating in COVID-19-related health-protective behaviors. Inversely, participants who reported getting news from Fox News or Facebook had more positive reactions to the tweets even after viewing the flags. Dhar et al. (2016) proposed a rumor control model where an "authenticated news agency" can flood a social network with counter statements that dilute the effects of misinformation (p. 56). Our study shows the limitations of counter statements in practice when individual users pick and choose who they believe is an authentic news source.

5.3 Flags change some participants' minds about COVID-19 misinformation

The flags also showed some effectiveness at changing people's opinions about the COVID-19 death tolls, but more notably in participants who believed numbers were over rather than undercounted. Twenty-one participants who initially felt that death tolls were overcounted changed their mind, though only 10% (8) from this subset changed their minds to believe the counts were accurate. Meanwhile, only sixteen participants who initially believed that death counts were undercounted changed their minds, and only around 4% (8) from that subset changed their minds to believe the counts were accurate. It appears that the flags more effectively changed minds about individual tweets than people's opinions overall. However, the ability of some participants to change their minds, combined with their lowered perceptions of the tweets, suggests that the flags did not activate a backfire effect, where people respond to fact-checking by becoming further entrenched in their original viewpoints. These results are especially encouraging because participants were sorted into groups to see tweets that aligned with their previously held beliefs, allowing us to test more directly for backfire effects.

Exposure to the tweets also changed the minds of several participants who initially felt that U.S. COVID-19 counts were accurate. Sixteen participants, 23% of those who started the experiment believing the counts were accurate, ended believing that the COVID-19 counts were under- or

overreported. While seven ended up disagreeing with the narrative they saw in the flagged tweets, nine participants ended by agreeing with the narrative they saw in the flagged tweets. Exposure to misinformation, even when it is paired with warning flags about its authorship and accuracy, may cause some participants to adopt more conspiratorial views. Though the flags lowered perceptions about the tweets regardless of participant perception of COVID-19 death counts, the misinformation may overpower the flags in some instances. Additionally, flags may work more effectively for participants who align with the misinformation than for those who hold differing or uncommitted beliefs, where their response is just as likely to be agreement as it is disagreement after they view the flagged tweet.

In addition to observing that participants can change their mind after viewing the flagged tweets, we found individual differences also influenced the likelihood participants would change their mind after exposure to the flags. Individual attitudes such as anomie (the view that there is a societal breakdown in norms) and an individual's trust in the government correlated with a greater likelihood that users would change their minds by the end of the experiment. Further, individuals changed their minds more frequently when they had higher levels of education and more preexisting conditions that make them vulnerable to COVID-19. Despite this effect of preexisting conditions, the experiment showed no effects for COVID-19 anxiety on either responses to the tweets or the likelihood that participants would change their minds, a finding that aligns with Laato et al. (2020).

5.4 Limitations and future research

Studying the impact, spread, and prevention of misinformation on social media is necessary as modern society relies on belief and trust in authorities and public health information to respond collectively to global crisis and tragedy. While this study is novel in its approach to filtering respondents to different versions of the survey instrument depending on their existing beliefs, there are some limitations. First, the window of data collection, September 8–10, 2020, was less than sixty days before a polarized national election in the United States. With Twitter's tendency to heighten partisan political communication, the ecosystem that Twitter users experienced during this window is more intense than normal which could impact how our participants responded to the flagged Tweets. Additionally, the spread of Twitter information is a complex contagion (Monsted et al. 2017) so it may be necessary to test multiple exposures to conspiracy theories and flagged content. While the current study did facilitate multiple exposures to a single tweet, future studies could add in longer periods of time between exposures to measure the persistence of the effects identified. We might also include full interactivity, like the ability to see who shared and commented, as well as full functionality, with other ways to respond as a participant. Other future studies might look at the source of the flags, whether government, the social media company, or other users, to see which source is most effective at correction and changing belief in conspiracy theories. Additional research might address the psychological perceptions of the flags as visual cues and how different colors, symbols, or language impact participants. Other flags that are more forceful in offering correct information may also allow the identification of the bright line for preventing the backfire effect (Wang et al. 2019). Finally, a realistic experiment that utilizes social network mining to monitor the real-world behavior of those exposed to flagged tweets would provide a glimpse of true behavior (follow, share, tweet, and retweets) rather than self-report data alone.

6 Conclusion

Identifying solutions for the COVID-19 infodemic requires a careful consideration of technical challenges alongside the human ones. The current study helped clarify the ways that human users respond to and interact with flagging techniques when content is identified as misinformation or identified as propagated by a bot account. There is evidence that flags can change engagement behaviors and most user attitudes toward misinformation, suggesting that it should be paired with automatic fact-checking strategies. There are still challenges that remain unresolved, with more regular users of social media showing an immunity to the flags. The stakes are high: Gruzd and Mai (2020) found that "power users" had an outsized impact on spreading the #FilmYourHospital conspiracy theory that COVID-19 is a hoax. When the power users (conservative politicians and right-wing political activists) encouraged their followers to film empty hospital rooms and waiting rooms, social media misinformation inspired an immediate threat to public safety. Multiple content and source flags may have lowered the spread and persuasive power of #FilmYourHospital messages. To ensure that everyone gets high-quality public health and safety information from credible and authoritative sources during large-scale crises, multiple strategies should be used to protect the integrity of our social media networks.

Funding No external funds were used for this study.

Declarations

Conflicts of interest The author declares there is no conflict of interest.

Data transparency https://osf.io/4gkzb/?view_only=34ba303363084a6b8773c160975d6ede

Ethical approval IRB Approval by UAH August 29, 2020 (EE202081).

References

Allington D, Duffy B, Wessely S, Dhavan N, Rubin J (2020) Health-protective behaviour, social media usage and conspiracy belief during the COVID-19 public health emergency. Psychol Med. https://doi.org/10.1017/S003329172000224X

Barua Z, Barua S, Aktar S, Kabir N, Li M (2020) Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. Prog Disaster Sci 8:1–9. https://doi.org/10.1016/j.pdisas.2020.100119

Barnett T, Bass K, Brown G (1996) Religiosity, ethical ideology, and intentions to report a peer's wrongdoing. J Bus Ethics 15:1161–1174. https://doi.org/10.1007/BF00412815

Bode L, Vraga E (2017) See something, say something: correction of global health misinformation on social media. Health Commun 33:1131–1140. https://doi.org/10.1080/10410236.2017.1331312

Borah P, Xiao X (2018) The importance of 'likes': The interplay of message framing, source, and social endorsement on credibility perceptions of health information on Facebook. J Health Commun 23:399–411

Bowles J, Larreguy H, Liu S (2020) Countering misinformation via WhatsApp: Evidence from the COVID-19 pandemic in Zimbabwe. Pre-print

Broniatowsky D, Jamiso A, Qi S, AlKulaib L, Chen T, Benton A, Quinn S, Dredze M (2018) Weaponized health communication: twitter bots and Russian trolls amplify the vaccine debate. Am J Public Health 108:1378–1384. https://doi.org/10.2105/AJPH.2018.304567

Bruder M, Haffke P, Neave N, Nouripanah N, Imhoff R (2013) Measuring individual differences in generic beliefs in conspiracy theories across cultures: conspiracy mentality questionnaire. Front Psychol 4:225. https://doi.org/10.3389/fpsyg.2013.00225

Bruzzese T, Ding C, Gao I, Romanos A, Dietz G (2020) Effect of confidence indicators on trust in AI-generated profiles. CHI Conference on Hum Factors in Computing Systems. https://doi.org/10.1145/3334480.3382842

Carey J, Chi V, Flynn D, Nyhan B, Zeitzoff T (2020) The effects of corrective information about disease epidemics and outbreaks: evidence from Zika and yellow fever in Brazil. Sci Adv 6:7449

Centers for Disease Control and Prevention (2020) Do I need to Take Extra Precautions Against COVID-19. Centers for Disease Control and Prevention. https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/index.html. Accessed 14 October 2020

Chen L, Wang X, Peng T (2018) Nature and diffusion of gynecologic cancer-related misinformation on social media: analysis of tweets. J Med Internet Res 20:e11515. https://doi.org/10.2196/11515

Chou W, Oh A, Klein W (2018) Addressing health-related misinformation on social media. JAMA 320:2417–2418. https://doi.org/10.1001/jama.2018.16865

Clement J (2020a) United States Twitter gender distribution. Statista. https://www.statista.com/statistics/678794/united-states-twitter-gender-distribution Accessed 13 October 2020

Clement J (2020b) Age distribution of global Twitter users. Statista. https://www.statista.com/statistics/283119/age-distribution-of-global-twitter-users/ Accessed 13 October 2020

Colliander J (2019) 'This is fake news': Investigating the role of conformity to other users' views when commenting on and spreading disinformation in social media. Comput Hum Behav 97:202–215. https://doi.org/10.1016/j.chb.2019.03.032

Dewan P, Kumaraguru P (2017) Facebook inspector: Towards automatic real-time detection of malicious content on Facebook. Soc Netw Anal Min 7:15. https://doi.org/10.1007/s13278-017-0434-5

Dhar J, Jain A, Gupta VK (2016) A mathematical model of news propagation on online social network and a control strategy for rumor spreading. Soc Netw Anal Min 6:57. https://doi.org/10.1007/s13278-016-0366-5

Ebrahimji A (2020) "Doctors say coronavirus myths on social media are 'spreading faster than the virus itself'," *CNN*, 1 September 2020. https://www.cnn.com/2020/09/01/business/coronavirus-myths-social-media-doctors-trnd/index.html. Accessed 14 October 2020

Edwards C, Edwards E, Spence P, Shelton A (2014) Is that a bot running the social media feed? Testing the differences in perceptions of communication quality for a human agent and a bot agent on Twitter. Comput Hum Behav 33:372–376. https://doi.org/10.1016/j.chb.2013.08.013

Ferrara E (2020) What types of COVID-19 conspiracies are populated by Twitter bots? First Monday 25:2–25. https://doi.org/10.5210/fm.v25i6.10633

Ford JB (2017) Amazon's Mechanical Turk: a comment. J Advert 46(1):156–158. https://doi.org/10.1080/00913367.2016.1277380

Forelle M, Howard P, Monroy-Hernández A, Savage S (2015) Political bots and the manipulation of public opinion in Venezuela. https://arxiv.org/abs/1507.07109

Gadde, V., & Beykpour, K. (2020). An update on our work around the 2020 US Elections. https://blog.twitter.com/en_us/topics/company/2020/2020-election-update.html Accessed 14 January 2021

Graefe A, Bohlken N (2020) Automated journalism: A meta-analysis of readers' perceptions of human-written in comparison to automated news. Media Commun. 8:50–59. http://dx.doi.org/https://doi.org/10.17645/mac.v8i3.3019

Grimmelikhuijsen S, Knies E (2017) Validating a scale for citizen trust in government organizations. Int Rev Adm Sci 83(3):583–601. https://doi.org/10.1177/2F0020852315585950

Gruzd A, Mai P (2020) Going viral: How a single tweet spawned a COVID-19 conspiracy theory on Twitter. Big Data Soc. https://doi.org/10.1177/2053951720938405

Jakesch M, French M, Ma X, Hancock J, Naaman M (2019) AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness. CHI. https://doi.org/10.1145/3290605.3300469

Kang C, Isaac M (2019) Zuckerberg says Facebook won't police political speech. New York Times. https://www.nytimes.com/2019/10/17/business/zuckerberg-facebook-free-speech.html Accessed 14 October 2020

Keith MG, Tay L, Harms PD (2017) Systems perspective of Amazon Mechanical Turk for organizational research: Review and recommendations. Front Psychol 8:1–19. https://doi.org/10.3389/fpsyg.2017.01359

Kim A, Dennis A (2019) Says who? The effects of presentation format and source rating on fake news in social media. MIS Q 43:1025–1039. https://doi.org/https://doi.org/10.25300/MISQ/2019/15188

Kouzy R, Abi Jaoude J, Kraitem A, El Alam MB, Karam B, Adib E, Zarka J, Traboulsi C, Aki EW, Baddour K (2020) Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. Cureus 12:e7255. https://doi.org/10.7759/2Fcureus.7255

Laato S, Islam A, Islam M, Whelan E (2020) What drives unverified information sharing and cyberchondria during the COVID-19 pandemic. Eur J Inf Syst 29:288–305. https://doi.org/10.1080/0960085X.2020.1770632

Levay KE, Freese J, Druckman JN (2016) The demographic and political composition of Mechanical Turk samples. Sage Open, January-March 2016:1–17. https://doi.org/10.1177/2158244016636433

Lewandowsky S, Ecker U, Seifert C, Schwarz N, Cook J (2012) Misinformation and its correction: Continued influence and successful debiasing. Psychol Sci Public Interest 13:106–131. https://doi.org/10.1177/1529100612451018

Lim Y, Lee-Won RJ (2017) When retweets persuade: The persuasive effects of dialogic retweeting and the role of social presence in organizations' Twitter-based communication. Telematics Inform 34:422–433

Lyons B, Merola V, Reifler J (2019) Not just asking questions: effects of implicit and explicit conspiracy information about vaccines and genetic modification. Health Commun 34(14):1741–1750. https://doi.org/10.1080/10410236.2018.1530526

Memon A, Carley K. (2020) Characterizing COVID-19 misinformation communities using a novel Twitter dataset. https://arxiv.org/abs/2008.00791

Metzger MJ, Hartsell EH, Flanagin AJ (2020) Cognitive dissonance or credibility? A comparison of two theoretical explanations for selective exposure to partisan news. Commun Res 47:3–28. https://doi.org/10.1177/0093650215613136

McCredie MN, Morey, LC (2019) Who are the turkers? A characterization of MTurk workers using the personality assessment inventory. Assessment 36(5):759–766. https://doi.org/10.1177/1073191118760709

Microsoft (2016) Excel [Computer software]. https://www.microsoft.com/enus/microsoft-365/microsoft-office

Miller J, Saunders K, Farhart C (2016) Conspiracy endorsement as motivated reasoning: the moderating roles of political knowledge and trust. Am J Pol Sci 60:824–844. https://doi.org/10.1111/ajps.12234

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:e0184148. https://doi.org/10.1371/journal.pone.0184148

O'Connor C, Murphy M (2020) Going viral: doctors must tackle fake news in the covid-19 pandemic. BMJ 369:1587. https://doi.org/10.1136/bmj.m1587

Oeldorf-Hirsch, A., Schmierbach, M., Appelman, A., & Boyle, M.P. (2020). For the birds: Media sourcing, Twitter, and the minimal effect on audience perceptions. Convergence: Int J Res into New Media Technol, 26.2, 350–368. https://doi.org/10.1177/1354856518780438

Oh H, Lee H (2019) When do people verify and share health rumors on social media? The effects of message importance, health anxiety, and health literacy. J Health Commun 24:837–847. https://doi.org/10.1080/10810730.2019.1677824

Pagoto S, Waring M, Xu R (2019) A call for a public health agenda for social media research. J Medical Internet Res 21:e16661. http://www.jmir.org/2019/12/e16661/

Pennycook G, McPhetres J, Zhang Y, Lu J, Rand D (2020) Fighting COVID-19 misinformation on social media: experimental evidence for a scalable accuracy nudge intervention. Psychol Sci 31:770–780. https://doi.org/10.1145/3290605.3300469

Quinn M (2020) Twitter removes tweet shared by Trump with false coronavirus statistics. *CBS* https://www.cbsnews.com/news/twitter-removes-trump-tweet-false-coronavirus-statistics/. Accessed 14 October 2020

Redmiles EM, Kross S, Mazurek ML (2019) How well do my results generalize? Comparing security and privacy survey results from MTurk, web, and telephone sampes. IEEE Symposium on Security and Privacy 2019:1326–1343

Rice J (2020) Awful archives: Conspiracy theory, rhetoric, and acts of evidence. Ohio State UP, Columbus

Scott GG, Brodie ZP, Wilson MJ, Ivory L, Hand CJ, Sereno SC (2020) Celebrity abuse on Twitter: The impact of tweet valence, volume of abuse, and the dark triad personality factors on victim blaming and perceptions of severity. Computers in Human Behaviors 103:109–119. https://doi.org/10.1016/j.chb.2019.09.020

Sharma M, Yadav K, Yadav N, Ferdinand KC (2017) Zika virus pandemic—analysis of Facebook as a social media health information platform. Am J Infect Control 45(3):301–302

Shao C, Ciampaglia G, Yang K, Flammini A, Menczer F (2018) The spread of low-credibility content by social media. Nat Commun 9:4787. https://doi.org/10.1038/s41467-018-06930-7

Shao C, Ciampaglia G, Varol O, Flammini A, Menczer F (2017). The spread of misinformation by social bots. https://arxiv.org/abs/1707.07592

Smith TW, Davern M, Freese J, Stephen LM (2019) General Social Surveys, 1972–2018: Cumulative Codebook / Principal Investigator, Tom W. Smith; Co-Principal Investigators, Michael Davern, Jeremy Freese and Stephen L. Morgan. -- Chicago: NORC, 2019. 3,758 pp., 28cm. -- (National Data Program for the Social Sciences Series, no. 25)

Solnick RE, Chao G, Ross R, Kraft-Todd GT, Kocher KE (2020) Emergency physicians and personal narratives improve the perceived effectiveness of COVID-19 public health recommendations on social media: a randomized experiment. Acad Emerg Med. https://doi.org/10.1111/acem.14188

Sommariva S, Vamos C, Mantzarlis A, Đào L, Tyson D (2018) Spreading the (fake) news: exploring health messages on social media and the implications for health professionals using a case study. Am Jour Health Ed 49:246–255. https://doi.org/10.1080/19325037.2018.1473178

The Lancet Infectious Diseases (2020). The COVID-19 infodemic. Lancet Infect Dis 20(8):875. https://doi.org/10.1016/S1473-3099(20)30565-X

Thomas KA, Clifford S (2017) Validity and Mechanical Turk: an assessment of exclusion methods and interactive experiments. Comput Hum Behav 77:184–197. https://doi.org/10.1016/j.chb.2017.08.038

Thorne J, Vlachos A (2018) Automated fact checking: Task formulations, methods, and future directions. https://arxiv.org/abs/1806.07687

Tsugawa S, Ohsaki H (2017) On the relation between message sentiment and its virality on social media. Socl Netw Anal Min, https://link-springer-com.elib.uah.edu/article/https://doi.org/10.1007/s13278-017-0439-0

"US Historical Data",The COVID Tracking Project. https://covidtracking.com/data/national Accessed14October 2020, 2020"US Historical Data," The COVID Tracking Project. Accessed 14 October 2020.

Vafeiadis M, Bortree D, Buckley C, Diddi P, Xiao A (2019) Refuting fake news on social media: nonprofits, crisis response strategies and issue involvement. J ProdBrand Manag 292:209–222. https://doi.org/10.1108/JPBM-12-2018-2146

Van der Meer T, Jin Y (2020) Seeking formula for misinformation treatment in public health crises: the effects of corrective information type and source. Health Commun 35:560–575. https://doi.org/10.1080/10410236.2019.1573295

Vraga E, Bode L (2017) I do not believe you: how providing a source corrects health misperceptions across social media platforms. Inf Commun Soc. https://doi.org/10.1080/1369118X.2017.1313883

Vraga E, Bode L (2017) Using expert sources to correct health misinformation in social media. Sci Commun 39:621–645. https://doi.org/10.1177/2F1075547017731776

Waddell T (2018) A robot wrote this? How perceived machine authorship affects news credibility. Digit Journal 6:236–255. https://doi.org/10.1080/21670811.2017.1384319

Wang Y, McKee M, Torbica A, Stuckler D (2019) Systematic literature review on the spread of health-related misinformation on social media. Soc Sci Med 240:1–12. https://doi.org/10.1016/j.socscimed.2019.112552

Wasike B (2017) Persuasion in 140 characters: Testing issue framing, persuasion, and credibility via Twitter and online news articles in the gun control debate. Comput Hum Behav 66:179–190

Waszak P, Kasprycka-Waszak W, Kubanek A (2018) The spread of medical fake news in social media-the pilot quantitative study. Health Policy and Technol 7:115–118. https://doi.org/10.1016/j.hlpt.2018.03.002

Wei W, Joseph K, Liu H, Carley KM (2016) Exploring characteristics of suspended users and network stability on Twitter. Soc Netw Anal Min. https://doi.org/10.1007/s13278-016-0358-5

Yang K, Torres-Lugo C, Menczer F (2020) Prevalence of low-credibility information on Twitter during the COVID-19 outbreak. arXiv. https://doi.org/https://doi.org/10.36190/2020.16

Yuan X, Schuchard R, Crooks A (2019) Examining emergent communities and social bots within the polarized online vaccination debate in Twitter. Soc Media Soc. https://doi.org/10.1177/2056305119865465

Zubiaga A, Heng J (2014) Tweet, but verify: epistemic study of information verification on Twitter. Soc Netw Anal Min. https://doi.org/10.1007/s13278-014-0163-y