Platform Design to Curb Misinformation

We propose and test a crowd-based content moderation approach to combat the spread of misinformation. In this method, users can flag posts they believe contain misinformation, with such flags serving as visible cues to others. These cues then influence users' moderation and sharing decisions. We assess users' willingness to participate in moderation and identify the underlying drivers. Thereafter, we examine whether such flags can address two primary causes for the spread of misinformation -lack of knowledge and lack of scrutiny, specifically, when posts are aligned with one's ideology. We developed a social media application and conducted a randomized controlled experiment. Participants were shown both true and false posts related to politics, COVID-19, and the Russia-Ukraine conflict, along with social cues like flags and share counts. Our results demonstrate that users with a higher level of knowledge are more likely to flag misinformation and that their flags are not influenced by their personal beliefs. The presence of these flags encourages other users, especially those with less information, to be more cautious about sharing content. The presence of flags also reduces the spread of misinformation, even when users' beliefs align with false posts. We also establish that users' flagging and sharing behavior are driven by their updated opinion regarding the accuracy of the post and the need for impression management. Our study reveals a key limitation in the current approach to platform moderation, where user feedback is not made visible to others, and demonstrates that a crowd-based approach can be effective in curbing misinformation.

Key words: Misinformation, Negative Cues, Social Influence, Platforms

1. Introduction

Social media platforms, including Facebook, are increasingly grappling with the challenge of misinformation spreading through their networks. In a 2020 survey by Pew Research, 64 percent of the respondents stated that social media platforms are affecting people negatively.[1] A respondent noted: "Social media is rampant with misinformation both about the coronavirus and political and social issues, and the social media organizations do not do enough to combat this." In spite of some approaches undertaken by social media platforms, misinformation remains ubiquitous; thus, there is a need to design and adopt additional or alternative strategies that would further help in curbing its spread.

[1] https://www.pewresearch.org/short-reads/2020/10/15/64-of-americans-say-social-media-have-a-mostly-negative-effect-on-the-way-things-are-going-in-the-u-s-today/

Existing content moderation approaches of major social media platforms are perceived to be silencing ‘protected speech’.[2] Such perceptions are fueled by the platform-driven centralized nature of content moderation strategy which lacks transparency.[3] As a result, platforms are increasingly seeking to incorporate crowd’s inputs to curb misinformation. Facebook provides users with a *Report* feature, which enables them to report the content they deem to be inappropriate or fake. The platform then evaluates the report and decides whether to remove the content, which may take up to three weeks (Jiménez Durán 2021), allowing misinformation to persist and potentially spread during this time. Similarly, X has introduced the Birdwatch program or Community Notes, where users can provide feedback on tweets, and where the platform selectively displays some of this feedback using its algorithm. Nevertheless, these designs also involve centralized evaluation and do not fully utilize the characteristics of social media to moderate content. For instance, contributors on X remain anonymous in the Birdwatch approach and must be approved by the platform.[4] Consequently, content moderation heavily relies on the rate of participation from a small group of users. Additionally, community notes are posted only after reaching a consensus among several contributors.[5]

An alternative approach to addressing these issues involves platforms adopting a decentralized method, relying primarily on the crowd to moderate misinformation. This would entail utilizing the crowd to identify and flag misinformation, potentially reducing the propensity to share such content. However, the efficacy of such decentralized content moderation strategy in curbing misinformation remains uncertain. Previous studies indicate that crowds can detect misinformation (Martel et al. 2022, Wang et al. 2021). However, these studies are based on anonymous reporting by the crowd. Whether the crowd can effectively curb misinformation in a social setting where user actions are visible to others remains unknown. Moreover, the effect of crowd based moderation on the content sharing behavior is also unknown. Prior studies on misinformation explore users’ motivation to share misinformation (Pennycook et al. 2021) and primarily focus on the effect of platform interventions such as nudges, warnings, etc. on user response (Jahanbakhsh et al. 2021). Whether crowd based content moderation will have the same effect remains an open question. In our study, we propose and assess the efficacy of a crowd-based content moderation strategy. In this approach, users can mark or flag posts they consider to be misinformation, making these flags visible to others with the aim of influencing subsequent user actions. These flags may serve

[2] https://techpolicy.press/debate-over-content-moderation-heads-to-the-supreme-court

[3] https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent; https://www.technologyreview.com/2020/11/06/1011769/social-media-moderation-transparency-censorship/

[4] https://help.twitter.com/en/using-x/community-notes

[5] https://www.socialmediatoday.com/news/x-expands-community-notes-program-adds-top-writer-badge/689270/

as cues about the nature of a post, potentially guiding others' decisions to either flag or share the content.[6]

To gauge the effectiveness of the proposed crowd-driven content moderation approach, it is essential to consider how it addresses the key factors influencing the spread of misinformation. A survey conducted by Pew Research found that 64 percent of respondents cited confusion about facts as the reason for consuming and spreading misinformation.[7] Only 26% of US adults could correctly classify factual statements[8] suggesting that people lack knowledge about new topics. Pennycook et al. (2021) present a similar argument, suggesting that users primarily spread misinformation due to confusion or a lack of knowledge about the accuracy of the information or simply forgetting to consider accuracy before sharing. Additionally, Bakshy et al. (2015) demonstrate that users are likely to spread misinformation when they are less inclined to scrutinize the correctness of posts that align with their beliefs. Therefore, the lack of knowledge and scrutiny, particularly when beliefs align with the misinformation, are major reasons for its spread. Hence, our study also focuses on the impact of a flag feature on these two drivers of misinformation. We address the following research questions (i) *Will users be motivated to moderate (flag) misinformation, and what role will users' knowledge and beliefs play in their willingness to flag?* (ii) *How will the presence of flags affect users' sharing behavior, and how will their knowledge and beliefs influence this behavior?* (iii) *What impact will the flag feature have on the overall spread of misinformation?*

We address the above questions through experiments using a Facebook-like social media platform we created, where we recruited 1458 participants from Amazon Mechanical Turk (MTurk). Prior work in the misinformation literature (e.g., Pennycook et al. 2020a, Jahanbakhsh et al. 2021 etc.) and other streams (such as Yin et al. 2023, Bapna et al. 2017 etc.) has generally used such laboratory experiments as their research approach. Moreover, since our research aims to propose and evaluate a new platform design feature, a laboratory experiment is the most feasible approach. To ensure consistency with previous studies in this area, we followed the criteria for recruiting MTurk participants (e.g., 95% hit rate) set by studies such as Pennycook et al. (2020a). The distribution of the demographics of the participants in our experiment is similar to that of the U.S. social media users.

In our experiments, the participants were assigned to the following categories: (i) Russia-Ukraine conflict, (ii) COVID-19, and (iii) U.S. politics. We chose these categories as they represent events

[6] Unlike platforms such as Reddit with traditional downvote capabilities, our proposed flag feature is distinct due to factors like the transparency of user identity, and its impact on social norms

[7] https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/

[8] https://newslit.org/about/mission/

where the widespread circulation of misinformation[9] has been documented and also because prior studies have used these topics to study misinformation (e.g., Dias et al. 2020). Posts were aggregated from PolitiFact (Gillin 2018), a third-party website that fact-checks news stories, and participants were presented with these posts (both true and false) in a random order.

Each post shown to an individual was accompanied by information on the number of other users who flagged and shared the post, as well as whether their friend had flagged or shared the same post. These numbers were randomly assigned to each post for every user. Upon seeing a post, participants had the option to share, flag, or take no action. Subsequently, users responded to a series of questions aimed at exploring possible underlying mechanisms. Specifically, we investigate how drivers such as update of priors regarding the accuracy of the post, self-presentation needs, and herding behavior influence flagging and sharing behavior of the participants. Finally, we collected specific information about each participant, including their demographics, beliefs, and knowledge related to the respective categories.

We find that users are more likely to flag posts that already have flags. We also find that users are less likely to share flagged content. Furthermore, we find that *High Knowledge types* flag content more actively compared to *Low Knowledge types*, particularly in the absence of other users' inputs (i.e., a post with zero flags). Furthermore, *High Knowledge types* are also more likely to flag misinformation than true information. Thus, our results suggest that while moderation is often initiated by the *High Knowledge types*, others follow.

We find that individuals, particularly the *Low Knowledge types*, refrain from sharing posts flagged by others. Plausible mechanisms for this behavior include updating of user priors that flagged posts constitute misinformation and the concern that sharing these posts may adversely impact their reputation. This suggests that flags can prevent users from spreading misinformation by addressing one principal cause of its spread, i.e., confusion regarding the veracity of the post (Pennycook et al. 2021).

We also find that users are less likely to flag a post when their beliefs align with the post. However, if such posts are false, the impact is insignificant. Generally, users are less willing to scrutinize the correctness of a post before sharing, specifically if it is aligned with their beliefs (Bakshy et al. 2015). However, we find that if such a post is flagged users refrain from sharing it. Thus, we demonstrate that flags can serve as an effective cue, which discourages users with low knowledge and/or aligned beliefs from sharing posts containing misinformation. The underlying mechanisms that explain the reduced sharing of a flagged post are updated priors regarding the

[9] While misinformation can include false as well as controversial information, for the purpose of this study, we focus only on factual information that has been verified as false.

accuracy of the post and the individual's self-presentation views. We also find that users are less likely to flag a post and more likely to share a post that has been shared by others after controlling for the presence of flags. This suggests that positive social media cues like shares help to propagate misinformation as users are even less likely to report misinformation and keep sharing misinformation in the presence of such cues.

We confirm the overall effect of users' flagging and sharing behavior on reducing the spread of misinformation by running a simulation using a snapshot of a network of Facebook users from the Stanford Large Network Dataset Collection (Leskovec and Mcauley 2012). The ego network is widely used in social media literature (e.g., Xu et al. 2020). We run the simulation using the flagging and sharing probabilities from our experiment in an echo-chamber-like setting where we consider users' beliefs to be homogeneous to assess the efficacy of the flag feature. The simulation results show that implementing the flag design - i.e., crowd-based moderation - can curb the spread of misinformation even when users' beliefs are homogeneous while still allowing the propagation of truthful information.

Our research contributes to the growing body of literature on misinformation. Previous studies have separately considered content moderation by crowd without considering social cues (Martel et al. 2022, Wang et al. 2021) or propagation of misinformation by users assuming the availability of moderated content (Jahanbakhsh et al. 2021, Moravec et al. 2020, Pennycook et al. 2020b). We add to this literature in several ways. First, we illustrate how our proposed design can leverage the crowd to both moderate content and prevent the sharing of misinformation. Additionally, our approach is effective even if users are not knowledgeable or if their beliefs align with the misinformation. Second, we demonstrate the efficacy of our approach in a social setup where cues are visible. Third, we establish the role of factors such as belief updates and self-presentation in this process.

Managerial Implications: Our study also provides managerial guidance to social media platforms for dealing with the misinformation challenge. Apart from demonstrating that social cues involving flags can be beneficial, our results also point to the limitations in the current approach of social platforms wherein they fail to make use of users' reports on misinformation. In most social media platforms including Facebook, the cues shown to users are only positive such as likes and shares, without any explicit feature to suggest to others that the post constitutes misinformation. We show that in the presence of only positive cues users are even less likely to report misinformation, creating a bandwagon effect and amplifying the spread of misinformation. While users have the option of reporting content to the platform, such reports are hidden from other users. The drawbacks of the current platform strategy of hidden reports or flags are twofold: (i) It does not alert other users to scrutinize posts before sharing, and (ii) In the current scenario, when cues available to users are only positive (such as likes and shares), the probability that users would report misinformation is low, particularly when the post is shared or liked by many, as posts shared by others are deemed to be correct. Therefore, platforms should consider introducing a flagging feature that provides visible cues to other users to help curb the spread of misinformation.

The balance of the paper is organized as follows. In the next section, we present the prior literature and identify gaps that motivate the investigation of crowd-driven content moderation. We also describe the theoretical underpinnings for flagging and sharing behavior. Section 3 provides a details of the experiment. Section 4 provides an overview of the the models used in our analyses and the results. We present the discussion of results in section 5, and conclude in section 6.

1.1. Theoretical Background

In this section, we elaborate upon the theories that may explain users' decisions to flag information, use such cues in their sharing decisions, and the subsequent impact of such decisions on the propagation of misinformation.

Flagging Behavior: The initial cues or the first flags are critical and fateful (Park et al. 2021) as they will influence the subsequent flags. Therefore, it is important to understand who provides these first flags and whether it is biased or correct. To understand whether and how users will offer initial flags, we draw from the Theory of Reasoned Action (TRA) developed by Ajzen and Fishbein (1975, 1980). TRA presents a framework to understand how people make behavioral decisions. TRA suggests that two key factors predict behavior: a person's perception of social norms and attitudes (e.g., beliefs, knowledge). Specifically, people are more likely to do something when they both view it positively themselves and believe it's valued by their community.

Based on the social norm component of the TRA, which states that one's actions are motivated by their usefulness to the community, users are likely to share information if they perceive this will help inform others about the veracity of the post. Prior literature shows that users' contributions on online platforms are driven by the intent to inform and help others (e.g., Fang and Zhang 2019). In the context of reviews, one of the primary motivations for users' feedback on the platforms is the individual's experience of a product or a service (Sen and Lerman 2007), wherein they intend to warn others or improve the quality of a product or service with ratings and reviews.

Drawing upon the theory of reasoned action Lin (2007) suggests that knowledge self-efficacy, which is correlated with one's expertise or knowledge, is critical in sharing information. Similarly, Experts or users with more knowledge were found to be more effective at transferring valuable knowledge to others, according to Constant et al. (1996). Along similar lines, Matzler et al. (2008) state that openness, which is also predicted by expertise (Cabrera et al. 2006), predicts sharing information with others. Accordingly, one may expect that users with knowledge are more likely to provide the initial flags. Therefore, building on insights from the theory of reasoned action, we hypothesize that users with more information or knowledge may be more aware of the veracity of a post and, therefore, more inclined to flag posts to inform others.

Again, drawing upon the TRA framework, individuals' beliefs shape their attitudes (Dillard 2002), which may subsequently motivate their actions. In the context of misinformation, prior studies, such as Pennycook et al. (2021), have argued that users' beliefs or alignment is a critical driver of their actions on social media. Pereira et al. (2023) state that partisanship or beliefs are associated with support of fake news. Moreover, research by Cohen (2003) suggest that partisanship may also override value-driven preferences. Beliefs, e.g., political partisanship, strongly influence both how people judge truthfulness and what they choose to share (Gawronski et al. 2023). For example, a pro-democrat user may be motivated to flag a pro-republican post as it may not align with their preferences, possibly due to their biased perception of truth. Therefore, using the preference-based account, wherein beliefs influence users' actions, it is also possible that the first flags are biased, i.e., motivated by users' beliefs. Hence, we theorize that users with misaligned beliefs are more likely to provide initial flags, motivated by their preferences, irrespective of the veracity of the post.

In summary, the expected motivation for providing negative cues on social media is not straightforward. It seems to depend not only on their knowledge but also on their ideological dispositions or beliefs. If the first flags are placed because the users have knowledge it is likely to be correct and make the feature effective in reducing misinformation. However, if the first flags are biased and motivated only by users' beliefs, they may provide noisy signals to other users. To summarize, initial cues are critical in predicting the quality of the final outcome, as subsequent users are likely to herd. Therefore, we evaluate if users with knowledge initiate flagging (and whether they are more likely to flag misinformation) or if the initials flags are biased by users' beliefs. The expected efficacy of flagging or moderation is likely to depend on whether we find users to provide the first flag based on users' knowledge or their beliefs.

Sharing in the Presence of Flags: To evaluate the efficacy of the flag feature in reducing misinformation, it is important to assess sharing behavior in the presence of flags. Particularly, we assess how users with low knowledge and aligned beliefs account for flags in their sharing behavior. These parameters are particularly important to assess as the spread of misinformation is primarily due to sharing by users with low information and when beliefs are aligned Pennycook et al. 2021.

Flagging systems may inhibit information-sharing behaviors by activating users' concerns regarding potential reputational harm that could result from being identified as disseminators of inaccurate content. Such reputational concerns align with the social norm component of the

Theory of Reasoned Action, whereby anticipated social norms significantly influence behavioral intentions (Fishbein Ajzen, 1975; Bock et al., 2005). Moreover, flags may help shape the opinion of the users with respect to the accuracy of the information. For example, Luo et al. (2022) and Chaiken (1987) show that social endorsement cues (such as likes or shares) on a post increase its credibility and are perceived as a sign of correctness. Likewise, the presence of flags may influence users' opinions regarding the correctness of the post and thereby modify their choice of action (i.e., sharing or flagging) on the post. While users may want to share accurate information, they may be confused about the accuracy of the post and share it without scrutiny (Pennycook et al. 2021, Avram et al. 2020). Along similar lines, prior studies such as Lee et al. (2020) have shown that trusting a post that constitutes misinformation is associated with poorer knowledge. The presence of flags may help users to update their priors regarding the veracity of the information, and as a consequence, they may refrain from sharing.

Users' characteristics, such as their political beliefs, may cause them to discount the value of negative reviews (Sen and Lerman 2007). Collective wisdom, as explored in previous research (Mannes 2009), may be underappreciated, leading users to ignore flags and still share content. The preference-based theory of misinformation sharing posits that people may deliberately share false content when it supports their beliefs, e.g., weighing partisan alignment more heavily than factual accuracy. Pennycook et al. (2021) find that this explains about 15 percent of the misinformation shares; individuals shared ideologically aligned headlines despite knowing they were false, suggesting they consciously valued partisan considerations above truthfulness. This same phenomenon could apply to flagged content, where users may choose to ignore flags if the content aligns with their beliefs and proceed to share it.

As beliefs strongly influence both how people judge truthfulness and what they choose to share, along with their ability to separate truth from false information (Gawronski et al. 2023), for flags to be effective among such users, they should be able to subside this belief based influence. Flags could potentially deter users with aligned beliefs by enabling them to better differentiate truth from false information, as users typically tend to pay heed to negative cues or feedback. Alternatively, flags may discourage users from sharing by triggering concerns about reputation damage from being seen as spreading misinformation. Thus, one may expect that users may refrain from sharing posts that are flagged by others, even if their beliefs are aligned, as it enables them to better differentiate between true and false information or if they perceive that sharing a flagged post may hurt their reputation. Therefore, users with aligned beliefs may refrain from sharing flagged posts.

Overall, the influence of flags on users' sharing behavior and the effectiveness of reducing misinformation through flagging are multifaceted and require further exploration. If users view flags as credible signals, their confusion regarding the accuracy of the post is likely to be resolved, i.e., individuals will update their priors regarding the accuracy of the post. Such a mechanism, if dominant, will help to curb the flow of misinformation. However, if such cues are merely seen as group-conforming instruments, individuals are likely to discount flags in their sharing decisions. Evaluating these underlying mechanisms, our study aims to add to the social media literature by highlighting users' responses to negative cues and, subsequently, the spread of misinformation.

Our study builds upon the Theory of Reasoned Action (TRA) to assess how both knowledge-based attitudes and social norms influence users' decisions to flag misinformation. Second, it integrates TRA with preference-based theories of information sharing to develop insights regarding what dominates users' behavior in providing negative cues on social media in the context of misinformation. Finally, our study advances our understanding of how social cues like flags are mediated by individual characteristics (knowledge levels and beliefs) to influence sharing decisions, building on previous work about social endorsement and collective wisdom. Finally, by examining how reputation concerns and accuracy motivations compete with partisan alignment in users' decisions to share flagged content, the study enriches theories about motivated reasoning and information sharing in social media environments. These theoretical advances would provide a more nuanced framework for understanding how platform design features like flagging systems shape the spread of misinformation.

2. Experimental Design

We deploy an experimental approach to assess the impact of flags on users' sharing and flagging behavior. Laboratory experiments are a powerful research tool, particularly in scenarios like ours, where we propose a new design feature for social media platforms to combat misinformation. In our context, conducting field studies is challenging, as existing platforms lack such features. Several prior studies in the Information Systems (e.g., Bapna et al. 2017, Adomavicius et al. 2012) and other streams of literature (e.g., Cao and Smith (2021), Pennycook et al. (2020a), etc.) have relied on an experimental framework and the Amazon Mechanical Turk (AMT) platform to assess user behavior in a social context. Moreover, most prior studies on misinformation have been conducted through surveys. Blascovich et al. (2002) suggest that one of the challenges of experiments in social science research is the "experimental control–mundane realism trade-off".[10] To overcome this challenge, we developed a social media application for our experiment, designed to provide users

[10] Experimental control refers to precise manipulation of independent variables and mundane realism refers to the extent to which an experiment replicates the real environment (Blascovich et al. 2002)

with an experience similar to that of Facebook. This application allows users to interact with social media posts in a manner akin to the real application. In the following subsections, we provide a detailed description of the participants, materials, and procedures.

2.1. Participants

Following prior literature (e.g., Kim et al. 2019) on fake news that rely on experimental settings, we recruited participants from the AMT platform and conducted interactive sessions using our social media application. Coppock (2019) suggest that the AMT platform produces similar results to nationally representative samples; therefore, we engaged a large sample of 1458 U.S. residents for our study. From this sample, we excluded 118 participants who failed attention checks or to complete the experiment to ensure the quality of the responses.

To ensure that the participants are representative of the U.S. population on social media, we compared the observed demographics of our participants to that of the social media platforms as shown by Statista.[11] The distribution of our participants in terms of age, education, and gender is similar to that of the social media users in the U.S. The mean age of the participants was 39 years, and 54 percent of the participants were women. In our sample of participants, about 46 percent of users are Democrats, and 74 percent of participants were pro-vaccine or vaccinated. Among the participants in the Russia-Ukraine category, 74 percent supported Ukraine, and 23 percent of participants were indifferent.

2.2. Procedure

We first familiarized the participants with the context, i.e., the meaning of flags and shares. If the participants chose to continue, they were directed to the social media application.

Inferring friends using Names Generator Technique: In order to create an immersive social media environment for the study participants, we also provide cues from their friends such as flags and shares. These serve as controls in our study and help confirm the validity of our experiment. As we do not have access to the actual social media networks of the participants, we use the names generator technique to collect friend names. Proposed by Burt (1984), this technique is widely utilized in social network research. It serves as a fundamental method for constructing a user's social networks and understanding the characteristics of these connections. The U.S. General Social Survey (GSS) employs this approach extensively to map out social networks and their tie characteristics. Similarly, many studies in the social network/media

[11] The distribution of participants' knowledge, age group, and gender are similar to the statistics provided by https://www.statista.com/statistics/1337578/us-distribution-leading-social-media-platforms-by-education/, https://www.statista.com/statistics/1337525/us-distribution-leading-social-media-platforms-by-age-group/, https://www.statista.com/statistics/1337563/us-distribution-leading-social-media-platforms-by-gender/ respectively.

literature, including Stadtfeld et al. (2019), leverage this technique to generate respondents' social networks for research purposes. Typically, this method is implemented via surveys and interviews to compile a list of friends, as noted by Burt (1984). Accordingly, participants in our setup, upon arriving at the landing page of the social media application, are prompted to provide the names of five of their social media friends. We use these names in a randomized way to show cues to the users, as described below.

2.3. Content

The participants were randomly assigned to see posts from the following categories: Politics, COVID-19, and Russia-Ukraine conflict. Several studies in the misinformation literature such as Pennycook and Rand (2021), Dias et al. (2020), etc., have used the context of politics and COVID-19 in their studies. These topics experienced a deluge of misinformation, and therefore, provide relevant context for our study. Posts were aggregated from PolitiFact (Gillin 2018), a third-party website that fact-checks news stories.

Posts and Cues: We used a total of 20 posts across the three categories (similar to Pennycook et al. 2020a). The number of posts used in our experiment is similar to that of prior studies in the literature (Pennycook et al. 2020a, Moravec et al. 2020). The posts were randomly selected from a pool of about fifty posts from Politifact. We follow a similar approach as that of Pennycook et al. (2020a) and show the participants a mix of pro-Republican and pro-Democrat posts related to politics. Following a similar pattern, we show a mix of pro-vaccination and anti-vaccination posts in the COVID category and pro-Russia and pro-Ukraine mix in the Russia-Ukraine group. Moreover, in each category, participants were shown both true and false posts in a random order. For example, a user in the Politics group will see both pro-republican and pro-democratic posts along with true and false posts of the respective types.

The posts were randomly assigned the number of flags and shares, and whether a friend had flagged or shared it. The number of flags (shares) was chosen from three categories: (i) zero flags (shares), (ii) low flags (shares) ranging from 3 to 5, (iii) high flags (shares) ranging from 1800 to 2000.[12] Major social media platforms such as LinkedIn, Instagram, etc. separately show a friend's endorsement from that of the crowd. For example, a post on Instagram is shown as "Liked by *Friend's name and 50 others*". Moreover, several studies in the prior literature (e.g., Wang et al. 2018) have shown a heterogeneous impact of a friend's cues versus the crowd's cues. Therefore, following the design of the social media platforms and the findings of the prior literature, we show friend and crowd cues separately. Given the design (space) constraints and the norm followed by

[12] We follow Jang et al. (2015), which suggests that the average number of likes on Instagram is approximately 1984.

social media platforms like LinkedIn and Facebook (in terms of likes and shares), we show either 0 or 1 friend flag (share). The name of the friend on the flag or share cue is randomly assigned. The participants could flag or share the posts or choose to do nothing.

Each participant interacted with six to eight different posts and chose to flag, share, or continue without taking any action. Next, we ask a series of questions to understand the participant's opinions about the post, knowledge about the context, and their beliefs. Finally, we also collect the demographics of the participants, and using Pennycook et al. (2020a), we ask a series of Cognitive Reflection Test (CRT) questions, which capture a user's disposition to think analytically even in psychological phenomena such as belief and identity (Stagnaro et al. 2018). We follow prior studies such as Osmundsen et al. (2021), Calvillo et al. (2020) for formulating the questions to test the knowledge of individuals in politics and COVID-19 instances. To understand the knowledge of users in the Russia-Ukraine category, we follow the approach of Huang (2015), wherein we ask questions about political figures and recent events in Russia and Ukraine. Similarly, for identifying the beliefs or ideological inclination of participants, we follow the approach of prior studies such as Pennycook et al. (2020a). The questions for knowledge and beliefs are presented in the supplementary material for each category along with the posts. Both CRT and knowledge tests had acceptable reliability (Cronbach's $\alpha > 0.6$).[13]

To further ensure the validity of the experiment, we compared the behavior of users in our study to that of the prior lab and field experiments, and users' self-reported characteristics and found it to be consistent.

3. Model and Results

Our objective is to determine the effects of social cues - flags and shares - on a user's propensity to flag and share posts and, subsequently, on the spread of misinformation, along with plausible underlying mechanisms. The summary statistics and the description of the variables used in our analysis are presented in Tables 1 and 2, respectively.

3.1. Who Flags First

To evaluate the effectiveness of the flags, it is important to ascertain whether users will initially flag a post. This initial action is significant because the first cue is considered “fateful” (Park et al. 2021), as subsequent cues may be influenced by earlier ones. For this analysis, we focus on a subsample of observations where posts were presented to users without any flags. We also assess the impact of users' knowledge and their beliefs (ideological alignment) on their flagging behavior.

[13] Cronbach's alpha or coefficient alpha, is the most common test score reliability measure.

Table 1 Summary Statistics

Variable	Mean	Std. Dev.	Min.	Max.	N
Share	0.28	0.45	0	1	8597
Flag	0.26	0.43	0	1	8597
HighFlag	0.34	0.47	0	1	8597
FriendFlagged	0.49	0.5	0	1	8597
HighShare	0.32	0.47	0	1	8597
FriendShare	0.50	0.5	0	1	8597
True	0.41	0.49	0	1	8597
Agree	2.69	1.21	1	5	8597
Correct	2.81	1.24	1	5	8597
Controversial	3.66	1.14	1	5	8597
Popular	2.30	1.26	1	5	8597
ShareLookGood	0.89	1.25	1	5	8597
FlaggingLookGood	0.92	1.30	1	5	8597
Discussion	3.38	1.16	1	5	8597
Interesting	3.18	1.21	1	5	8597
View	0.61	0.49	0	1	8597
Knowledge	0.463	0.499	0	1	8597
BeliefAligned	0.446	0.497	0	1	8597

Table 2 Variables Description

Variable	Description
Share	= 1 if user shares the post, 0 otherwise
Flag	= 1 if user flags the post, 0 otherwise
HighFlag	= 1 if the post shown to the user has high number (1800-2000) of flags, 0 otherwise
FriendFlagged	= 1 if the post shown to the user has friend's flag, 0 otherwise
HighShare	= 1 if the post shown to the user has high number (1800-2000) of shares, 0 otherwise
FriendShare	= 1 if the post shown to the user has friend's share, 0 otherwise
True	= 1 if the post shown to the user is true, 0 otherwise
BeliefAligned	= 1 if the post shown to the user aligns with their beliefs, 0 otherwise
Knowledge	= 1 if the user if the number of knowledge questions correctly answered by the user is above the median value, 0 otherwise
Agree	varies from 1 (strongly disagree with the post) to 5 (strongly agree with the post)
Correct	varies from 1 (strongly disagrees that the post is correct) to 5 (strongly agrees that the post is correct)
Controversial	varies from 1 (strongly disagrees that the post is controversial) to 5 (strongly agrees that the post is controversial)
Popular	varies from 1 (strongly disagrees that the post is popular) to 5 (strongly agrees that the post is popular)
ShareLookGood	varies from 1 (strongly disagrees that sharing the post will make the user look good) to 5 (strongly agrees that sharing the post will make the user look good)
FlaggingLookGood	varies from 1 (strongly disagrees that flagging the post will make the user look good) to 5 (strongly agrees that flagging the post will make the user look good)
Discussion	varies from 1 (strongly disagrees that the post will generate discussion) to 5 (strongly agrees that the post will generate discussion)
Interesting	varies from 1 (strongly disagrees that the post is interesting) to 5 (strongly agrees that the post is interesting)
View	=1 if user clicks on view more button on the post, 0 otherwise

The efficiency of the flag feature hinges on users flagging content correctly, i.e., not flagging true information due to biases in their ideological alignment but based on their understanding of the content's veracity. Thus, we evaluate the flagging behavior for both true and false posts.

As before, our unit of data involves a post and a user. We use the following model specification:

$$\begin{aligned} DV_{ip} = {} & \sigma_0 + \sigma_1 Knowledge_i X False_p + \sigma_2 Knowledge_i + \sigma_3 BeliefAligned_i X False_p + \\ & \sigma_4 BeliefAligned_i + \sigma_5 High\_Share_{ip} + \sigma_6 Friend\_Share_{ip} + \sigma_7 Controls_i + \phi_p + \zeta_{ip} \quad (1) \end{aligned}$$

where DV_{ip} represents the dependent variables in our model, i.e., whether user i flags post p; $Knowledge_i$ is a dummy which takes the value 1 if the user is more knowledgable (i.e., above the sample mean), 0 otherwise; $BeliefAligned_i$ takes the value 1 if the beliefs of the user are aligned with the post, 0 otherwise. $False_p$ dummy takes the value 0 if the post p is true and 1 otherwise. $High\_Share_{ip}$ ($Friend\_Share_{ip}$) takes the value 1 if post p is shown to the user i with a high number of shares (friend's shares), zero otherwise. $Controls_i$ represent controls used in the model, i.e., user i's education, IQ, race, age, minutes spent on social media, self-reported propensity to share without reading, and the number of shares on the post; ϕ_p stands for post-fixed effects and ζ_{ip} represents the error term. As post characteristics are absorbed in the fixed effect specification, we do not include any post-specific control in the model.

We estimate this model using logistic regression with robust standard errors clustered on both participants and posts and our results as presented in specifications 2 and 3 of Table 3. Specification 2 of Table 3 includes post-fixed effects and specification 3 is shown without post fixed effects. The results indicate that *High Knowledge types* (as shown by the variable $Knowledge$) are more likely to initiate the flags on a post or be the first one to flag a post. Interestingly, we also find that if users' beliefs are aligned, and the post has zero flags, they refrain from flagging the posts as shown by the coefficient of the variable $BeliefAligned$.

In specification 1 of Table 3, we interact $Knowledge$ with $False$ dummy to understand whether the *High Knowledge types* are more likely to flag only misinformation or even the true ones; we find that such users flag first and flag misinformation as shown by the coefficient of $Knowledge$ X $False$; the coefficient of $Knowledge$ in specification 1 is insignificant. On average users are less likely to flag posts when their beliefs are aligned (as shown by the coefficient of $BeliefAligned$ in specification 2); however, for misinformation, i.e., false posts, the bias is insignificant (as shown by the coefficient of $BeliefAligned$ X $False$ in specification 1).

We find that *High Knowledge types* are the ones who are likely to flag first and more likely to flag misinformation than *Low Knowledge types*. Our findings suggest that users are likely to make use of the flag feature objectively, i.e., based on their knowledge, and not based on their beliefs. We also

Table 3 Who Flags First

	(1) Flag	(2) Flag	(3) Flag
Knowledge	-0.131	0.309^{**}	0.321^{***}
	(0.21)	(0.13)	(0.10)
Knowledge X False	0.694^{***}		
	(0.24)		
BeliefAligned	-0.504	-0.666^{**}	-0.516^{*}
	(0.43)	(0.31)	(0.31)
BeliefAligned X False	-0.270		
	(0.59)		
HighShare	-0.205^{**}	-0.190^{*}	-0.169^{**}
	(0.10)	(0.10)	(0.08)
FriendShared	0.044	0.030	0.043
	(0.18)	(0.18)	(0.17)
Constant	-2.427^{**}	-2.571^{**}	-2.840^{***}
	(0.99)	(1.09)	(1.08)
Observations	1423	1423	1423
Pseudo R^2	0.168	0.164	0.088
PostFE	Yes	Yes	No
OtherControls	Yes	Yes	Yes

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

find that users are less likely to flag when the post has a high number of shares. This suggests that users may be reluctant to flag a popular post. In the next section, we explore plausible mechanisms to understand users' motivation to flag posts and the effect of existing flags on their flagging and sharing behavior.

3.2. Impact of Existing Flags on Flags and Shares

Prior literature suggests that existing social cues such as likes and shares may update the opinion of the users regarding the veracity of the post (Luo et al. 2022) and also impact how others may perceive them (Hennig-Thurau et al. 2004). Similarly, existing flags may affect users' propensity to flag or share. Thus, we analyze the impact of existing flags on the subsequent flagging and sharing behavior of a user with the following model specification:

$$DV_{ip} = \sigma_0 + \sigma_1 High\_Flag_{ip} + \sigma_2 Friend\_Flag_{ip} + \sigma_3 High\_Share_{ip} + \sigma_4 Friend\_Share_{ip} + \sigma_5 Controls_i + \phi_p + \zeta_{ip} \quad (2)$$

where DV_{ip} represents the dependent variables in our model, i.e., share or flag by user i on post p; $High\_Flag_{ip}$ and $High\_Share_{ip}$ take the value 1 if post p is shown to the user i with a high number of flags and shares respectively, zero otherwise; $Friend\_Flag_{ip}$ and $Friend\_Share_{ip}$ take the value 1 if post p is shown to the user i with friend's flag and share respectively, zero otherwise; Controls used in the model represented by $Controls_i$ include user i's knowledge, belief, education, IQ, race, age, minutes spent on social media per day, and self-reported propensity to share without reading; ϕ_p stands for post-fixed effects and ζ_{ip} represents the error term.

The analyses for the impact of cues on users' flagging behavior are shown in specification 2 of Table 4. The coefficient of $HighFlag$ is positive and significant, suggesting that users are more likely to flag posts that are flagged by the crowd. This result shows that existing flags can encourage users to moderate content. Similarly, users are also likely to flag posts that are flagged by friends as suggested by the coefficient of $FriendFlagged$. However, users are less likely to flag posts that are shared by many others as shown by the estimates of $HighShare$. While Wang et al. (2021) show that post content can increase the propensity to identify misinformation, our result suggests how different cues can influence user contribution to content moderation.

The effect of cues on the users' sharing behavior is shown in specification 1 of Table 4. We find that users are less likely to share posts that are flagged by the crowd ($HighFlag$). This suggests that users are likely to recognize user generated cues for misinformation and curb the sharing of such content. Previous studies (Dias et al. 2020, Moravec et al. 2020, Pennycook et al. 2019, 2020a, Jahanbakhsh et al. 2021) show such effect only for warnings from external providers or the platform. A friend's flag ($FriendFlagged$) also produces a marginal negative effect on sharing. The baseline in our analysis is the control condition wherein there are zero or low flags (shares). For robustness, we also use a model (discussed in section 4.5) where the baseline is zero flags (shares).

Overall, we find that existing flags attract more flags and reduce subsequent sharing behavior. This result highlights that users take flags into account in their behavior. However, the efficacy of this approach to curb misinformation likely depends on two important factors: how users with aligned beliefs and *Low Knowledge types* utilize flags in their flagging and sharing behavior, as these users play a central role in spreading misinformation (Bakshy et al. 2015, Pennycook et al. 2021). Given that existing flags influence the behavior of subsequent users who see the post, it is also important to assess who initiates the flags (Park et al. 2021), and whether these flags are initiated based on users' information or biased by their beliefs. We investigate these aspects in the subsequent sections.

The current design of platforms like Facebook has a Report feature that enables a user to report misinformation to the platform and the cues seen by users are in the form of positive endorsements such as likes and shares. Our results suggest that it is likely that the misinformation that is spreading (i.e., with more shares) is even less likely to be reported in the absence of negative cues

Table 4 Effect of Cues on Users' Sharing and Flagging Action

	(1) Share	(2) Flag
HighFlag	-0.205***	0.308***
	(0.067)	(0.075)
FriendFlagged	-0.076*	0.264***
	(0.040)	(0.050)
HighShare	0.166***	-0.166***
	(0.046)	(0.040)
FriendShared	0.124***	-0.009
	(0.039)	(0.051)
Knowledge	0.071	0.092
	(0.072)	(0.079)
BeliefAligned	0.379*	-0.520**
	(0.195)	(0.239)
Observations	8597	8597
Pseudo R^2	0.075	0.130
PostFE	Yes	Yes
OtherControls	Yes	Yes

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

in the current setting of Facebook.

We also validate whether users' behavior in the experiment is aligned with their actual social media behavior as reported in prior studies. We find that users are more likely to share posts that are shared by their friends ($FriendShared$), and by other users ($HighShare$) as shown in specification 1 of Table 4. This is in line with the findings of prior studies such as Bakshy et al. (2012), Epstein et al. (2022), etc. Moreover, as stated by studies such as Allcott and Gentzkow (2017), we find that users are more likely to share when their beliefs are aligned with the post ($BeliefAligned$). The results in specification 1 of Table 4 also highlight that users are more likely to share if they report spending more time on social media as shown by the coefficients of the variable $Mins\_SocialMedia$ (Chang and Hsiao 2014). The above results suggest that users' behavior in the experiment is representative of their actual social media behavior as stated by prior studies.

3.3. Role of Knowledge and Beliefs

Current research indicates that misinformation is primarily disseminated because individuals are either confused about or unaware of the truthfulness of the information (Pennycook et al. (2021)), or because they do not make an effort to verify the accuracy of information before sharing it, especially if it conforms to their existing beliefs (Bakshy et al. 2015). Thus, it is important to assess how the presence of flags would influence the spread of misinformation for less knowledgeable individuals and for individuals with aligned beliefs. Similarly, knowledge and belief may also impact users' flagging behavior. Thus, we assess how users provide and make use of the flags when the beliefs are aligned, and users are of *Low Knowledge types*. We show these results in Tables 5 and 6.

Specification 2 of Table 5 suggests that the presence of flags increases the likelihood of a user flagging the post, irrespective of their beliefs (calculated using the coefficients of $HighFlag$ and $HighFlag$ X $BeliefAligned$). Similarly, specification 2 of Table 6 shows that the coefficient of the interaction term $HighFlag$ X $Knowledge$ is insignificant. This suggests that existing flags increase users' propensity to flag the post, for both types (low and high knowledge types). Taken together, these two results underscore that a flagged post is more likely to be flagged by others.

Table 5 Moderating Effect of Belief

	(1) Share	(2) Flag
HighFlag	-0.239***	0.314***
	(0.084)	(0.076)
HighFlag X BeliefAligned	0.067	-0.014
	(0.098)	(0.106)
FriendFlagged	-0.138**	0.223***
	(0.057)	(0.078)
FriendFlag X BeliefAligned	0.129	0.102
	(0.080)	(0.138)
HighShare	0.167***	-0.167***
	(0.046)	(0.040)
FriendShared	0.123***	-0.010
	(0.038)	(0.051)
Observations	8597	8597
Pseudo R^2	0.076	0.130
PostFE	Yes	Yes
OtherControls	Yes	Yes

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Specification 1 of Table 5 suggests that even if beliefs are aligned users are less likely to share posts flagged by other users (calculated using the coefficients of $HighFlag$ and $HighFlag$ X $BeliefAligned$). Specification 1 of Table 6 shows that existing flags can curb the sharing behavior of *Low Knowledge types*; however, it has an insignificant impact on the sharing behavior of *High Knowledge types* (as shown by the coefficients of $HighFlag$ and $HighFlag$ X $Knowledge$). These findings help us answer the second part of our research question: ***How will the presence of flags impact users' sharing behavior for low knowledge types, and when beliefs are aligned?***

Table 6 Moderating Effect of Knowledge

	(1) Share	(2) Flag
HighFlag	-0.354***	0.347***
	(0.090)	(0.090)
HighFlag X Knowledge	0.314***	-0.081
	(0.111)	(0.113)
FriendFlagged	-0.099	0.322***
	(0.062)	(0.076)
FriendFlag X Knowledge	0.051	-0.121
	(0.097)	(0.098)
HighShare	0.164***	-0.167***
	(0.046)	(0.040)
FriendShared	0.122***	-0.010
	(0.038)	(0.050)
Observations	8597	8597
Pseudo R^2	0.076	0.131
PostFE	Yes	Yes
OtherControls	Yes	Yes

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Our results highlight that existing flags on a post can curb the sharing behavior of users even when their beliefs are aligned and they are of *Low Knowledge type*. These findings are encouraging as flags can overcome the effect of any confusion and beliefs and effectively curb sharing in both these scenarios.

3.4. Mechanisms

In addition to intrinsic factors, i.e., users' knowledge and beliefs, the impact of flags on a post on their sharing and flagging actions could be driven by factors such as updating of priors (Luo et al. 2022) and self-presentation concerns (Sundaram et al. 1998, Marwick and Boyd 2011).

To explore underlying mechanisms at play, we asked the users a series of questions and collected their responses on a 5-point Likert scale (Strongly disagree - Strongly agree). The questions, corresponding mechanisms, and the support from literature are shown in Table 7. To investigate the mechanisms, we conducted a mediation analysis (Foerderer et al. 2018) using a two-step approach. In the first step, we estimate the model shown in Equation 2 with dependent variables capturing user responses to questions corresponding to different mechanisms. We use an ordered logistic regression model as the responses are on an ordered 5-point Likert scale. In step 2, we included the evaluations from step 1 as control variables in the regressions for both the flagging and sharing decision.

Table 7 Questions Related to Mechanisms

Questions/Action	Mechanisms
I think this post is correct (Yaqub et al. 2020)	Update of Priors
I agree with this post (Yaqub et al. 2020)	Update of Priors
I think this post is controversial (Kim and Ihm 2020)	Update of Priors
I think this post is interesting (Bakshy et al. 2011, Yaqub et al. 2020)	Self-Presentation
I think this post is popular (Berger 2014, Ritson and Elliott 1999)	Self-Presentation
I think this post will generate discussion among my friends (Yaqub et al. 2020)	Self-Presentation
Flagging this post will make me look good (Yaqub et al. 2020, Berger 2014)	Self-Presentation
Sharing this post will make me look good (Yaqub et al. 2020, Berger 2014)	Self-Presentation

3.4.1. Update of Priors: Specifications 1, 2, and 3 of Table 8 suggest that users believe that the posts flagged by the crowd ($HighFlag$) constitute misinformation and are controversial,[14] regardless of ground truth, and that they are less likely to agree with such posts. Thus, the presence of flags on a post impacts users' opinions regarding the veracity of the post, which, in turn, may drive users' decision to flag the post as they want to inform others. Interestingly, though, the same is not true for friends' flags, i.e., users do not update their priors in terms of the correctness or being agreeable, however, they do associate the presence of friends' flags on a post with it being controversial and this opinion update may increase their propensity to flag the post.

At the same time, users believe that the posts shared by the crowd are correct and less controversial and they are more likely to agree with such posts. As a result, they are less likely to flag such posts. Friend's shares have a similar but marginal effect on users' beliefs. Overall, crowd flags are effective in terms of shaping users' opinion regarding the correctness of the posts. This points to the inefficiency of the current strategy used by platforms such as Facebook. Currently, users may report a post to the platforms if they deem it to be incorrect, however, other users do not see these reports. The only cue available to them is in the form of social endorsement (likes and shares); so, if a fake piece of information is shared by many they are less likely to be reported as social endorsement is associated with accuracy. This underscores the need to provide the users with additional cues to assess the accuracy of the posts.

3.4.2. Self-Presentation: Specifications 4-6 of Table 8 suggest how cues, i.e., flags and shares, impact users' notion that the posts are popular, generate discussion, and are interesting for others. Posts flagged by the crowd ($HighFlag$) are perceived as popular (specification 4) but not interesting enough to generate discussion. However, posts shared by friends and the crowd ($HighShare$ and $FriendShared$) are considered to be popular and interesting to generate discussion.

[14] Studies such as Kim and Ihm (2020) suggest that controversial posts are the ones that trigger debate among users, regardless of the veracity of the post.

Specification 8 of Table 8 suggests that users believe that sharing posts that are flagged by many others will adversely impact their reputation, whereas sharing the post shared by many others will positively affect their impression on others. Similarly, Specification 9 of Table 8 suggests that users believe that flagging posts that are flagged by many others and their friends will make them look good, whereas flagging the post shared by many others will negatively affect their social appearance or persona. We extend our analysis in Table 8 to assess whether flags have a differential impact on the above mechanisms for users with varying knowledge and beliefs. The estimates are presented in Table 9. We find that the behavior of users with more knowledge and aligned beliefs is consistent with our results discussed above.

Together, these results suggest that cues associated with a post affect users' impression of how the post and their actions on the post will be perceived by others. Thus, users' decisions to flag or share the post are also likely driven by their need to manage their impressions. This underscores the importance of evaluating the user response in the presence of social cues to capture their actual behavior on social media. As previous studies on misinformation (e.g., Moravec et al. 2020) do not explicitly show such cues while judging user response, these studies are not able to assess the underlying mechanisms for user response. Furthermore, in the absence of cues, user response may not accurately represent their actual social media behavior.

4. Overall Effect Using Simulations

Our experiment helps us assess the individual behavior of users based on their knowledge, beliefs, and other social cues. To further confirm that our design can be effective, it is important to assess how these behaviors will come together for users in a network to prevent the spread of misinformation, especially when the beliefs are aligned with misinformation. To understand the overall effect of flags on the diffusion of true versus false information, we use the probabilities of sharing and flag obtained from the experiment and simulate the diffusion of information on a network (4039 nodes and 88234 edges of anonymous social media users) from Facebook collected by Stanford Large Network Dataset Collection (Leskovec and Mcauley 2012). This social media ego network is used by several papers in the current literature (e.g., Xu et al. 2020).

The algorithm used for the simulation is presented in Table 1. We randomly select the seed node that shares the post first and then calculate the total number of shares and flags received by the post at the end of 15 time periods. Once a post is shared by a node, the post is visible to its neighbors or friends along with the associated cues at the given time. Based on the state of the post (i.e., number of shares and flags) for a given node to which it is visible, we use the probability of node n's actions from our experiment and generate random draws to determine the choice, which can be sharing, flagging, or neither.

Table 8 Impact of Cues on Users Priors and Self-Presentation Views

	Agree	Correct	Controversial	Popular	Discussion	Interesting	ShareLookGood	FlagLookGood
HighFlag	-0.231***	-0.233***	0.196***	0.384***	-0.011	-0.120**	-0.171***	0.210***
	(0.057)	(0.052)	(0.035)	(0.049)	(0.036)	(0.057)	(0.046)	(0.054)
FriendFlagged	-0.053	-0.033	0.103***	0.049	0.090*	-0.044	-0.086***	0.172***
	(0.038)	(0.038)	(0.031)	(0.047)	(0.054)	(0.046)	(0.030)	(0.040)
HighShare	0.168***	0.165***	-0.095**	1.183***	0.201***	0.162***	0.208***	-0.139***
	(0.049)	(0.043)	(0.040)	(0.068)	(0.038)	(0.046)	(0.047)	(0.044)
FriendShared	0.063*	0.053	0.004	0.097**	0.092***	0.135***	0.062	0.013
	(0.038)	(0.034)	(0.038)	(0.048)	(0.022)	(0.032)	(0.038)	(0.046)
Observations	8597	8597	8597	8597	8597	8597	8597	8597
Pseudo R^2	0.069	0.067	0.050	0.042	0.017	0.019	0.157	0.239
PostFE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
OtherControls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table 9 Impact of Cues on Users Priors and Self-Presentation Views (with Knowledge and Beliefs Interaction)

	Agree	Correct	Controversial	Popular	Discussion	Interesting	ShareLookGood	FlagLookGood
HighFlag	-0.288***	-0.259***	0.201***	0.311***	-0.052	-0.203***	-0.271***	0.306***
	(0.076)	(0.078)	(0.072)	(0.068)	(0.070)	(0.079)	(0.056)	(0.064)
HighFlag X Knowledge	0.002	0.018	-0.046	0.129	0.023	0.051	0.154*	-0.168**
	(0.090)	(0.101)	(0.084)	(0.086)	(0.080)	(0.077)	(0.082)	(0.070)
HighFlag X BeliefAligned	0.127	0.041	0.038	0.027	0.067	0.133	0.072	-0.047
	(0.101)	(0.115)	(0.084)	(0.096)	(0.079)	(0.104)	(0.070)	(0.104)
FriendFlagged	-0.054	-0.033	0.102***	0.049	0.090*	-0.044	-0.087***	0.172***
	(0.038)	(0.038)	(0.031)	(0.047)	(0.054)	(0.046)	(0.030)	(0.040)
HighShare	0.169***	0.165***	-0.094**	1.183***	0.201***	0.164***	0.209***	-0.139***
	(0.049)	(0.043)	(0.041)	(0.068)	(0.038)	(0.046)	(0.047)	(0.044)
FriendShared	0.063*	0.053	0.005	0.096**	0.091***	0.134***	0.061	0.015
	(0.038)	(0.034)	(0.038)	(0.048)	(0.022)	(0.032)	(0.038)	(0.046)
Observations	8597	8597	8597	8597	8597	8597	8597	8597
Pseudo R^2	0.069	0.067	0.050	0.042	0.017	0.019	0.157	0.240
PostFE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
OtherControls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

To test our feature, we use the boundary condition, wherein we use the probabilities to share or flag when the beliefs are aligned. This helps us assess the efficiency of our flag feature in an echo-chamber-like environment. If the action of the node is *flag*, the cues are updated for the friends and others accordingly; if the node chooses to *share*, the post is made visible to the friends of node n with updated cues.

The simulation is run for 4 scenarios: (i) true posts with no flag feature (ii) false posts with no flag feature, (iii) true posts with a flag feature (iv) false posts with a flag feature. In cases (i) and (ii) where the flag feature is not available, users can only share the post or do nothing. The cues shown to the users are the number of crowd and friend's share. In cases (iii) and (iv) where the flag feature is present, users can share, flag, or do nothing. The cues shown to the users are the crowd and friends' shares and flags. We simulated each scenario to obtain the total number of shares and the number of users the post was visible to at the end of 15 time periods. Following the prior studies (Pierri et al. 2020, Pham et al. 2020, Juul and Ugander 2021) our measure of diffusion are the number of users who were exposed to the post and the number of users who shared the post. Additionally, we also control for the influence of the seed node which triggers the sharing using its centrality measures (Samanta et al. 2021, Zhang et al. 2020). To assess the diffusion of true and false posts in the presence of flags we use the following model specification:

$$\begin{aligned} DV_p = \sigma_0 + \sigma_1 Flags\_Present_p + \sigma_2 Degree\_Centrality_p + \sigma_3 Betweenness\_Centrality_p + \\ \sigma_4 EigenVector\_Centrality_p + \zeta_p \quad (3) \end{aligned}$$

where DV_p represents the dependent variables in our model: (i) total number of shares on post p (ii) total number of users who viewed post p; $Flags\_Present_p$ represents whether the post was with the flag feature. $Degree\_Centrality_p$, $Betweenness\_Centrality_p$, $EigenVector\_Centrality_p$ represents the centrality measures of the seed node that first shares the post in the network; ζ_p represents the error term. The results are presented in Tables 10 for total user views (spread) and shares. Specification 1 of Table 10 suggests that having the flag feature can reduce the spread of false posts significantly. Consistently, the results presented in specification 2 of Table 10 suggest that having the flag feature can reduce the sharing of false posts significantly. These findings help us answer our research question: ***How will the flag feature impact misinformation?*** Using the simulation and the experiments, we show that the proposed flag feature will indeed be able to curb misinformation significantly.

Algorithm 1: Simulation Algorithm to Estimate the Diffusion of True vs. False Information

```
Input: Graph G = SNAP Facebook Network with 4039 nodes and 88234 edges; Flagging
       and Sharing Probabilities
Output: Total number of nodes that viewed the post ($Total_{Views}$), shared the post
        ($Total_{Shares}$) and flagged the post ($Total_{Flags}$)
Set the $VisibleState_n$, $TimeVisibleState_n$, $FriendFlag_n$, $FriendShare_n$, $AnyFlag_n$,
  $AnyShare_n$ for each node $n$ to 0 and $ActionState_n$ to $Nothing$,
Set the $PShare_n$ and $PFlag_n$ based on whether the post is false or true
Set $T_{Max}$ =15, $Total_{Shares}$ =0, $Total_{Flags}$ =0, and $Total_{Viewed}$ =0
Select a random seed node ($S_n$) that shares the post at $T = 1$.
Set the $VisibleState$ and $TimeVisibleState$ of ($S_n$) to 1 and $ActionState$ to $Share$
For all friends of $S_n$, set the $VisibleState$ to 1 and $TimeVisibleState$ to 2
Increment the $Total_{Views}$ by number of friends of $n$ whose $VisibleState$ is altered
for $T = 2$ to $T_{Max}$ do
    for nodes where $VisibleState$ =1,$ActionState$ =Nothing and $TimeVisibleState >= T - 1$
     do
        • Based on the values of $FriendFlag_n$, $FriendShare_n$, $AnyFlag_n$, $AnyShare_n$ for node $n$
        re-evaluate the probability of sharing ($PShare_n$) and flagging ($PFlag_n$) of the post by node $n$
        • Using the updated probabilities of sharing ($PShare_n$) and flagging ($PFlag_n$) of the post by
        n, draw n's action from the set: $Share, Flag, Nothing$ and update the Action state of node $n$
        ($Action_n$)
        if $Action_n$ == $Share$ then
            Increment the $Total_{Shares}$ by 1
            Increment the $AnyShare_n$ for non-friends and $FriendShare_n$ for friends by 1
            Make the post visible to the friends ($f$) of $n$ who have not seen it yet by
            updating their $VisibleState_f$ to 1, $TimeVisibleState_f$ to $T$
            Increment the $Total_{Views}$ by number of friends of $n$ whose $VisibleState$ is altered
        end
        if $Action_n$ == $Flag$ then
            Increment the $Total_{Flags}$ by 1
            Increment the $AnyFlag_n$ for non-friends and $FriendFlag_n$ for friends by 1
        end
    end
end
```

5. Discussion of Results

Current moderation techniques for dealing with misinformation, which rely on platform intervention, lack efficiency and are often accused of silencing free speech. For example, YouTube removed

Table 10 **Simulation: Impact of Flags on Sharing and Spread of Information**

	(1) Spread	(2) Shares
Flag X False	-0.415**	-0.831***
	(0.163)	(0.165)
Flag	0.141	0.312**
	(0.126)	(0.127)
Degree_Centrality	0.003***	0.003***
	(0.001)	(0.001)
Betweenness_Centrality	0.041**	0.042**
	(0.021)	(0.021)
EigenVector_Centrality	0.030	0.048
	(0.334)	(0.337)
Constant	6.859***	5.647***
	(0.089)	(0.090)
N	1455	1455
Adj. R^2	0.065	0.141

Standard errors in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

several videos spreading misinformation due to a policy created in 2020. However, recently it announced that it will stop removing such content. YouTube stated that "*In the current environment, we find that while removing this content does curb some misinformation, it could also have the unintended effect of curtailing political speech without meaningfully reducing the risk of violence or other real-world harm*".[15] As noted by Google, such platform initiated policy to curb misinformation has not been effective and curtails freedom of speech of users. UN guidelines also call for transparency in content moderation.[16] This creates a need for investigating alternate strategies that are moderated by the crowd. Our study aims to address this issue, by proposing and evaluating the efficiency of a crowd based measure and provides managerial implications for social media platforms.

Existing literature has primarily analyzed platform design focused on nudges, emphasizing source information or third-party fact checks (Dias et al. 2020, Guess et al. 2020, Pennycook et al. 2019). Our study is the first, to the best of our knowledge, to investigate a crowd based strategy to curb misinformation. We propose a flag feature, which provides users the ability to provide visible feedback to other users, which is a deviation from the current design which keeps a user's feedback hidden. X has taken a step in this direction by introducing the Birdwatch program, however, it

[15] See https://www.npr.org/2023/06/02/1179864026/youtube-will-no-longer-take-down-false-claims-about-u-s-elections

[16] https://www.ohchr.org/en/stories/2021/07/moderating-online-content-fighting-harm-or-silencing-dissent

involves platform intervention to a large extent. Moreover, our proposed strategy is different from Birdwatch in two significant ways: (i) We rely on users' self-presentation mechanism in flagging and managing misinformation, however, X's Birdwatch does not ultimately reveal the identity of the Birdwatch commentors to the viewers, and (ii) The cognitive resource and load required in commenting versus flagging (i.e., click a button) is significantly different, and thus, the outcomes may vary (Jiang et al. 2016) considerably.

Our findings underscore that users with more knowledge actively flag misinformation, specifically in the absence of it, and others follow. Prior studies have shown that users with low information and aligned beliefs are major contributors to the spread of misinformation (Bakshy et al. 2015, Pennycook et al. 2021). The presence of flags makes other users more mindful of sharing content, specifically for users who have less information and when the beliefs are aligned. Therefore, flags can significantly reduce the spread of misinformation, by helping users with low information in their sharing decision and the users with aligned beliefs in scrutinizing it.

Users' actions on social media platforms are primarily dictated by their beliefs, knowledge, and their self-presentation concerns (Hennig-Thurau et al. 2004, Pennycook et al. 2021, Luo et al. 2022). The net outcome of how users behave is dependent on both their own characteristics (as mentioned above) and whether the cues they are providing are positive or negative (Amabile 1983). We find that the first users to flag a post do so objectively based on their knowledge, and are less likely to be biased by their beliefs. The first flag is critical, as suggested by Park et al. (2021), and thus, it needs to be objective and correct, for the design to be efficient. The users do so as they believe that flagging will make them look good. This is in line with the theory proposed by Amabile (1983), which states that people provide negative reviews to differentiate themselves, thus appealing to their self-presentation needs. Overall, it is an encouraging outcome that flags are primarily driven by users' knowledge and self-presentation.

Users, specifically the ones with less knowledge, respond correctly to such flags. They update their priors assuming that flagged posts constitute misinformation and that sharing such posts will harm their reputation. Interestingly, flags do not impact the sharing behavior of users with more knowledge. These results underscore that flags are more likely to curb the sharing behavior for users with less information and not for the ones with a high level of knowledge. This shows that the concern raised by Bakshy et al. (2015) and highlighted by Pew research[17] is likely to be addressed by our proposed design.

Generally, users are less likely to scrutinize a post for accuracy before sharing if their beliefs are aligned with the post (Pennycook et al. 2021). We show the presence of flags and curb the sharing

[17] https://www.pewresearch.org/journalism/2016/12/15/many-americans-believe-fake-news-is-sowing-confusion/

behavior of users, even when beliefs are aligned. More flags make users believe that sharing such posts will adversely impact their reputation, which, in turn, discourages them from sharing. The finding highlights that self-presentation triumphs over beliefs, making flags an effective feature for reducing misinformation.

Our study contributes to the misinformation literature by uncovering and assessing the efficacy of the flag feature in a social context to curb misinformation. This is a simple tool that platforms can implement to complement their existing ones. In addition to the misinformation literature, our study also contributes to the literature on social influences. Whether and how users respond to negative cues in a social setting is understudied in the existing literature. Prior studies have focused on negative reviews (Sen and Lerman 2007, Lim and Van Der Heide 2015), which are largely motivated by external factors such as product quality, experience, etc. Specifically, in the context of misinformation, we show that users objectively make use of the flag feature based on their information level. Moreover, they do not discount the negative cues from the crowd and make use of them to judge the accuracy of the post, despite their ideological alignment or beliefs.

The benefit of our proposed strategy is fourfold: (i) Existing strategies lack transparency in explaining moderation decisions and are criticized for violating free speech, as demonstrated by recent actions taken by YouTube.[18] Given that the crowd itself is moderating in our approach, such concerns will be mitigated, (ii) In the platform controlled strategy, a group of human moderators often encounter harmful content, which is detrimental to their well-being and mental health[19] (Arsht and Etcovitch 2018). When the crowd moderates the content, such concerns are likely to be addressed in the proposed crowd-driven strategy. (iii) Given the limited number of human resources employed by a social media platform, engaging the masses in moderation has the potential to mitigate delays, scalability issues, and improve the detection of post intent, (iv) Finally, we analyze the efficacy of the flag, and show that social media platforms can incorporate this simple feature to help reduce the spread of misinformation.

6. Conclusion

Mitigating the large-scale propagation of misinformation is one of the biggest challenges in the current digital age. Our study investigates the efficiency of a community-based strategy to moderate content in dealing with misinformation. Most studies in the current literature have focused on platform-initiated strategies such as nudges, third-party fact-check warnings, etc., in exploring mechanisms to curb the spread of fake news; however, little is known about the efficacy of a decentralized crowdsourced approach in a setting with the opportunity to provide social cues

[18] https://www.npr.org/2023/06/02/1179864026/youtube-will-no-longer-take-down-false-claims-about-u-s-elections

[19] https://blog.google/documents/83/information_quality_content_moderation_white_paper.pdf/

without anonymity. Our research uses a social media application we created to provide users with a realistic setting for the experiment. A series of posts with a randomly assigned number of flags and shares were shown to the users in a random order and the users could choose to share, flag, or do nothing.

Overall, we demonstrate that users, when given the charge of moderation in the presence of social cues, are efficient in sorting true information from false, thus sending reliable signals to other users. We provide key insights into how flags of the crowd and friends shape user perceptions about a post's accuracy, popularity, and its characteristics of being interesting to generate discussions. We further show that users believe that a flagged post involves misinformation and refrain from sharing it for both impression management and altruistic motives. Our study informs the misinformation literature about the efficacy of user-based content moderation. We also add to the literature on social influence by showing how sharing is influenced by negative feedback. Finally, our study has important practical implications for social media platforms on how to leverage social cues to combat the spread of misinformation. One limitation of our study is that we are unable to assess how users' content creation behavior is affected by the introduction of flagging feature on the social media platform; future studies can build on our work to investigate the potential change of user behavior.

References

Adomavicius G, Curley SP, Gupta A, Sanyal P (2012) Effect of information feedback on bidder behavior in continuous combinatorial auctions. *Management Science* 58(4):811–830.

Allcott H, Gentzkow M (2017) Social media and fake news in the 2016 election. *Journal of economic perspectives* 31(2):211–236.

Amabile TM (1983) Brilliant but cruel: Perceptions of negative evaluators. *Journal of Experimental Social Psychology* 19(2):146–156.

Arsht A, Etcovitch D (2018) The human cost of online content moderation. *Harvard Law Review Online, Harvard University, Cambridge, MA, USA. Retrieved from https://jolt. law. harvard. edu/digest/the-human-cost-ofonline-content-moderation* .

Avram M, Micallef N, Patil S, Menczer F (2020) Exposure to social engagement metrics increases vulnerability to misinformation. *arXiv preprint arXiv:2005.04682* .

Bakshy E, Hofman JM, Mason WA, Watts DJ (2011) Everyone's an influencer: quantifying influence on twitter. *Proceedings of the fourth ACM international conference on Web search and data mining*, 65–74.

Bakshy E, Messing S, Adamic LA (2015) Exposure to ideologically diverse news and opinion on facebook. *Science* 348(6239):1130–1132.

Bakshy E, Rosenn I, Marlow C, Adamic L (2012) The role of social networks in information diffusion. *Proceedings of the 21st international conference on World Wide Web*, 519–528.

Bapna R, Gupta A, Rice S, Sundararajan A (2017) Trust and the strength of ties in online social networks. *MIS quarterly* 41(1):115–130.

Berger J (2014) Word of mouth and interpersonal communication: A review and directions for future research. *Journal of consumer psychology* 24(4):586–607.

Blascovich J, Loomis J, Beall AC, Swinth KR, Hoyt CL, Bailenson JN (2002) Immersive virtual environment technology as a methodological tool for social psychology. *Psychological inquiry* 13(2):103–124.

Burt RS (1984) Network items and the general social survey. *Social networks* 6(4):293–339.

Cabrera A, Collins WC, Salgado JF (2006) Determinants of individual engagement in knowledge sharing. *The International Journal of Human Resource Management* 17(2):245–264.

Calvillo DP, Ross BJ, Garcia RJ, Smelter TJ, Rutchick AM (2020) Political ideology predicts perceptions of the threat of covid-19 (and susceptibility to fake news about it). *Social Psychological and Personality Science* 11(8):1119–1128.

Cao J, Smith EB (2021) Why do high-status people have larger social networks? belief in status-quality coupling as a driver of network-broadening behavior and social network size. *Organization Science* 32(1):111–132.

Chaiken S (1987) The heuristic model of persuasion. *Social influence: the ontario symposium*, volume 5, 3–39.

Chang TS, Hsiao WH (2014) Time spent on social networking sites: Understanding user behavior and social capital. *Systems Research and Behavioral Science* 31(1):102–114.

Cohen GL (2003) Party over policy: The dominating impact of group influence on political beliefs. *Journal of personality and social psychology* 85(5):808.

Constant D, Sproull L, Kiesler S (1996) The kindness of strangers: The usefulness of electronic weak ties for technical advice. *Organization science* 7(2):119–135.

Coppock A (2019) Generalizing from survey experiments conducted on mechanical turk: A replication approach. *Political Science Research and Methods* 7(3):613–628.

Dias N, Pennycook G, Rand DG (2020) Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review* 1(1).

Dillard JP (2002) *The persuasion handbook: Developments in theory and practice* (Sage).

Epstein Z, Lin H, Pennycook G, Rand D (2022) How many others have shared this? experimentally investigating the effects of social cues on engagement, misinformation, and unpredictability on social media. *arXiv preprint arXiv:2207.07562* .

Fang C, Zhang J (2019) Users' continued participation behavior in social q&a communities: A motivation perspective. *Computers in Human Behavior* 92:87–109.

Foerderer J, Kude T, Mithas S, Heinzl A (2018) Does platform owner's entry crowd out innovation? evidence from google photos. *Information Systems Research* 29(2):444–460.

Gawronski B, Ng NL, Luke DM (2023) Truth sensitivity and partisan bias in responses to misinformation. *Journal of Experimental Psychology: General* 152(8):2205.

Gillin J (2018) Politifact's guide to fake news websites and what they peddle. politifact.

Guess AM, Lerner M, Lyons B, Montgomery JM, Nyhan B, Reifler J, Sircar N (2020) A digital media literacy intervention increases discernment between mainstream and false news in the united states and india. *Proceedings of the National Academy of Sciences* 117(27):15536–15545.

Hennig-Thurau T, Gwinner KP, Walsh G, Gremler DD (2004) Electronic word-of-mouth via consumer-opinion platforms: what motivates consumers to articulate themselves on the internet? *Journal of interactive marketing* 18(1):38–52.

Huang H (2015) International knowledge and domestic evaluations in a changing society: The case of china. *American Political Science Review* 109(3):613–634.

Jahanbakhsh F, Zhang AX, Berinsky AJ, Pennycook G, Rand DG, Karger DR (2021) Exploring lightweight interventions at posting time to reduce the sharing of misinformation on social media. *Proceedings of the ACM on Human-Computer Interaction* 5(CSCW1):1–42.

Jang JY, Han K, Lee D (2015) No reciprocity in" liking" photos: Analyzing like activities in instagram. *Proceedings of the 26th ACM conference on hypertext & social media*, 273–282.

Jiang T, Hou Y, Wang Q (2016) Does micro-blogging make us "shallow"? sharing information online interferes with information comprehension. *Computers in Human Behavior* 59:210–214.

Jiménez Durán R (2021) The economics of content moderation: Theory and experimental evidence from hate speech on twitter. *Available at SSRN 4044098* .

Juul JL, Ugander J (2021) Comparing information diffusion mechanisms by matching on cascade size. *Proceedings of the National Academy of Sciences* 118(46):e2100786118.

Kim A, Moravec PL, Dennis AR (2019) Combating fake news on social media with source ratings: The effects of user and expert reputation ratings. *Journal of Management Information Systems* 36(3):931–968.

Kim Em, Ihm J (2020) More than virality: Online sharing of controversial news with activated audience. *Journalism & Mass Communication Quarterly* 97(1):118–140.

Lee JJ, Kang KA, Wang MP, Zhao SZ, Wong JYH, O'Connor S, Yang SC, Shin S (2020) Associations between covid-19 misinformation exposure and belief with covid-19 knowledge and preventive behaviors: cross-sectional online study. *Journal of medical Internet research* 22(11):e22205.

Leskovec J, Mcauley J (2012) Learning to discover social circles in ego networks. *Advances in neural information processing systems* 25.

Lim Ys, Van Der Heide B (2015) Evaluating the wisdom of strangers: The perceived credibility of online consumer reviews on yelp. *Journal of Computer-Mediated Communication* 20(1):67–82.

Lin HF (2007) Effects of extrinsic and intrinsic motivation on employee knowledge sharing intentions. *Journal of information science* 33(2):135–149.

Luo M, Hancock JT, Markowitz DM (2022) Credibility perceptions and detection accuracy of fake news headlines on social media: Effects of truth-bias and endorsement cues. *Communication Research* 49(2):171–195.

Mannes AE (2009) Are we wise about the wisdom of crowds? the use of group judgments in belief revision. *Management Science* 55(8):1267–1279.

Martel C, Allen J, Pennycook G, Rand DG (2022) Crowds can effectively identify misinformation at scale. *Perspectives on Psychological Science* 17456916231190388.

Marwick AE, Boyd D (2011) I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13(1):114–133.

Matzler K, Renzl B, Müller J, Herting S, Mooradian TA (2008) Personality traits and knowledge sharing. *Journal of economic psychology* 29(3):301–313.

Moravec PL, Kim A, Dennis AR (2020) Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research* 31(3):987–1006.

Osmundsen M, Bor A, Vahlstrup PB, Bechmann A, Petersen MB (2021) Partisan polarization is the primary psychological motivation behind political fake news sharing on twitter. *American Political Science Review* 115(3):999–1015.

Park S, Shin W, Xie J (2021) The fateful first consumer review. *Marketing Science* 40(3):481–507.

Pennycook G, Bear A, Collins ET, Rand DG (2020a) The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* 66(11):4944–4957.

Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand DG (2019) Understanding and reducing the spread of misinformation online. *Unpublished manuscript: https://psyarxiv. com/3n9u8* .

Pennycook G, Epstein Z, Mosleh M, Arechar AA, Eckles D, Rand DG (2021) Shifting attention to accuracy can reduce misinformation online. *Nature* 592(7855):590–595.

Pennycook G, McPhetres J, Zhang Y, Lu JG, Rand DG (2020b) Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31(7):770–780.

Pennycook G, Rand DG (2021) The psychology of fake news. *Trends in cognitive sciences* .

Pereira A, Harris E, Van Bavel JJ (2023) Identity concerns drive belief: The impact of partisan identity on the belief and dissemination of true and false news. *Group Processes & Intergroup Relations* 26(1):24–47.

Pham DV, Nguyen GL, Nguyen TN, Pham CV, Nguyen AV (2020) Multi-topic misinformation blocking with budget constraint on online social networks. *IEEE Access* 8:78879–78889.

Pierri F, Piccardi C, Ceri S (2020) Topology comparison of twitter diffusion networks effectively reveals misleading information. *Scientific reports* 10(1):1372.

Ritson M, Elliott R (1999) The social uses of advertising: an ethnographic study of adolescent advertising audiences. *Journal of Consumer research* 26(3):260–277.

Samanta S, Dubey VK, Sarkar B (2021) Measure of influences in social networks. *Applied Soft Computing* 99:106858.

Sen S, Lerman D (2007) Why are you telling me this? an examination into negative consumer reviews on the web. *Journal of interactive marketing* 21(4):76–94.

Stadtfeld C, Vörös A, Elmer T, Boda Z, Raabe IJ (2019) Integration in emerging social networks explains academic failure and success. *Proceedings of the National Academy of Sciences* 116(3):792–797.

Stagnaro MN, Pennycook G, Rand DG (2018) Performance on the cognitive reflection test is stable across time. *Judgment and Decision making* 13(3):260–267.

Sundaram DS, Mitra K, Webster C (1998) Word-of-mouth communications: A motivational analysis. *ACR North American Advances* .

Wang C, Zhang X, Hann IH (2018) Socially nudged: A quasi-experimental study of friends' social influence in online product ratings. *Information Systems Research* 29(3):641–655.

Wang SA, Pang MS, Pavlou PA (2021) Seeing is believing? how including a video in fake news influences users' reporting of the fake news to social media platforms. *MIS Quarterly (Forthcoming), Fox School of Business Research Paper* .

Xu X, Zhu C, Wang Q, Zhu X, Zhou Y (2020) Identifying vital nodes in complex networks by adjacency information entropy. *Scientific reports* 10(1):2691.

Yaqub W, Kakhidze O, Brockman ML, Memon N, Patil S (2020) Effects of credibility indicators on social media news sharing intent. *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–14.

Yin D, de Vreede T, Steele LM, de Vreede GJ (2023) Decide now or later: making sense of incoherence across online reviews. *Information Systems Research* 34(3):1211–1227.

Zhang L, Luo M, Boncella RJ (2020) Product information diffusion in a social network. *Electronic Commerce Research* 20:3–19.