Understanding Effects of Algorithmic vs. Community Label on Perceived Accuracy of Hyper-partisan Misinformation

CHENYAN JIA, Moody College of Communication, The University of Texas at Austin, USA
ALEXANDER BOLTZ*, Department of Government, The University of Texas at Austin, USA
ANGIE ZHANG*, School of Information, The University of Texas at Austin, USA
ANQING CHEN, Department of Electrical and Computer Engineering, The University of Texas at Austin, USA
MIN KYUNG LEE, School of Information, The University of Texas at Austin, USA

Hyper-partisan misinformation has become a major public concern. In order to examine what type of misinformation label can mitigate hyper-partisan misinformation sharing on social media, we conducted a 4 (label type: algorithm, community, third-party fact-checker, and no label) X 2 (post ideology: liberal vs. conservative) between-subjects online experiment (N = 1,677) in the context of COVID-19 health information. The results suggest that for liberal users, all labels reduced the perceived accuracy and believability of fake posts regardless of the posts' ideology. In contrast, for conservative users, the efficacy of the labels depended on whether the posts were ideologically consistent: algorithmic labels were more effective in reducing the perceived accuracy and believability of fake conservative posts compared to community labels, whereas all labels were effective in reducing their belief in liberal posts. Our results shed light on the differing effects of various misinformation labels dependent on people's political ideology.

CCS Concepts: • **Human-centered computing → Empirical studies in HCI**.

Additional Key Words and Phrases: accuracy, misinformation label, hyper-partisan misinformation, social media

ACM Reference Format:
Chenyan Jia, Alexander Boltz, Angie Zhang, Anqing Chen, and Min Kyung Lee. 2022. Understanding Effects of Algorithmic vs. Community Label on Perceived Accuracy of Hyper-partisan Misinformation. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 371 (November 2022), 27 pages. https://doi.org/10.1145/3555096

1 INTRODUCTION

Hyper-partisan misinformation has become a major public concern that can exacerbate partisan disagreement over even basic facts [34, 45]. Partisans tend to believe news that aligns with their beliefs regardless of truthfulness due to confirmation bias [32]. Such confirmation bias has proven to be one major factor affecting social media users' belief in news articles [32]. Partisans may share

*Both authors contributed equally to this research.

Authors' addresses: Chenyan Jia, chenyanjia@utexas.edu, Moody College of Communication, The University of Texas at Austin, 300 W. Dean Keeton, A0900, Austin, Texas, USA, 78712–1069; Alexander Boltz, alexboltz@utexas.edu, Department of Government, The University of Texas at Austin, USA; Angie Zhang, angie.zhang@austin.utexas.edu, School of Information, The University of Texas at Austin, USA; Anqing Chen, benjamin.c0427@gmail.com, Department of Electrical and Computer Engineering, The University of Texas at Austin, USA; Min Kyung Lee, minkyung.lee@utexas.edu, School of Information, The University of Texas at Austin, USA.

2573-0142/2022/11-ART371 $15.00
https://doi.org/10.1145/3555096

ideologically consistent but false information on social media, partially because their attention focuses more on the political alignment of content instead of accuracy [38, 62]. Previous studies have identified partisanship as a stronger predictor of misinformation sharing than veracity among Twitter users [21].

The widespread sharing of hyper-partisan misinformation on social media is problematic and can be connected to a couple factors. First, news consumption on social media is often less mindful because of the entertainment-seeking goals of social media use [29, 31]. Frequent social media users often process information with hedonic mindsets [64] and thus are less likely to think critically than when they are in a utilitarian mindset [59]. Additionally, social media platforms further accelerate the proliferation of false and misleading information because of numerous sharing features of the platforms and the sheer number of users [16, 55]. Researchers have expressed their fears of the potential for social media platforms being leveraged by propagandists with ulterior motives such as confusing voters with information overload, preventing them from being able to distinguish truths from falsehoods [17, 36]. One recent study suggests that prior exposure to fake political posts, even for extremely implausible posts, can increase people's perceived accuracy of such misinformation [44].

Researchers and practitioners have recognized the urgent need for seeking solutions to combat the spread of misinformation. Many previous studies have shown that misinformation labels (e.g., stop signs, disclaimers, or warnings) can reduce people's believability of fake posts and decrease the propensity to share misinformation [38, 75]. Most past studies have examined the effect of labels attributed to third-party fact-checkers [38, 43] or human moderators hired by social media platforms. For instance, previous work examined the impact of third-party fact-checkers' labels on Twitter or Facebook posts containing false election claims during the 2020 Presidential election [76]. Although those labels provide possible solutions to mitigate misinformation sharing, this approach to labelling has a pressing limitation: in spite of its high credibility and reliability, third-party fact-checkers' labels heavily rely on human moderation and can not provide real-time intervention [75]. Thus by the time posts get manually labelled, the false information may have already spread.

One alternative approach is to leverage real-time interventions such as automated labelling (i.e., through misinformation detection algorithms) or community-based methods to assist users in distinguishing false information from factual content. While the development of misinformation detection algorithms continues to expand [25–27] and Twitter has launched a preliminary pilot (Birdwatch[1]) to test crowdsourced misinformation reporting, not much additional work has yet explored how to use these techniques to provide real-time feedback to users. Very few studies have examined how artificial intelligence (AI) and human-related labels differ in reducing people's believability in misinformation. One pioneering paper explores how the presence of an algorithmic misinformation detection warning and a fact-checker's warning will influence people's ability to detect misinformation [51]. Results show that those participants made more correct decisions identifying misinformation with the algorithmic warning (78.3%) than without any warning (70.1%), but their accuracy in identification of real news was similar between the two conditions [51]. Another work examined the impact of real-time labels such as misinformation detection algorithms and other users from the community [75]. Results showed that both the algorithmic label and community users' label succeeded in persuading people to avoid sharing false headlines, but both were less effective compared to third-party labels. Neither of the papers, however, measured the participants' political stance as a factor and thus overlooked whether there may have been a difference in misinformation labels' impact on posts dependent on people's political ideology. To

[1]blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html

address this gap, our paper intends to further investigate the different effects of real-time labels (i.e., labels attributed to algorithms and community) and third-party fact-checkers' labels on partisans' perceptions of hyper-partisan social media posts. Since 2020, the COVID-19 pandemic has posed an unprecedented risk to society, resulting in a polarized political landscape [46]. Prior work found initial evidence that people's misperceptions about COVID-19 are associated with people's political ideology [46]. Given the emerging nature of such a topic, our paper focuses on COVID-19 related social media posts.

In recent years, there has been a prevalent expectation that AI will have a unique advantage in online political contexts [73]. Public discourse reflects that AI is often perceived as more fair, objective, unbiased, and as having a less political agenda under the assumption of neutrality [20]. People tend to hold positive stereotypes about machine infallibility and neutrality [56]. Such heuristics may generate an advantage for AI-related products in political settings, particularly where these products may assist in mitigating hyper-partisan misinformation. Previous studies have examined AI-related decisions vs. human decisions in the context of news recommendation and online content moderation [70, 73]. Initial evidence suggests that people's perceptions of news bias may be attenuated when news is attributed to a machine cue [70] because people are inclined to believe that stories written or selected by a machine must be objective and free from political bias [28, 67].

Nevertheless, not many studies have examined algorithmic labels vs. human labels in the context of hyper-partisan misinformation sharing on social media. One pioneering work finds no difference between the effect of algorithmic and other users in the community's credibility indicators in reducing people's sharing intention of fake news headlines [75]. Adding to prior work, our study develops a social media-like website to further explore people's perceptual and behavioral responses to social media posts instead of using news headlines as stimuli, as was done in many past studies [38, 48, 75]. Understanding how users interact with social media posts is particularly valuable because misinformation embedded in users' posts can be more rapidly spread due to ease of access to the platform. Therefore, one focus of our study is to compare the differing effects of algorithmic- and community-based misinformation labels on partisan's perceptions of political social media posts. Using COVID-19 social media posts, we examined the following overarching questions: 1) Can misinformation labels reduce partisans' perceived accuracy and believability of fake posts? 2) Do different types of misinformation labels (algorithm, community, third-party fact-checker) exhibit different levels of influence on partisans' perceptions and sharing intentions of fake political posts? 3) How will partisans perceive different misinformation labels?

Consistent with prior work, our findings confirm that both algorithmic and third-party misinformation labels can reduce people's perceived accuracy and belief in COVID-19 related fake posts regardless of source ideology. In terms of community labels, we find that it reduced only liberal participants' beliefs in COVID-19 related fake posts regardless of source ideology but not for conservative participants. For conservative participants, community labels reduced the believability of fake liberal posts but were not effective for fake conservative posts. However, algorithmic and third-party fact-checker indicators reduced conservatives' belief in fake conservative posts. One interesting finding of our study is that the algorithmic labels perform as well as third-party fact-checker labels in reducing partisans' belief in COVID-19 related fake posts.

2 RELATED WORK

2.1 Hyper-Partisan Misinformation

Hyper-partisan misinformation refers to misleading information with strong partisan bias [16]. Hyper-partisan misinformation often portrays itself as functionally indistinguishable from 'real

news' but is more often a mix of genres, combining news, entertainment, and politically-charged opinions [39]. Misinformation entered the national spotlight during the 2016 US Federal Elections, and in its aftermath, misinformation has gained even more attention during the COVID-19 pandemic [45]. An internal study conducted by Twitter discovered that political content with a conservative bias is routinely favored by Twitter's algorithm [11]. The pandemic has become an important, yet divisive, political issue, with 60% of surveyed voters indicating it was 'the most important' or 'an important' factor in the 2020 Presidential election, and only 38% of the 'most important' group voting for Trump [13]. Studies also find that belief in COVID-19 misinformation is highly correlated with distrust in science [1].

Past misinformation studies have explored more traditional political issues such as immigration [23, 61], former President Trump [12], mandatory vaccinations [19], gun control [18], and abortion [33]. As the pandemic has become a global issue, it is essential to examine COVID-19 related hyper-partisan misinformation as COVID-19 also represents a unique political phenomenon. Furthermore, when analyzing health and vaccination-related issues, the tone and approach of interventions appear to be even more crucial and sensitive [19].

A large body of literature has explored why people share hyper-partisan misinformation and how to reduce such information sharing. One widely acknowledged explanation is that partisans will preferentially trust news that is consistent with their existing political ideology regardless of its truthfulness [16]. Past work suggests that hyper-partisan misinformation is the context where such politically motivated reasoning is more likely to occur [16, 30]. In other words, partisans value political alignment more than veracity when sharing misinformation [21]. Thus, without intervention, partisans tend to believe the news that aligns with their beliefs regardless of its truth due to confirmation bias [32]. Another possible reason for this tendency is that moral and highly emotional framing of news stories also drives sharing of hyper-partisan misinformation over social media, engaging with readers' morality through selective framing [74].

Some prior studies suggest that engaging in active reasoning and internal deliberation [5] or shifting attention to accuracy [45] can decrease people's belief in hyper-partisan false news and inadvertent misinformation sharing. Other studies find that a reduction of emotion intensity in news stories can result in less hyper-partisan misinformation sharing behavior [74]. Even so, not much work has explored how real-time interventions such as algorithmic and community-based labels can reduce partisans' confirmation bias.

2.2 Labelling Fake News

While some earlier studies find fake news labels to be generally ineffective [18, 37, 66], a growing body of literature suggests the opposite [5, 38, 75]. Interventions such as fake news labels - including warnings, credibility indicators, and disclaimers - have been proven to be effective in reducing people's believability and sharing intention of fake news [38, 43, 75].

Past work suggests that the presence of a misinformation label on news headlines that align with users' beliefs can trigger cognitive activity such as increased attention and increased time spent considering the headline [37], and thus may reduce people's beliefs in misinformation. One recent work found that not only can a misinformation label with a detailed warning message ('declared fake by 3rd party fact-checkers') significantly reduce people's believability of misinformation, but a simple stop sign can also reduce people's believability of misinformation by triggering people's gut reaction and natural intuition [14, 38].

Studying the proliferation of hyper-partisan news over social media is also an expanding field, with sharing intentions originating from politically polarizing new sources (such as Infowars or Breitbart) gaining increasing research focus. The sharing of news headlines from such low-credibility hyper-partisan news sites drives significant sharing activity, however, interventions

such as a simple label can increase user discernment and reduce sharing behavior [45]. Strong interventions (e.g., an intervention labelling potential misinformation as 'rated false') have been shown to cause stronger effectiveness compared to weaker interventions (e.g., labelling potential misinformation as 'disputed') in counter-acting belief in misinformation [12].Furthermore, labelling was found to be most effective when appealing to both the mind's system 1 (automatic) as well as system 2 (deliberate) cognition, demanding the activation of the mind's intuition as well as its rational thinking core [38].

Most prior work examined the effect of labels provided by third-party fact-checkers [38]. While the third-party fact-checking indicator is reliable, it often heavily relies on human moderation and can only verify news at a slow pace and on a small scale [75]. Algorithmic fact-checking is becoming an increasingly viable alternative with benefits such as a massively increased breadth, a reduction in potential bias, and increased applicability to alternative cultures and contexts.

Other studies also indicate that users are highly cognisant of the content they consume on social media and are sensitive to potential misinformation [4, 23, 32, 33, 51, 69]. In these cases where labelling can affect users' perceptions of fake news though, the matter of labelling misinformation is further complicated by a potential "implied truth" on *unlabelled* news items that users may assume after seeing that some posts have accompanying labels indicating misinformation [43]. This implied truth effect could backfire if the correction of false beliefs results in increased misconceptions [7]. One possible explanation for this phenomenon is the higher level of awareness and thoughtfulness that exists when performing in an active study that demands the users' attention, compared to the passive nature of consuming social media content.

A prior study has found that political beliefs, interest in politics, and education all heavily affect belief in news generally, however - surprisingly - believability did not vary significantly between true news and labelled false news [61]. Furthermore, the researchers found that young, educated, and left-oriented users distrusted any news (true or false) on social media to a greater degree. In another study, researchers found that when labels indicating a news source's ideology matched the user's ideology, this significantly increased the user's trust in news from that source [18]. However, this study also discovered that, generally, adding credibility labels to articles - such as indicating that the news article was disputed - was not effective in combating misinformation and its spread. Similarly, one study found that news confirming users' political beliefs produced strong confirmation bias effects that were too large to overcome with labelling [38].

Many social media platforms have been active in educating their user bases on the potential perils of misinformation, to desirous results [60]. However, as Zannettou [76] found, the levels and types of engagement with misinformation labeled posts on social media platforms has variations depending on partisan leaning. Notably, Zannettou [76] studied a collection of tweets and misinformation interventions primarily about the 2020 U.S. Presidential election, finding that Republican users were responsible for 72% of re-shared tweets with misinformation interventions (compared to 11% by Democrats). This study, however, examined real tweets which were limited to one type of label attributed to third-party fact-checkers. This indicates a need to study how alternative intervention labels on hyper-partisan misinformation may impact Republican users vs. Democrat users. Because of the aforementioned benefits to alternative types of misinformation detection, the need to investigate the effectiveness of those types is clear. Adapting previous literature written investigating the accuracy and believability of misinformation in the context of multiple types of labelling systems is valuable because it could provide insight into the future of misinformation detection.

2.3 Community vs. Algorithmic Misinformation Labels

Given the limitations in scalability of third-party fact-checkers, researchers and companies continue to explore the potential and effectiveness of automated techniques for misinformation detection. One method currently being studied, notably being tested in the real world as well, is crowd-sourcing techniques for detecting misinformation in real-time: Twitter is currently piloting a community-driven, real-time misinformation detection effort (Birdwatch[2]). A separate study has also investigated the accuracy of crowd-sourced misinformation detection compared to third-party fact-checkers, finding that crowd-sourced detection can be as reliable and accurate [2].

Another alternative to third-party fact-checkers in misinformation detection is misinformation detection algorithms. In recent years, researchers have had success in creating misinformation detection algorithms, using various machine learning methods to identify features and characteristics of misinformation in order to perform things like identifying categories of misinformation [26] and distinguishing between fake and real news article titles [25] to promising results. As misinformation detection algorithms advance in their detection accuracy, it is valuable to understand how such algorithms would be received by users of a social media platform.

Seo et al. [51] investigated the effectiveness and trustworthiness of false news warnings based on its attribution source, comparing warnings ascribed to third-party fact-checkers websites vs. a machine learning (ML) algorithm. They conducted two experiments to assess the effect of the warnings on participants' abilities to detect fake and true news and participants' trust in warnings [51]. In their first experiment, while all warning types led to higher correct detection of misinformation, only the fact-checking warning led to participants being able to detect more true news as well. In the second experiment, they enhanced the ML warning with more information about how the algorithm worked and removed source labels from the fact-checking warning. This enhanced ML warning resulted in the highest correct detection of both fake and true news. However, despite the efficacy of all warnings in increasing recognition of misinformation, participants still held low trust in warnings [51]. This low trust may indicate that users require additional or different ways of understanding how an automated detection method works before they feel comfortable trusting it. Additionally, this study does not examine the differences between these labels and a misinformation label attributed to other public or community users.

Most other past work on misinformation labels has primarily explored how users react to warnings attributed to third-party fact-checkers but not algorithmic labels or other public or community-based labels [12, 33, 43]. However, one recent study did include testing user perceptions of various credibility indicators including a "Public" and AI credibility indicator: Yaqub et al. [75] looked into how label interventions could affect people's tendency to share fake stories. The authors manipulated the source that disputed the stories, testing 1) fact-checkers, 2) news media, 3) public ("a majority of Americans"), and 4) AI against people's willingness to share headlines. The most effective indicator in reducing the sharing of fake headlines was fact-checkers, while AI was the least effective. There was also no significant difference in effect between the label attributed to the public compared to the AI label [75]. It is possible that the public label was not as effective as it was described as being "a majority of Americans", rather than a more specific group of people leading participants to view it as too general. The AI label may have also been viewed skeptically as the use of the term "Artificial Intelligence" in describing the indicator may not have been well-received by participants who did not know the term or were familiar with it but uncertain about how the algorithm worked. In their experiment, however, the participants had a high rate of recall failure which may have affected their findings in the efficacy of the labels that they selected.

[2]blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html

Our study intends to further examine this comparison between third-party fact-checker labels and automated, real-time labels (i.e. algorithmic labels and community labels), particularly in the context of social media posts. Previous studies, including [75] have only examined the effect of these labels when shown with news article headlines. We also contribute to literature investigating how the effectiveness of these labels differ dependent on people's party affiliation and the ideology of the posts they view.

3 RESEARCH QUESTIONS AND HYPOTHESES

Our study intended to examine the following overarching research questions:

- **RQ1** Can misinformation labels reduce partisans' perceived accuracy and believability of fake posts?
- **RQ2** Do different types of misinformation labels (algorithm, community, third-party fact-checker) exhibit different levels of influence on partisans' perceptions and sharing intention of fake political posts?
- **RQ3** How will partisans perceive different misinformation labels?

Building on prior work [38, 75], our study hypothesized that all three labels would be effective in reducing people's perceived accuracy and believability of fake posts regardless of the source ideology:

- **H1**: Algorithmic labels, community labels, and third-party fact-checkers' labels will be effective in reducing people's perceived accuracy and believability of fake posts that both align (**H1a**) and do not align (**H1b**) with their beliefs.

Previous studies have examined whether machine vs. human sources of online information affect the perceived bias and credibility of news [57, 67, 70]. Those studies find initial evidence that people's perceptions of news bias may be attenuated when news is attributed to a machine cue. Some scholars have named such a phenomenon as the "machine heuristic", which refers to the mental shortcut where people tend to consider machines as being more mechanical, objective, and ideologically unbiased than humans [56]. Users often attribute bias to other human users rather than algorithms because they believe the algorithm is less likely to have a political agenda [41].

One recent study found that when information challenges people's views, they tend to reject such information and perceive it as less credible and more biased, regardless of whether its source is AI or a human [73]. In other words, in regards to cross-cutting messages that do not align with partisans' beliefs, AI does not perform better than humans [73]. This can possibly be explained by the fact that news that challenges people's opinion usually receives little cognitive activity and cues such that any label, machine or human, will be more likely to be ignored [37]. Thus, we predicted the following hypothesis:

- **H2a**: For posts that align with their beliefs, the algorithmic misinformation labels will be more effective in reducing people's perceived accuracy of fake posts than community misinformation labels.
- **H2b**: For posts that do not align with their beliefs, there will be no difference in effectiveness between algorithmic misinformation labels and community misinformation labels in reducing people's believability of fake posts.

Based on previous research that suggests third-party fact-checkers' labels were the most effective among misinformation labels [75], we further predicted that:

- **H3**: For posts that align with their beliefs, the third-party fact-checkers' labels will be more effective in reducing people's perceived accuracy and believability of fake posts than algorithmic (**H3a**) and community (**H3b**) labels.

- **H4**: For posts that do not align with their beliefs, the third-party fact-checkers' labels will be more effective in reducing people's perceived accuracy and believability of fake posts than algorithmic (**H4a**) and community (**H4b**) labels.

Recent studies suggest that there was a dissociation between accuracy judgment and people's sharing intention [45]. People may still share news that they do not necessarily believe in [45]. Therefore, we also intended to examine the impact of labels on people's sharing, liking, and commenting intentions and people's perceptions of labels as exploratory analysis:

- How will different labels affect people's sharing, liking, and commenting intentions?
- Will people have different perceptions of algorithmic, community, and third-party fact-checkers' labels?

4 METHOD

4.1 Participants

We recruited 2257 participants using MTurk Toolkit on CloudResearch, an online participant pool that aggregates multiple market research platforms [35]. Participants were all from the United States. Participants were required to have a HIT approval rate greater than 95% and be over 18 years old. After ruling out people who had moderate political views (n = 353), were under the age of 18 (n = 3), failed the recall (n = 151), failed the embedded attention check question (n = 67), and spent less than four minutes (n = 6), 1677 participants remained in the data analysis.

		Source Ideology		
		Conservative	**Liberal**	**Total**
Conservative Participant	**Algorithm**	112	96	208
	Community	92	111	203
	Third-party	104	97	201
	No Label	90	107	197
Liberal Participant	**Algorithm**	106	100	206
	Community	99	121	220
	Third-party	115	115	230
	No Label	110	102	212
Total		828	849	1677

Table 1. Number of Participants in Each Condition

Label	**Description**
Algorithm	**Label**: You may want to know this post's accuracy is disputed by a misinformation detection algorithm.
	Hyper-Text : We worked with third-party fact-checkers to develop an algorithm. Our algorithm helps us detect misinformation quickly and accurately.
Community	**Label**: You may want to know this post's accuracy is disputed by other users checking misinformation.
	Hyper-Text : We worked with our users to create a community- based system. Our designated users help us detect misinformation quickly and accurately.
Third-party	**Label**: You may want to know this post's accuracy is disputed by third-party fact-checkers.
	Hyper-Text : We worked with third-party fact-checkers. Our goal is to detect misinformation accurately.

Table 2. Label and Hyper-Text Description

The mean age of the participants was 40.24 years old (*SD*=12.22, *Median*=38). Among 1677 participants: 744 (44.4%) were male, 915 (54.6%) were female, and 18 people chose other categories. When asked to self-report their political leaning, 868 of the participants self-reported as liberal-leaning (52%), and 809 (37%) as conservative-leaning [3]. Participants were compensated $1.50 US dollars for completing the experiment (Median completion time = 12.5 minutes excluding participants that kept their browsers open over 30 minutes).

4.2 Experiment Design

We conducted a 4 (labels: algorithm, community, third-party fact-checkers, no label) X 2 (source ideology: liberal vs. conservative) between-subjects design online experiment (N = 1677). Each participant was randomly assigned to one of the 8 conditions and read 12 posts. The 6 true and 6 false posts displayed had their veracity verified by major fact-checking organizations and were used as stimuli. The number of participants in each condition is shown in Table 1.

4.2.1 Manipulation Check. Several statistical tests were conducted to check whether randomization was effective and successful. One-way ANOVA showed there were similar sample distributions in terms of age, race, gender, education, Twitter usage, COVID-19 knowledge, political leaning, and party affiliation across 8 conditions.

Two manipulation checks were embedded in the experiment. Since the understanding of the label plays a primary role in our experiment, the first manipulation check used the recall question as a filter to filter out people who could not successfully recall the label (n = 151). Participants who correctly answered the recall question "could you recall the labels under the posts you just read?" remained in the data analysis. The second manipulation check tested whether the manipulation of source ideology was effective. Participants needed to answer the question "what, if any, is the political bias of this post?" on a 7-point scale with 1 representing 'extremely liberal' and 7 representing 'extremely conservative' (adapted from [40]) to rate the perceived bias of each post. Repeated measures ANOVA showed that our manipulation of source ideology was successful. Repeated measures ANOVA was used because participants were repeatedly measured on the same dependent variables for 12 posts. Participants who were assigned to the conservative source ideology conditions rated the perceived bias of posts (M= 5.07, SE=.02) significantly higher than those who were assigned to the liberal source ideology conditions (M= 3.45, SE=.02), $p < .001$.

4.2.2 Label Design. All the fake posts were correctly labelled as misinformation based on the ground truth. Only the description of the entity that labeled the posts varied across the conditions (algorithmic, community, or third-party fact-checkers' labels). This design decision was made to understand the effect of the label type without confounding factors that can come from differences in the posts that different label types may tag as misinformation.

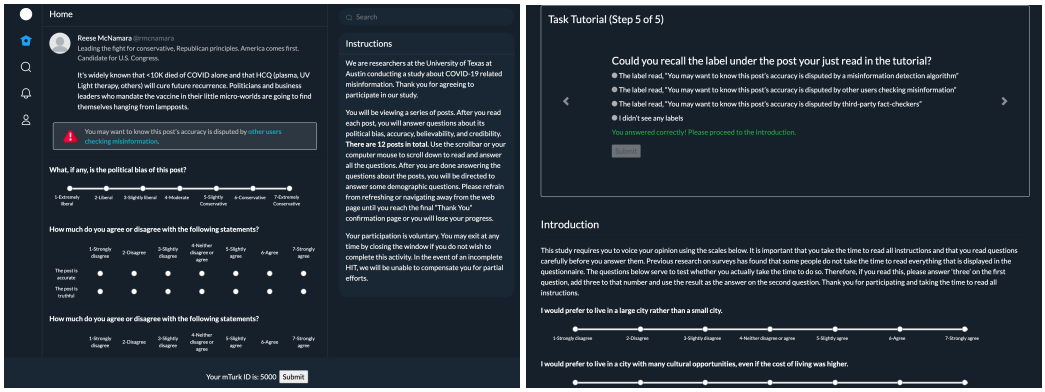
The label designs were inspired by real-life social media fact check labels. The official misinformation policies of Twitter[4], Facebook[5], and Instagram[6] were instrumental, as the labels contained many phrases lifted directly from the policy statements. In addition to the direct labelling of the post as potentially misleading, participants also had access to hover-able hypertext that expands on how the treatments are intended to function in a real-world context. Detailed label and hyper-text descriptions are shown in Table 2.

[3]More detailed demographic data - including splits based on participant political ideology - can be found in the Supplemental Materials

[4]blog.twitter.com/en_us/topics/product/2021/introducing-birdwatch-a-community-based-approach-to-misinformation.html

[5]www.facebook.com/business/help/2593586717571940

[6]about.instagram.com/blog/announcements/combatting-misinformation-on-instagram



(a) Website Interface (b) Tutorial and Attention Check Interface

Fig. 1. Website and Tutorial Interface

To create our label wording, we carefully reviewed labels of past studies and real labels used by social media companies for misinformation. Past studies such as [33, 38, 44, 51] characterized misinformation as being "disputed" by third-party fact-checkers. Social media companies such as Twitter and Facebook have used stronger language in declaring misinformation as verified by independent fact-checkers or experts such as adding an additional step and message before people can read a post labeled as misinformation (Twitter[7]) and blurring content with the message "False Information"and labeling content with the message "The primary claims in the information are factually inaccurate" (Facebook[8]). Because of the imperfect nature and potential biases of misinformation detection by algorithms or community, we chose to use subtle language to nudge the reader to consider the presence of misinformation, telling them that they "*may want to know* this post's accuracy is *disputed* by (source)" as opposed to language that declares a post to be false[9].

4.2.3 Social Media Website Development. In order to simulate the real social media consumption environment, participants were asked to participate in the experiment on a website developed by our researchers. We created a website that emulates the Twitter environment. The website was developed in Python and hosted on Heroku, a cloud application platform; the data was recorded in a Heroku Postgres database. The interface of the website is shown in Figure 1.

Adding to previous experimental work on Twitter's misinformation labels [76], we chose to simulate Twitter as our social media platform because Twitter has taken multiple approaches to identify misinformation including third-party and community-driven fact-checking[4]. Additionally, most past studies have studied misinformation within the context of Facebook news headlines, while our work adds an additional layer of understanding by investigating misinformation labels in the context of another news format, social media posts on Twitter. Finally, we believe that simulating Twitter allows our study to have practical implications for companies including Twitter which has previously employed third-party fact-checkers for misinformation labels and has recently been exploring platform community-based fact-checking.

[7]https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information
[8]https://techcrunch.com/2020/03/03/trump-coronavirus-hoax-fact-check/
[9]Although Birdwatch was officially announced by Twitter in January of 2021, many of the specific details regarding the UI were not released outside the private beta until June/July, and remains constantly updating. Because this study began collecting data in March 2021, the specific UI and wording of Birdwatch was unknown to the researchers until after the study was designed.

4.3 Procedures

Once the participants reached the landing page of our experiment website, they were randomly assigned to one of the 8 conditions. They began with a tutorial walking them through an example of the task ahead. The tutorial specifically included a slide that explained some posts included a misinformation detection label (in non-control conditions) to ensure the participants would be aware of the label and its attribution (by an algorithm, third-party fact-checkers, or other platform users) during the experiment. To ensure the participant actively read the tutorial, we concluded the tutorial with a question asking them to recall what source the label was attributed to. Participants could not proceed to the main experiment until they answered this attention check question correctly, and participants were allowed multiple attempts. A satisfactory manipulation recall rate (91.00%) was achieved, after concluding the tutorial.

After passing the tutorial and two attention check questions, participants proceeded to the main experiment where they were exposed to 12 social media posts. During the main experiment, participants were asked to evaluate a series of questions related to post believability and accuracy, as well as intention to share, comment, or like. Following the main experiment, participants completed a post-test questionnaire, where they were asked to evaluate the label effectiveness, objectiveness, whether the label was objective or politically unbiased, and if they belonged to one of the three label conditions. They were also asked a series of questions regarding their demographics, COVID-19 concern, and beliefs on machine learning and social media.

4.4 Stimuli

Stimuli were collected directly from actual posts on Twitter.com. We collected 6 true and 6 false COVID-19 related social media posts. Both true and fake COVID-19 related posts were verified to be true or fake using third-party fact-checkers such as Snopes.com and PolitiFact, or large reputable health organizations such as the CDC or WHO. Each post was edited to have a similar length and readability score tested by the Flesch-Kincaid Grade Level (M = 11.10, SD = 1.64). The topics of these posts vary from vaccines, treatments, mortality rate, and masks.

Previous studies have presented different numbers of stimuli for participants to review [12, 18, 32, 33, 37, 38, 45, 47, 51, 61]. For instance, Kim et al. [32] use 8 stimuli in their study; Moravec et al. [37] use 10 stimuli; Moravec et al. [38] and Kim and Dennis [31] use 12 news headlines; and Seo et al. [51] use 24 stimuli. The number of 12 stimuli was consistent with many prior studies [31, 38]. We did not include more stimuli because many prior studies focus exclusively on presenting news headlines, whereas the longer format of social media posts in our study requires more active reading by participants, lengthening the experiment[10].

Most political-related Twitter experimental studies use elites/politicians or media organizations as the account types [63]. Substantial evidence suggests that elite messages can transfer to public consciousness and conversation [15]. Adapted from previous studies [63], our study uses political candidate bios to be our source to indicate the posts' ideology. We chose politically neutral posts and then manipulated the source ideology. For each post, we included a bio indicating the party affiliation and identity of the candidate (e.g., 'Morgan Lang: Liberal. Let's leave the Earth a better place than we found it. Candidate for U.S. Congress.'; 'Blake Walls: Conservative. No taxation without Representation. No big government! Candidate for U.S. Congress'). In order to reduce the impact of news source-specific effect, we followed previous research [32] and randomly assigned gender-neutral names to political candidates. All the names and bios were randomly assigned to

[10]Specifically, the average number of characters of 12 social media posts in our stimuli is 193.42. The average number of characters of 12 news headlines used by Moravec et al. [38] and Kim and Dennis [31] is 76.08.

posts for each participant. Consistent with prior work [32, 61], we used default profile pictures to avoid potential gender or race bias.

Profile biographies were standardized to a similar Flesch-Kincaid Grade Level Score (Conservative: M = 5.48, SD = 2.19; Liberal: M = 5.31, SD = 1.70). The bios themselves consisted first of a direct description of the user's intended political affiliation (liberal/conservative), followed by a statement or two of coded language related to their affiliation to increase the platform immersion. For example, a liberal user may indicate that 'Transgender Lives Matter' in their bio. After that, the biography indicated that the user was a candidate for US Congress; this was chosen to simulate a national-level politician with a realistic large social media reach, in contrast to a local-level politician without such reach.

4.5 Pre-Test of the Source Ideology Manipulation

We manipulated the profile bio of neutral social media posts to indicate the post ideology. In order to make sure the manipulation of the source ideology was successful, we conducted a pre-test (N=82) to test whether the direction of ideology was what we expected. Participants [11] were either randomly assigned to a group where all the neutral posts were assigned to liberal bios or a group where all the posts were assigned to conservative bios. We asked participants to answer the question "what, if any, is the political bias of this post?" on a 7-point scale with 1 representing 'extremely liberal' and 7 representing 'extremely conservative' (adapted from [40]). Repeated measures ANOVA showed that our pre-test was successful. Posts assigned to conservative bios (M= 5.16, SE=.13) were rated significantly higher than those assigned to liberal bio (M= 3.50, SE=.15), F (1, 80) = 50.65, p < .001.

4.6 Measures

4.6.1 Post-level measures. After each post, we asked the following questions to understand participants' perception of the post.

Political Bias. Political bias was measured by asking participants, "What, if any, is the political bias of this post?" on a 7-point scale ranging from 1 indicating 'extremely liberal' to 7 indicating 'extremely conservative'.

Perceived Accuracy. Perceived accuracy was measured as an index consisting of two items by asking how much do participants agree or disagree with the following statements - "the post is accurate" and "the post is truthful" (adapted from [45]) on a 7-point scales from 1 (Strongly disagree) to 7 (Strongly agree). The two items were highly correlated and can be averaged to form a reliable index (Cronbach's α = .99).

Perceived Believability. Perceived believability was measured as an index consisting of two items by asking how much participants agree or disagree with the following statements - "the post is believable" and "the post is credible" on a 7-point scales from 1 (Strongly disagree) to 7 (Strongly agree), adapted from [38]. The two items were highly correlated and can be averaged to form a reliable index (Cronbach's α = .96).

Social Media Behaviors. Participants were asked to evaluate the likelihood of them performing three typical social media actions: liking the post, sharing the post, and commenting on the post (adapted from [38]). Participants evaluated the likelihood on a 7-point scale ranging from 1 representing 'extremely unlikely' to 7 representing 'extremely likely'.

4.6.2 Post-survey measures. In the post-survey, we asked the following questions to understand the participants' overall perceptions of the label. We also asked a recall-based manipulation check question.

[11] Detailed demographic data regarding the pre-test participants can be found in the Supplemental Materials

Label Perceptions. In the post-test questionnaire, participants were asked to agree or disagree with three statements: 'the label is mechanical', 'the label is objective', and 'the label is politically unbiased'. Participants rated each of the statements on a 7-point scale ranging from 1 indicating 'strongly disagree' to 7 indicating 'strongly agree'.

Label Effectiveness. In the post-test questionnaire, participants were asked, "How effective is the label at indicating potential misinformation?" and evaluated the label on a 7-point scale with 1 designating 'extremely ineffective' and 7 designating 'extremely effective' (adapted from [38]).

Recall. Participants were asked to recall the specific label condition they were assigned and were given five answer choices: one for each of the three label conditions, one for no labels, and one for if the participant did not remember.

Additional variable details and descriptions are available in the Supplemental Materials.

5 RESULTS

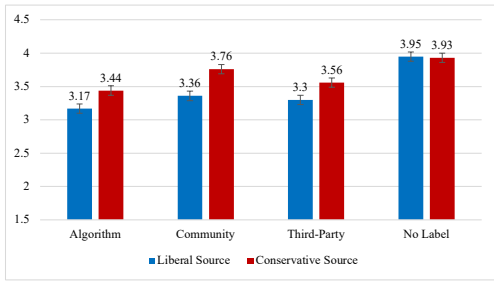
We used linear mixed models to analyze data controlling for participant ID and post ID as random effects. Results showed that there was a significant 3-way interaction effect among the participant's political leaning, source ideology, and label, $F(7,10046)= 6.26$[12], $p< .001$ on the perceived accuracy of fake posts. There was a significant main effect of the labels, $F(3,10046)=55.77$, $p< .001$, and a significant main effect of the participant's political leaning, $F(1,10046)=1328.22$, $p< .001$, on the perceived accuracy of fake posts.

H1 predicted that algorithmic labels, community labels, and third-party fact-checkers' labels would be effective in reducing people's perceived accuracy and believability of fake posts that both align (**H1a**) and do not align (**H1b**) with their beliefs. As shown in Figure 2 and Table 3, multiple pairwise comparisons using the Bonferroni post-hoc test revealed that compared with the control (no label) condition, for both conservative- and liberal participants, algorithmic labels significantly reduced the perceived accuracy of fake posts that aligned with their political beliefs, $p< .001$. For liberal participants, community labels significantly reduced the perceived accuracy of fake posts that aligned with their beliefs, $p< .001$. For conservative participants, however, there was no significant difference between the community label condition ($M=3.76$, $SD= .08$) and the control ($M=3.93$, $SD= .08$) condition, $p= .73$ in terms of the perceived accuracy. Third-party fact-checkers' labels significantly reduced the perceived accuracy of ideologically agreeable fake posts for both conservative- and liberal partisans, $p< .001$. Thus, (**H1a**) was partially supported. For fake posts that do not align with people's beliefs, multiple pairwise comparisons showed that all three labels significantly reduced the perceived accuracy of fake posts compared with the no label condition, $p< .001$. as shown in Figure 2 and Table 3. Thus, (**H1b**) was supported.

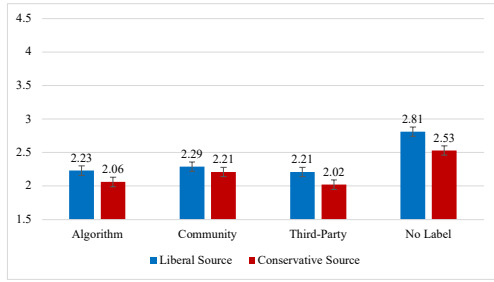
For perceived believability, the pattern stayed the same, as shown in Figure 2. There was a significant main effect of the labels, $F(3,10046)=44.14$, $p< .001$, and a significant main effect of participant's political leaning, $F(1,10046)=1223.66$, $p< .001$, on the perceived accuracy of fake posts. There was a significant 3-way interaction effect among the participant's political leaning, source ideology, and labels on the perceived believability, $F(7,10046)= 6.42$, $p< .001$. Pairwise comparisons showed that for both liberal and conservative participants, all three labels significantly reduced the perceived believability of fake posts that do not align with their political beliefs, $p< .001$. For fake posts that align with participants' political beliefs, both algorithmic and third-party fact-checkers' labels significantly reduced the perceived believability, $p< .001$; yet the significant effect only existed for liberal participants for the community label condition, $p< .001$.

H2a predicted that for posts that align with their beliefs, the algorithmic labels would be more effective in reducing people's perceived accuracy of fake posts than community labels, which was

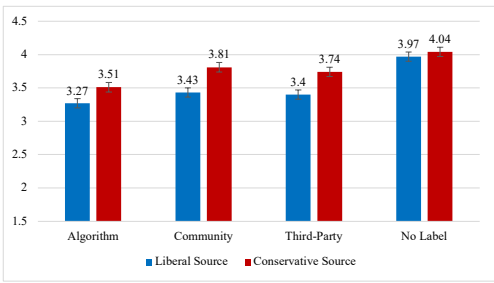
[12]We converted the data set from a wide format to a long format and obtained 10,062 data points for fake posts.



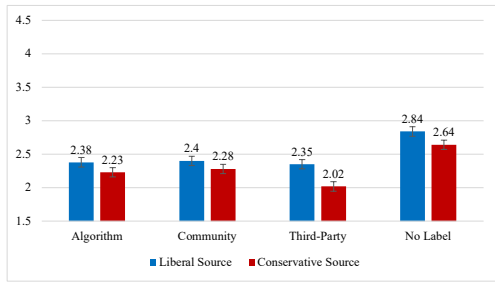
(a) Perceived Accuracy (Conservative Participant)



(b) Perceived Accuracy (Liberal Participant)



(c) Perceived Believability (Conservative Participant)



(d) Perceived Believability (Liberal Participant)

Fig. 2. The perceived accuracy and believability of fake posts with different labels

partially supported for conservative participants. For conservatives, when the source ideology aligned with their political beliefs, algorithmic misinformation labels (M=3.44, SD= .07) were more effective in reducing the perceived accuracy compared to community labels (M=3.93, SD= .08), p= .019. For liberal participants, however, algorithmic misinformation labels (M=2.297, SD= .07) were not significantly different from community labels (M=2.293, SD= .06) when the posts aligned with their political beliefs.

H2b predicted that for posts that do not align with their beliefs, there would be no difference in effectiveness between algorithmic misinformation labels and community misinformation labels in reducing people's believability of fake posts, which was supported. For conservatives, when the source ideology does not align with their political beliefs, algorithmic labels (M=3.17, SD= .08) were not significantly different from community labels in terms of perceived accuracy of fake posts (M=3.36, SD= .07), p= .35. For liberals, algorithmic labels (M=2.06, SD= .06) were not significantly different from community labels in terms of perceived accuracy of fake posts (M=2.21, SD= .07), p= .54.

H3 predicted that for posts that align with their beliefs, the third-party fact-checkers' labels would be more effective in reducing people's perceived accuracy and believability of fake posts than algorithmic (**H3a**) and community (**H3b**) labels, which was not supported. Algorithmic labels performed as well as the third-party fact-checkers' labels (**H3a**), as shown in Table 3. Likewise, there was no significant difference in the effect of third-party fact-checkers' labels and community labels (**H3b**).

H4 predicted that for posts that do not align with their beliefs, the third-party fact-checkers' labels would be more effective in reducing people's perceived accuracy and believability of fake posts than algorithmic (**H4a**) and community (**H4b**) labels, which was not supported. There was no significant difference between the effect of third-party fact-checkers' labels and algorithmic

Source Ideology	Label Comparison	Perceived Accuracy		Perceived Believability	
		Mean Difference	*p*	*Mean Difference*	*p*
Conservative Participant					
Conservative	Algorithm / No Label	-.49	< .001***	-.53	< .001***
	Community / No Label	-.18	.73	-.23	.27
	Third-Party / No Label	-.37	< .001***	-.30	< .001***
	Algorithm / Community	-.32	.019*	-.29	.044*
	Algorithm / Third-Party	-.13	1.00	-.22	.21
	Community / Third-Party	.19	.48	.07	1.00
Liberal	Algorithm / No Label	-.78	< .001***	-.70	< .001***
	Community / No Label	-.58	< .001***	-.55	< .001***
	Third-Party / No Label	-.65	< .001***	-.58	< .001***
	Algorithm / Community	-.20	.35	-.16	.86
	Algorithm / Third-Party	-.13	1.00	-.13	1.00
	Community / Third-Party	.07	1.00	.03	1.00
Liberal Participant					
Conservative	Algorithm / No Label	-.48	< .001***	-.41	< .001***
	Community / No Label	-.32	< .001***	-.36	< .001***
	Third-Party / No Label	-.51	< .001***	-.51	< .001***
	Algorithm / Community	-.15	.54	-.05	1.00
	Algorithm / Third-Party	.03	1.00	.10	1.00
	Community / Third-Party	.08	1.00	.14	.71
Liberal	Algorithm / No Label	-.51	< .001***	-.46	< .001***
	Community / No Label	-.51	< .001***	-.44	< .001***
	Third-Party / No Label	-.60	< .001***	-.49	< .001***
	Algorithm / Community	.00	1.00	-.02	1.00
	Algorithm / Third-Party	.03	1.00	.03	1.00
	Community / Third-Party	.18	.23	.05	1.00

Table 3. Pairwise comparisons of perceived accuracy and believability * $p < .05$, ** $p < .01$, *** $p < .001$

labels (**H4a**). No significant difference was found between third-party fact-checkers' labels and community labels (**H4b**).

We conducted exploratory analyses to test the effect of labels on people's sharing, liking, and commenting intention. We used linear mixed models while controlling for participant ID and post ID as random effects. Results showed that there was a significant main effect of labels on sharing, $F(3,10046)=10.64$, $p< .001$, and liking intention, $F(3,10046)=13.37$, but no main effect of labels on commenting intention, $F(3,10046)= .85$, $p = .47$. There was a significant main effect of source ideology and participant's ideology on sharing, commenting, and liking intention, $p< .001$. There was also a significant 3-way interaction effect among the participant's political leaning, source ideology, and labels on sharing, liking, and commenting intention of fake posts, $p< .001$. We also reported the results of pairwise comparisons in Table 4 and Table 6 (in Appendix). The differences among label conditions were very small as people had overall low sharing, liking, and commenting intention.

Additional analyses were conducted to further examine the effect of labels on people's perceptions of misinformation labels including label effectiveness, objectiveness, and whether the labels were mechanical and politically unbiased or not. Results showed that there were significant main effects of the label type, $p< .001$ and a significant main effect of participant's political leaning on people's

Source Ideology	Label	Sharing Intention		Liking Intention		Commenting Intention	
		Mean	*SE*	*Mean*	*SE*	*Mean*	*SE*
Conservative Participant							
Conservative	Algorithm	2.43	.07	2.17	.06	2.37	.06
	Community	2.63	.08	2.13	.07	2.14	.07
	Third-Party	2.21	.07	1.93	.06	2.11	.07
	No Label	2.65	.08	2.31	.07	2.34	.07
Liberal	Algorithm	2.08	.07	1.76	.06	1.92	.07
	Community	2.14	.07	2.01	.06	2.02	.06
	Third-Party	2.07	.07	1.88	.06	2.01	.07
	No Label	2.26	.07	2.01	.06	2.11	.06
Liberal Participant							
Conservative	Algorithm	1.46	.05	1.36	.04	1.79	.06
	Community	1.56	.05	1.46	.04	1.94	.06
	Third-Party	1.41	.05	1.37	.04	2.03	.06
	No Label	1.61	.05	1.51	.04	1.70	.06
Liberal	Algorithm	1.60	.05	1.56	.05	1.89	.06
	Community	1.65	.05	1.51	.04	1.96	.06
	Third-Party	1.68	.05	1.50	.04	1.83	.06
	No Label	1.87	.06	1.67	.05	2.07	.06

Table 4. The Means (*SE*) of sharing, liking, and commenting intention of fake posts

label perceptions (effectiveness, politically unbiased, objectiveness, mechanical), $p < .001$. There was also a significant 3-way interaction effect among the participant's political leaning, source ideology, and label perceptions, $p < .001$. Pairwise comparisons showed that for both conservative and liberal participants, algorithmic labels were overall perceived as significantly more effective, politically unbiased, objective, and mechanical than community labels for posts that both aligned and did not align with their beliefs. For conservative participants, algorithmic labels were even perceived as more politically unbiased, objective, and mechanical than third-party fact-checkers' labels. For liberal participants, however, the results were mixed. When liberal participants viewed conservative posts, they perceived third-party fact-checkers' labels as more politically unbiased, and more effective than algorithmic labels; when liberal participants viewed liberal posts, they considered algorithmic labels as more politically unbiased, as shown in Table 7 and Table 8 (in Appendix).

6 DISCUSSION

6.1 Theoretical Implications

Our study sheds light on the differing effects of various misinformation labels depending on partisan alignment. One interesting finding was that algorithmic misinformation labels outperformed community misinformation labels when conservatives viewed ideologically consistent posts. One possible explanation for why algorithmic labels were more effective for hyper-partisan misinformation than community labels is that people may have positive stereotypes about machine infallibility. Previous research suggests that the machine heuristic may reduce the perceived news bias [67, 70]. Our results provided more evidence on the machine heuristic assumption [56].

Another possible explanation for such results is conservative partisans' mistrust of other social media users. Past work indicates the tendency of conservatives to be less trusting of media outlets, social media platforms, and journalists than liberals [49, 65, 72]. Between 1997 to 2016, trust in mainstream media by Democrats had decreased to 50% while trust by Republicans had gone down from 41% to 14% [72]. Such mistrust by conservatives extends into perceptions of technology companies and social media platforms as well: 90% of Republicans believe that social media platforms intentionally censor political opinions (vs. 59% of Democrats), and 69% of Republicans believe major technology companies favor liberal views over those of conservatives (vs. 19% of Democrats believing the sites support conservative views over liberal ones) [65]. In a survey asking about attitudes towards journalists, conservative users on social media agreed more than liberals regarding statements that journalists were immoral, ego-driven, and affected biased mainstream media due to their own politics [49]. It is possible that the mistrust in social media platforms and journalists is projected onto our community labels by conservative participants who are wary of *who* other users may be if selected by the platform. Conversely, they may view the algorithm as being a somewhat neutral source due to our explanation describing the algorithm as being created in conjunction with other platforms, which could make it appear more acceptable.

Conservative users' skepticism towards social media platforms can also explain why conservative participants in our study rated all types of labels as less effective, objective, and politically unbiased than liberal participants. Conservatives' low trust in labels can be associated with another finding from our study that conservative participants rated fake posts with labels as more accurate and believable than liberal participants. Our results also showed that conservative participants have overall higher perceived accuracy and believability of fake posts than liberal participants even for unlabeled posts. Such a result is not surprising as past work suggests conservatives are more likely to fall for misinformation than liberals [46].

Another interesting result indicated that both algorithmic and third-party fact-checker indicators reduced people's perceived accuracy and believability of fake posts regardless of the post's ideology. In fact, for conservative participants, algorithmic labels were even considered to be more politically unbiased, objective, and mechanical than third-party fact-checkers' labels. This drop in belief of third-party fact-checkers being unbiased reflects trends found in surveys conducted by Pew Research over attitudes held towards third-party fact-checkers. According to one Pew Research survey, 70% of Republicans felt fact-checkers favored one side (while only 29% of Democrats felt fact-checkers favored one side) [68]. Additionally, when asked their views on social media companies flagging inaccurate information, just 27% of Republicans said they approved of this type of activity (vs. 73% of Democrats that at least somewhat approve of this practice) [65]. Of the social media companies that have explained how their fact-checking process works, they typically outline a partnership with third-party fact-checking companies to verify potential misinformation [2], so it can be surmised that these results indicate mistrust of social media third-party fact-checkers already in place. Thus our results around the algorithmic label being seen as less biased than third-party fact-checkers combined with the algorithmic label reducing the accuracy of fake posts by conservative participants indicate a promising direction for additional work around the design of algorithmic labels. Many conservative users might be disapproving of labelling misinformation on social media platforms, but they may still pay attention to the labels if labels are attributed to sources that users deem acceptable or neutral.

Our work also broadens the field of misinformation studies by examining social media posts instead of news headlines. Past studies related to interventions with misinformation have primarily tested people's identification and rating of misinformation through the format of news headlines. These studies present the headlines as they would appear on Facebook: in a standalone format with a photo, headline, and brief description [12, 48], and they may include a user's name and default

profile photo that is intentionally blurred or nondescript [31, 33]. In contrast, our study explores a new space: we build a website to simulate people's perceptual and behavioral responses on social media and examine a new information format (user-level posts). We intentionally mimic Twitter, an environment where users are more likely to use functions such as "retweeting" to share articles and thus disseminate information much more rapidly compared to headlines on Facebook. Our web tool also displays names, account handles, and short bios to deepen a realistic social media platform experience. We include in the selection of posts displayed to participants some posts that have an insertion of personal beliefs or opinions (e.g., *"It's widely known that <10k died of COVID alone and that HCQ (plasma, UV Light therapy, others) will cure future reoccurrence. Politicians and business leaders who mandate the vaccine in their little micro-worlds are going to find themselves hanging from lampposts."*), which is atypical of headlines used in previous studies but common in social media posts.

Our analyses reveal small but non-negligible effects of labels on people's sharing, liking, and commenting intention. The small differences among label conditions were not surprising as people had overall low sharing, liking, and commenting intention of those fake posts. This result was consistent with findings from past work. One recent study shows a discrepancy between people's sharing intention and accuracy judgment of fake posts [45]. This may also explain why Yaqub et al. [75]'s study using sharing intent as the outcome variable exhibits some different patterns with our study. Their findings suggest that AI was least effective in reducing peoples' sharing intentions while ours suggest that algorithmic labels perform as well as third-party fact-checker labels and outperform community labels in some cases.

6.2 Practical Implications

Practically, we shed light on what types of misinformation labels are effective in nudging partisans to more mindfully assess and share social media content. Such results can have implications for widely used social media platform designs in the industry about future mechanisms for misinformation detection and label design. Currently, fact-checking performed on social media websites is done by third-party fact-checkers which may be slow and fail to quickly identify harmful misinformation before it spreads, and recently, companies such as Twitter have begun to explore community-based fact-checking. Our study shows that misinformation identification attributed to algorithms or other platform users can have the same effect as third-party fact-checkers for improving a person's ability to accurately identify misinformation. This can allow for rapid detection of misinformation to prevent the spread of misinformation more effectively than current techniques.

However, the types of errors by real-time misinformation detection and people's perceptions of errors must be carefully assessed to gauge the real-world trust and impact of them before deploying these systems. Researchers have made strides in automated misinformation detection techniques such as creating larger and more robust data sets for algorithm training [71] and developing hybrid models that consider multiple dimensions of misinformation such as content, metadata, source, and user engagement [50]. However, achieving a highly accurate and accountable misinformation detection algorithm remains challenging [53] due to the wide diversity of topics and features that cannot be encompassed in one data set [54] and the very nature of misinformation being to mislead and deceive users [52]. Thus, although real-time detection methods offer the potential benefit of scaling to catch high volumes of misinformation ahead of harmful dissemination, they are also prone to error or biases which may contribute to harmful misinformation spreading or even decreased user trust in labels or platforms.

The platform that we simulated, Twitter, has a very distinct user base in terms of both who uses the platform as well as who is actually actively producing Tweets and other forms of content. Twitter's users tend to be more Democratic than Republican, with the users producing the vast

majority of Tweets also skewing Democratic [9]. If aware of this information, the knowledge about users on Twitter could have also influenced conservative participants, and these participants may have been skeptical of the heterogeneity of users they imagined to be involved in labelling misinformation. Republicans also believe that social media platforms intentionally censor political viewpoints they find objectionable at a much higher rate (90%) compared to Democrats (59%) [65]. 69% of Republicans also believe that technology companies support the views of liberals over conservatives, compared to only 22% that believe that they are supported equally [65]. These factors may result in a general distrust of community-based labelling among conservatives, believing that hand-selected users will skew liberal and be proxies for liberal-biased companies.

Less work exists exploring how users perceive misinformation in more informal, conversational settings such as Tweets on the Twitter platform where posts are often written and viewed with more fleeting attention and frequently integrate user opinion when discussing current events [77]. Additionally, while studying misinformation through how users can detect false news headlines is valuable, misinformation embedded in users' posts can be more rapidly spread due to ease of access to the platform by almost everyone to post or share posts. Understanding misinformation in the context of forms like Tweets is of particular importance as users can and often do combine their own political opinions with misinformation, potentially exacerbating hyper-partisan tendencies and masking signals of misinformation.

Furthermore, prior studies found that ideological extremists are more likely to spread misinformation on Twitter compared to other social media platforms such as Facebook [24]. The reasoning behind this was that Facebook has an implied 'real name' policy providing a more normative social application built on personal and identity-linked information utilizing many of your real-world social connections. Twitter, on the other hand, is comparatively weak in rich personal interactions, resulting in the platform being rife with trolling and other anti-social behavior [42].

Twitter's launching of its 'Birdwatch' pilot in January 2021 means that user- or community-based misinformation verification processes may become increasingly common. Community-based platforms are potentially susceptible to abuse and misuse, and more needs to be studied regarding the potential for political exploitation, given social media's current role in American society. As the fight against misinformation develops, companies will be forced to adapt and create more innovative methods to confront users with the idea that their content may be misleading and betraying them. However, recent work has shown that crowd-sourcing for misinformation detection does not necessarily result in the abuse of these reporting mechanisms, but can potentially produce reliable misinformation identification [2, 16]. Birdwatch relies on community-based labeling whereas our study examines the effects of three types of labels (algorithmic, community, and third-party labels).

The results of this paper indicate that all three types of labels were generally effective in reducing the believability of fake posts, both for posts that align with the participant's political ideology, as well as for posts that don't. This is interesting because it demonstrates that the previously under-researched algorithmic and community-based misinformation detection platforms are also effective in dissuading social media users. Our results confirmed that for conservative participants, algorithmic labels are more effective than community labels in reducing the perceived accuracy of conservative posts. However, this is not the case for liberal participants, for whom both community labels and algorithmic labels were effective. Despite this difference, the opportunity still opens for real-world social media platforms to install such community-based misinformation detection algorithms. The advantages of community-based (faster, more egalitarian, more transparent, larger-scale) and algorithmic (even faster, automated, highly manageable) techniques over more traditional third-party misinformation detection methods are obvious. Research has indicated that crowd-based misinformation detection can even be as accurate as third-party independent fact-checkers [2].

With these results, real-world platforms such as Birdwatch should feel more comfortable with rolling out these features in the near future.

6.3 Limitations

Despite our theoretical and practical contributions, this work had certain limitations. First, our study exposes participants to 6 true posts and 6 false posts, a greater ratio of false posts than any real social media platform. Even though such a ratio is commonly used in previous misinformation studies, it can potentially make people overall more skeptical to those posts by exposing them to several fake posts at one time. Additionally, although we selected our total post count (12) shown to participants based on the number of posts used in prior misinformation studies, we recognize that this is only a small amount of posts a social media user may be exposed to in a day. We did an analysis of the effectiveness of the labels on each fake post. The main effect of label condition was significant across each fake post except for one single post (*"It's widely known that <10k died of COVID alone and that HCQ (plasma, UV Light therapy, others) will cure future reoccurrence. Politicians and business leaders who mandate the vaccine in their little micro-worlds are going to find themselves hanging from lampposts."*). The interaction effect of the label condition and the participant's political ideology (i.e.,"algorithmic labels were more effective than community labels when conservative users view conservative posts") was significant only when we tested it with all fake posts; over a single post, it was either marginal significant or not significant possibly due to the small sample size.

Second, the set of posts we displayed to participants was limited to one context, COVID-19. While examining misinformation in this context is novel due to the emerging nature of the pandemic and information shared around it, the topics on social media platforms are diverse and may not carry the same societal polarization as COVID-19. Thus, this selection of COVID-19 posts may affect the generalizability of our results beyond COVID-19 related misinformation. Conservatives and people that ingest a steady diet of conservative news are at a higher susceptibility to incorrectly believe COVID-19 misinformation [8, 58]. This does, however, potentially limit the implications of this study, because of COVID-19's role as a highly politicized ideological issue in American society, meaning the findings may not be exportable to other issues [22].

Third, we asked about political bias under each post, which may reinforce people's impression of source ideology, potentially leading to a less natural mindset than the real-world social media setting. Even though our website closely simulated social media platforms, it still has functional limits such as the inability to simulate true sharing behavior. Future work can measure people's actual sharing behavior on similar simulated websites instead of self-reported sharing intentions.

A fourth limitation is the default Twitter profile pictures we used in the study. While it may be emblematic of 'bot' accounts with low credibility, the default icon was chosen to reduce potential gender and racial bias with human images. We also acknowledge that the stimuli we use were restricted to the text format so the results may not be generalized to social media posts in other formats such as images.

Fifth, even though we found that labels attributed to algorithms are effective in reducing the perceived accuracy of fake posts, we admit that algorithmically generated labels run the risk of making errors, including biased errors against specific groups [3, 6]. Thus, while algorithms may be faster in labeling misinformation than professional fact-checkers, any algorithmic system designed for misinformation detection must consider how to reduce harm in users from mistakes, recover from mistakes made, and ensure that the algorithm itself is not biased in its detection such that errors or failures to detect misinformation do not exacerbate harm on specific populations.

Finally, our participants were recruited from CloudResearch and thus were not nationally representative [10]. Moreover, workers from crowd-sourcing platforms such as CloudResearch and

Amazon Mechanical Turk are known to have relatively high information literacy and technical expertise than the general population [75]. Such sample characteristics may affect the generalizability of our results.

7 CONCLUSION

Our work added a new dimension to hyper-partisan misinformation studies by examining the impacts of algorithmic, community vs. third-party fact-checkers' labels depending on people's political ideology. Our results showed that both algorithmic and third-party fact-checkers' labels can reduce people's perceived accuracy and believability of fake posts regardless of the post's ideology, with no significant difference. We also found that algorithmic labels were more effective in reducing people's believability of fake posts than community labels when conservative users view ideologically-consistent posts. Our work sheds light on the effectiveness of real-time labels, which provides important theoretical and practical implications for automated and community-based misinformation detection approaches.

ACKNOWLEDGEMENT

We thank our participants who provided valuable insights. Our research was supported by UT Austin School of Information and Good Systems[13], a UT Austin Grand Challenge to develop responsible AI technologies.

REFERENCES

[1] Jon Agley and Yunyu Xiao. 2021. Misinformation about COVID-19: evidence for differential latent profiles and a strong association with trust in science. *BMC Public Health* 21, 1 (2021), 1–12.
[2] Jennifer Allen, Antonio A Arechar, Gordon Pennycook, and David G Rand. 2020. Scaling up fact-checking using the wisdom of crowds. *Preprint at https://doi. org/10.31234/osf. io/9qdza* (2020).
[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. ProPublica (2016). *Google Scholar* (2016), 23.
[4] Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer. 2020. Exposure to social engagement metrics increases vulnerability to misinformation. *arXiv preprint arXiv:2005.04682* (2020).
[5] Bence Bago, David G Rand, and Gordon Pennycook. 2020. Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of experimental psychology: general* 149, 8 (2020), 1608.
[6] Solon Barocas and Andrew D Selbst. 2016. Big data's disparate impact. *Calif. L. Rev.* 104 (2016), 671.
[7] Adam J Berinsky. 2017. Rumors and health care reform: Experiments in political misinformation. *British journal of political science* 47, 2 (2017), 241–262.
[8] Dustin P Calvillo, Bryan J Ross, Ryan JB Garcia, Thomas J Smelter, and Abraham M Rutchick. 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 (2020), 1119–1128.
[9] Pew Research Center. 2021. How Democrats and Republicans Use Twitter. https://www.pewresearch.org/politics/2020/10/15/differences-in-how-democrats-and-republicans-behave-on-twitter/
[10] Jesse Chandler, Cheskie Rosenzweig, Aaron J Moss, Jonathan Robinson, and Leib Litman. 2019. Online panels in social science research: Expanding sampling methods beyond Mechanical Turk. *Behavior research methods* 51, 5 (2019), 2022–2038.
[11] Rumman Chowdhury and Luca Belli. 2021. Examining algorithmic amplification of political content on Twitter. https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent
[12] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 2020. Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior* 42, 4 (2020), 1073–1095.
[13] CNN. 2020. National Results 2020 Presidential Exit Polls. https://www.cnn.com/election/2020/exit-polls/president/national-results

[13]https://goodsystems.utexas.edu

[14] Nicholas Dias, Gordon Pennycook, and David G Rand. 2020. Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media. *Harvard Kennedy School Misinformation Review* 1, 1 (2020).
[15] James N Druckman. 2001. The implications of framing effects for citizen competence. *Political behavior* 23, 3 (2001), 225–256.
[16] Ziv Epstein, Gordon Pennycook, and David Rand. 2020. Will the crowd game the algorithm? Using layperson judgments to combat misinformation on social media by downranking distrusted sources. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–11.
[17] Robert Faris, Hal Roberts, Bruce Etling, Nikki Bourassa, Ethan Zuckerman, and Yochai Benkler. 2017. Partisanship, propaganda, and disinformation: Online media and the 2016 US presidential election. *Berkman Klein Center Research Publication* 6 (2017).
[18] Mingkun Gao, Ziang Xiao, Karrie Karahalios, and Wai-Tat Fu. 2018. To label or not to label: The effect of stance and credibility labels on readers' selection and perception of news articles. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–16.
[19] Anat Gesser-Edelsburg, Alon Diamant, Rana Hijazi, and Gustavo S Mesch. 2018. Correcting misinformation by health organizations during measles outbreaks: a controlled experiment. *PloS one* 13, 12 (2018), e0209505.
[20] Tarleton Gillespie. 2014. The relevance of algorithms. *Media technologies: Essays on communication, materiality, and society* 167, 2014 (2014), 167.
[21] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 US presidential election. *Science* 363, 6425 (2019), 374–378.
[22] Guy Grossman, Soojong Kim, Jonah M Rexer, and Harsha Thirumurthy. 2020. Political partisanship influences behavioral responses to governors' recommendations for COVID-19 prevention in the United States. *Proceedings of the National Academy of Sciences* 117, 39 (2020), 24144–24153.
[23] Michael Hameleers and Toni GLA van der Meer. 2020. Misinformation and polarization in a high-choice media environment: How effective are political fact-checkers? *Communication Research* 47, 2 (2020), 227–250.
[24] Toby Hopp, Patrick Ferrucci, and Chris J Vargo. 2020. Why do people share ideologically extreme, false, and misleading content on social media? A self-report and trace data–based analysis of countermedia content dissemination on Facebook and Twitter. *Human Communication Research* 46, 4 (2020), 357–384.
[25] Benjamin Horne and Sibel Adali. 2017. This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11.
[26] Seyedmehdi Hosseinimotlagh and Evangelos E Papalexakis. 2018. Unsupervised content-based identification of fake news articles with tensor decomposition ensembles. In *Proceedings of the Workshop on Misinformation and Misbehavior Mining on the Web (MIS2)*.
[27] Suchita Jain, Vanya Sharma, and Rishabh Kaushal. 2016. Towards automated real-time detection of misinformation on Twitter. In *2016 International conference on advances in computing, communications and informatics (ICACCI)*. IEEE, 2015–2020.
[28] Chenyan Jia and Thomas J Johnson. 2021. Source Credibility Matters: Does Automated Journalism Inspire Selective Exposure? *International Journal of Communication* 15 (2021), 22.
[29] Thomas J Johnson and Barbara K Kaye. 2015. Reasons to believe: Influence of credibility on motivations for using social networks. *Computers in human behavior* 50 (2015), 544–555.
[30] Dan M Kahan. 2017. Misconceptions, misinformation, and the logic of identity-protective cognition. (2017).
[31] Antino Kim and Alan R Dennis. 2019. Says who? The effects of presentation format and source rating on fake news in social media. *Mis quarterly* 43, 3 (2019).
[32] Antino Kim, Patricia L Moravec, and Alan R Dennis. 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 (2019), 931–968.
[33] Jan Kirchner and Christian Reuter. 2020. Countering fake news: A comparison of possible solutions regarding user acceptance and effectiveness. *Proceedings of the ACM on Human-computer Interaction* 4, CSCW2 (2020), 1–27.
[34] David MJ Lazer, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, et al. 2018. The science of fake news. *Science* 359, 6380 (2018), 1094–1096.
[35] Leib Litman, Jonathan Robinson, and Tzvi Abberbock. 2017. TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods* 49, 2 (2017), 433–442.
[36] Ruibo Liu, Chenyan Jia, and Soroush Vosoughi. 2021. A transformer-based framework for neutralizing and reversing the political polarity of news articles. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–26.
[37] Patricia Moravec, Randall Minas, and Alan R Dennis. 2018. Fake news on social media: People believe what they want to believe when it makes no sense at all. *Kelley School of Business Research Paper* 18-87 (2018).

[38] Patricia L Moravec, Antino Kim, and Alan R Dennis. 2020. Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research* 31, 3 (2020), 987–1006.
[39] Rachel R Mourão and Craig T Robertson. 2019. Fake news as discursive integration: An analysis of sites that publish false, misleading, hyperpartisan and sensational information. *Journalism Studies* 20, 14 (2019), 2077–2095.
[40] Sean A Munson and Paul Resnick. 2010. Presenting diverse political opinions: how and how much. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1457–1466.
[41] Sarah Myers West. 2018. Censored, suspended, shadowbanned: User interpretations of content moderation on social media platforms. *New Media & Society* 20, 11 (2018), 4366–4383.
[42] Mustafa Oz, Pei Zheng, and Gina Masullo Chen. 2018. Twitter versus Facebook: Comparing incivility, impoliteness, and deliberative attributes. *New media & society* 20, 9 (2018), 3400–3419.
[43] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* 66, 11 (2020), 4944–4957.
[44] Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general* 147, 12 (2018), 1865.
[45] Gordon Pennycook, Ziv Epstein, Mohsen Mosleh, Antonio A Arechar, Dean Eckles, and David G Rand. 2021. Shifting attention to accuracy can reduce misinformation online. *Nature* 592, 7855 (2021), 590–595.
[46] Gordon Pennycook, Jonathon McPhetres, Bence Bago, and David G Rand. 2020. Predictors of attitudes and misperceptions about COVID-19 in Canada, the UK, and the USA. *PsyArXiv* 10 (2020), 1–25.
[47] Gordon Pennycook, Jonathon McPhetres, Yunhao Zhang, Jackson G Lu, and David G Rand. 2020. Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological science* 31, 7 (2020), 770–780.
[48] Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50.
[49] Jennifer Rauch. 2019. Comparing progressive and conservative audiences for alternative media and their attitudes towards journalism. *Alternative media meets mainstream politics: Activist nation rising* 19 (2019).
[50] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. Csi: A hybrid deep model for fake news detection. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 797–806.
[51] Haeseung Seo, Aiping Xiong, and Dongwon Lee. 2019. Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation. In *Proceedings of the 10th ACM Conference on Web Science*. 265–274.
[52] Karishma Sharma, Feng Qian, He Jiang, Natali Ruchansky, Ming Zhang, and Yan Liu. 2019. Combating fake news: A survey on identification and mitigation techniques. *ACM Transactions on Intelligent Systems and Technology (TIST)* 10, 3 (2019), 1–42.
[53] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data* 8, 3 (2020), 171–188.
[54] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD explorations newsletter* 19, 1 (2017), 22–36.
[55] Craig Silverman. 2016. This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed news* 16 (2016).
[56] S Shyam Sundar and Jinyoung Kim. 2019. Machine heuristic: When we trust computers more than humans with our personal information. In *Proceedings of the 2019 CHI Conference on human factors in computing systems*. 1–9.
[57] Edson C Tandoc Jr, Lim Jia Yao, and Shangyuan Wu. 2020. Man vs. machine? The impact of algorithm authorship on news credibility. *Digital Journalism* 8, 4 (2020), 548–562.
[58] Xian Teng, Yu-Ru Lin, Wen-Ting Chung, Ang Li, and Adriana Kovashka. 2021. Characterizing User Susceptibility to COVID-19 Misinformation on Twitter. *arXiv preprint arXiv:2109.09532* (2021).
[59] Jason Bennett Thatcher, Ryan T Wright, Heshan Sun, Thomas J Zagenczyk, and Richard Klein. 2018. Mindfulness in information technology use: Definitions, distinctions, and a new measure. *MIS Quarterly* 42, 3 (2018), 831–848.
[60] Daniel Thomas. 2017. Facebook to tackle fake news with educational campaign. https://www.bbc.com/news/technology-39517033
[61] André Calero Valdez and Martina Ziefle. 2018. Believability of News. In *Congress of the International Ergonomics Association*. Springer, 469–477.
[62] Jay J Van Bavel and Andrea Pereira. 2018. The partisan brain: An identity-based model of political belief. *Trends in cognitive sciences* 22, 3 (2018), 213–224.
[63] Emily Van Duyn and Jessica Collier. 2019. Priming and fake news: The effects of elite discourse on evaluations of news media. *Mass Communication and Society* 22, 1 (2019), 29–48.

[64] Guido M Van Koningsbruggen, Tilo Hartmann, Allison Eden, and Harm Veling. 2017. Spontaneous hedonic reactions to social media cues. *Cyberpsychology, Behavior, and Social Networking* 20, 5 (2017), 334–340.
[65] Emily A. Vogels, Andrew Perrin, and Monica Anderson. 2020. Most Americans Think Social Media Sites Censor Political Viewpoints. https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints/
[66] Emily K Vraga, Leticia Bode, and Melissa Tully. 2020. Creating news literacy messages to enhance expert corrections of misinformation on Twitter. *Communication Research* (2020), 0093650219898094.
[67] T Franklin Waddell. 2019. Can an algorithm reduce the perceived bias of news? Testing the effect of machine attribution on news readers' evaluations of bias, anthropomorphism, and credibility. *Journalism & Mass Communication Quarterly* 96, 1 (2019), 82–100.
[68] Mason Walker and Jeffrey Gottfried. 2020. Republicans far more likely than Democrats to say fact-checkers tend to favor one side. https://www.pewresearch.org/fact-tank/2019/06/27/republicans-far-more-likely-than-democrats-to-say-fact-checkers-tend-to-favor-one-side/
[69] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.
[70] Sai Wang. 2021. Moderating Uncivil User Comments by Humans or Machines? The Effects of Moderation Agent on Perceptions of Bias and Credibility in News Content. *Digital Journalism* 9, 1 (2021), 64–83.
[71] William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648* (2017).
[72] Harry Wessel. 2018. Messengers of the right: conservative media and the transformation of American politics. *Political Science Quarterly* 133, 1 (2018), 181–183.
[73] Magdalena Wojcieszak, Arti Thakur, João Fernando Ferreira Gonçalves, Andreu Casas, Ericka Menchen-Trevino, et al. 2021. Can AI Enhance People's Support for Online Moderation and Their Openness to Dissimilar Political Views? *Journal of Computer-Mediated Communication* (2021).
[74] Weiai Wayne Xu, Yoonmo Sang, and Christopher Kim. 2020. What drives hyper-partisan news sharing: Exploring the role of source, style, and content. *Digital Journalism* 8, 4 (2020), 486–505.
[75] Waheeb Yaqub, Otari Kakhidze, Morgan L Brockman, Nasir Memon, and Sameer Patil. 2020. Effects of credibility indicators on social media news sharing intent. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–14.
[76] Savvas Zannettou. 2021. " I Won the Election!": An Empirical Analysis of Soft Moderation Interventions on Twitter. *arXiv preprint arXiv:2101.07183* (2021).
[77] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing twitter and traditional media using topic models. In *European conference on information retrieval*. Springer, 338–349.

APPENDIX

Received July 2021; revised November 2021; accepted February 2022

Source Ideology	Label	Perceived Accuracy		Perceived Believability	
		Mean	*SE*	*Mean*	*SE*
Conservative Participant					
Conservative	Algorithm	3.44	.07	3.51	.07
	Community	3.76	.08	3.81	.08
	Third-Party	3.56	.08	3.74	.08
	No Label	3.93	.08	4.04	.08
Liberal	Algorithm	3.17	.08	3.27	.08
	Community	3.36	.07	3.43	.07
	Third-Party	3.30	.08	3.40	.08
	No Label	3.95	.07	3.97	.07
Liberal Participant					
Conservative	Algorithm	2.06	.06	2.23	.07
	Community	2.21	.07	2.28	.07
	Third-Party	2.03	.06	2.14	.06
	No Label	2.53	.06	2.64	.06
Liberal	Algorithm	2.30	.07	2.38	.07
	Community	2.29	.06	2.40	.06
	Third-Party	2.21	.06	2.35	.06
	No Label	2.81	.07	2.84	.07

Table 5. The Means (*SE*) of perceived accuracy and believability of fake posts

Source Ideology	Label Comparison	Sharing Intention		Liking Intention		Commenting Intention	
		Mean Diff.	p	Mean Diff.	p	Mean Diff.	p
Conservative Participant							
Conservative	Algorithm / No Label	-.22	.24	-.14	.86	.03	1.00
	Community / No Label	-.03	1.00	-.17	.45	-.20	.28
	Third-Party / No Label	-.44	$< .001^{***}$	$-.371^{*}$	$< .001^{***}$	-.23	.10
	Algorithm / Community	-.20	.40	.04	1.00	.22	.10
	Algorithm / Third-Party	.22	.20	.24	$.05^{*}$	.26	$.03^{*}$
	Community / Third-Party	.41	.00	.20	.21	.04	1.00
Liberal	Algorithm / No Label	-.18	.39	-.24	$.03^{*}$	-.18	.26
	Community / No Label	-.12	1.00	.00	1.00	-.08	1.00
	Third-Party / No Label	-.19	.35	-.12	.99	-.10	1.00
	Algorithm / Community	-.06	1.00	-.24	$.03^{*}$	-.10	1.00
	Algorithm / Third-Party	.00	1.00	-.12	1.00	-.09	1.00
	Community / Third-Party	.07	1.00	.12	.98	.01	1.00
Liberal Participant							
Conservative	Algorithm / No Label	-.14	.16	-.16	$.03^{*}$	.09	1.00
	Community / No Label	-.05	1.00	-.06	1.00	.24	$.03^{*}$
	Third-Party / No Label	-.20	$.01^{**}$	-.15	$.05^{*}$	.33	$< .001^{***}$
	Algorithm / Community	-.10	.85	.06	.55	-.16	.43
	Algorithm / Third-Party	.05	1.00	-.01	1.00	-.24	$.02^{*}$
	Community / Third-Party	.15	.13	.09	.80	-.09	1.00
Liberal	Algorithm / No Label	-.27	$< .001^{***}$	-.11	.64	-.18	.21
	Community / No Label	$-.23^{*}$	$.01^{**}$	-.16	.07†	-.11	1.00
	Third-Party / No Label	-.20	$.05^{*}$	-.17	$.05^{*}$	-.24	$.02^{*}$
	Algorithm / Community	-.04	1.00	.05	1.00	-.07	1.00
	Algorithm / Third-Party	-.08	1.00	.06	1.00	.06	1.00
	Community / Third-Party	-.03	1.00	.01	1.00	.13	.58

Table 6. Pairwise comparisons of sharing, liking, and commenting intention † $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Source Ideology	Label	Effective		Politically Unbiased		Objective		Mechanical	
		Mean	*SE*	*Mean*	*SE*	*Mean*	*SE*	*Mean*	*SE*
Conservative Participant									
Conservative	Algorithm	4.35	.07	3.76	.07	4.32	.06	4.86	.06
	Community	3.78	.07	3.22	.07	3.64	.07	4.24	.06
	Third-Party	4.10	.07	3.20	.07	3.89	.06	4.22	.06
Liberal	Algorithm	4.19	.07	3.73	.07	4.20	.07	4.98	.06
	Community	3.91	.07	3.32	.07	3.94	.06	4.15	.06
	Third-Party	4.30	.07	3.49	.07	3.80	.07	4.31	.06
Liberal Participant									
Conservative	Algorithm	5.46	.04	4.92	.06	5.09	.06	5.03	.06
	Community	5.17	.04	4.69	.06	4.74	.06	4.28	.06
	Third-Party	5.67	.04	5.20	.06	5.21	.05	4.38	.06
Liberal	Algorithm	5.39	.05	5.28	.06	5.05	.06	4.73	.06
	Community	5.07	.05	4.62	.06	4.65	.05	3.94	.06
	Third-Party	5.24	.05	4.99	.06	4.88	.05	4.65	.06

Table 7. The Means (*SE*) of label perceptions

Source Ideology	Label Comparison	Effective		Politically Unbiased		Objective		Mechanical	
		Mean Diff.	*p*	Mean Diff.	*p*	Mean Diff.	*p*	Mean Diff.	*p*
Conservative Participant									
Conservative	Algorithm / Community	.57	$< .001^{***}$	.54	$< .001^{***}$	.68	$< .001^{***}$	.61	$< .001^{***}$
	Algorithm / Third-Party	.25	$.02^{*}$	.56	$< .001^{***}$	.42	$< .001^{***}$	.64	$< .001^{***}$
	Community / Third-Party	-.31	$.01^{**}$	.02	1.00	-.25	$.02^{*}$	.02	1.00
Liberal	Algorithm / Community	.28	$.01^{**}$	.41	$< .001^{***}$	.26	$.01^{*}$	.83	$< .001^{***}$
	Algorithm / Third-Party	-.11	.79	.25	$.05^{*}$	.39	$< .001^{***}$	.67	$< .001^{***}$
	Community / Third-Party	-.39	$< .001^{***}$	-.17	.26	.13	.43	-.16	.13
Liberal Participant									
Conservative	Algorithm / Community	.29	$< .001^{***}$	.23	$.03^{*}$	.35	$< .001^{***}$	.75	$< .001^{***}$
	Algorithm / Third-Party	-.21	$< .001^{***}$	-.29	$< .001^{***}$	-.12	.31	.65	$< .001^{***}$
	Community / Third-Party	-.50	$< .001^{***}$	-.51	$< .001^{***}$	-.47	$< .001^{***}$	-.10	.67
Liberal	Algorithm / Community	.32	$< .001^{***}$	.66	$< .001^{***}$	.40	$< .001^{***}$	.79	$< .001^{***}$
	Algorithm / Third-Party	.15	.10	.29	$< .001^{***}$	.17	$.07^{\dagger}$	.08	1.00
	Community / Third-Party	-.18	$.02^{*}$	-.37	$< .001^{***}$	-.23	$.01^{**}$	-.71	$< .001^{***}$

Table 8. Pairwise comparisons of labels perceptions $^{\dagger}\ p < .10$, $^{*}\ p < .05$, $^{**}\ p < .01$, $^{***}\ p < .001$