Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness

JAN KIRCHNER and CHRISTIAN REUTER, Technical University of Darmstadt, Science and Technology for Peace and Security (PEASEC), Germany

Since the emergence of so-called fake news on the internet and in social media, platforms such as Facebook have started to take countermeasures, and researchers have begun looking into this phenomenon from a variety of perspectives. A large number of scientific work has investigated ways to detect fake news automatically. Less attention has been paid to the subsequent step, i.e., what to do when you are aware of the inaccuracy of claims in social media. This work takes a user-centered approach on means to counter identified mis- and disinformation in social media. We conduct a three-step study design on how approaches in social media should be presented to respect the users' needs and experiences and how effective they are. As our first step, in an online survey representative for some factors to the German adult population, we enquire regarding their strategies on handling information in social media, and their opinion regarding possible solutions — focusing on the approach of displaying a warning on inaccurate posts. In a second step, we present five potential approaches for countermeasures identified in related work to interviewees for qualitative input. We discuss (1) warning, (2) related articles, (3) reducing the size, (4) covering, and (5) requiring confirmation. Based on the interview feedback, as the third step of this study, we select, improve, and examine four promising approaches on how to counter misinformation. We conduct an online experiment to test their effectiveness on the perceived accuracy of false headlines and also ask for the users' preferences. In this study, we find that users welcome warning-based approaches to counter fake news and are somewhat critical with less transparent methods. Moreover, users want social media platforms to explain why a post was marked as disputed. The results regarding effectiveness are similar: Warning-based approaches are shown to be effective in reducing the perceived accuracy of false headlines. Moreover, adding an explanation to the warning leads to the most significant results. In contrast, we could not find a significant effect on one of Facebook's current approaches (reduced post size and fact-checks in related articles).

CCS Concepts: • **Human-centered computing → Interactive systems and tools**; **Collaborative and social computing**; **Empirical studies in interaction design**.

Additional Key Words and Phrases: fake news; disinformation; misinformation; user acceptance; survey; design; online experiment

ACM Reference Format:
Jan Kirchner and Christian Reuter. 2020. Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 140 (October 2020), 27 pages. https://doi.org/10.1145/3415211

Authors' address: Jan Kirchner, jan_kristoffer.kirchner@stud.tu-darmstadt.de; Christian Reuter, reuter@peasec.tu-darmstadt.de, Technical University of Darmstadt, Science and Technology for Peace and Security (PEASEC), Prankratiusstraße 2, Darmstadt, 64289, Germany.

2573-0142/2020/10-ART140 $15.00
https://doi.org/10.1145/3415211

1 INTRODUCTION

In recent years, social media networks such as Facebook and Twitter serve increasingly as a vital source for news and information [50]. The propagation of information becomes independent from professional journalism, which facilitates the spread of dubious and fabricated content. Driven by the continuous drop in circulation and the attention-based online market, professional journalism can also participate in similar phenomena like the spread of false rumors or clickbait. In the context of crises and rumors, research showed that social media could facilitate the spread of misinformation [28] - also during Corona pandemic [26]. Previous research also indicated that false rumours are a main barrier to not use social media in crisis situations [51]. Further, social media can become an accomplice to the intentional spread of misinformation: These disinformation can be defined as “false, inaccurate, or misleading information designed, presented and promoted to intentionally cause public harm or for profit” [13]. It can be part of so-called information operations. One famous example is the infiltration of the Twitter discourse in the context of the #BlackLivesMatter movement [4].

Starbird et al. [59] argue that such information operations in social media are also a concern of Computer Supported Cooperative Work (CSCW), and also of Crisis Informatics [50], an interdisciplinary field with roots in CSCW [59]. Furthermore it is also a topic of IT Peace Research [48] as an interdisciplinary field that addresses the "role of IT in peace and security from a theoretical, empirical and technical perspective", according to Reuter [48]. Our paper focuses particularly on the aspect of misinformation from the perspective of CSCW.

Although disinformation has existed for a long time, the recent proceedings in social media have made it a popular topic in public discourse, also known under the term *Fake News*. Allcott and Gentzkow [2] have defined fake news as “news articles that are intentionally and verifiably false and could mislead readers” [2]. The contemporary discourse seems to define fake news as referring to viral posts based on fictitious accounts made to look like news reports [2]. Tandoc et al. [61] examined a broad spectrum of 34 scholarly definitions. In our paper, we follow their suggestion to exclude satire and parody from the definition. In the following we understand fake news to mean all forms of incorrect, inaccurate, or misleading information that are designed, presented, and promoted in such a way that they are intended to cause public damage or to generate profit. It was widely discussed in public discourse, in the context of the 2016 US presidential elections, the Brexit referendum, and the 2017 German parliamentary elections. Fake news is mostly politically motivated, e.g., as part of information operations [47]. Such information operations can even serve as a method of hybrid warfare [48]. Nevertheless, because it often draws much attention, there are also monetary motives for spreading fake news. Although the effects of fake news are still a controversial topic, and research suggests that only a few users are susceptible [14], many people seem to have encountered fake news and interacted with them [49]. There have been cases where fake news have caused severe consequences. For example, in 2013, a fake tweet from a hacked Associated Press (AP) account caused a loss of USD$130 billion in market capitalization in the US stock market [44], while fake news from the #PizzaGate conspiracy theory led to a shooting at the pizza place in question [1].

Since fake news appears just like real news at first glance, users cannot easily discern between the two. As such, programs to educate users and increase media literacy are discussed, for example, in the European Union [13]. There are also dedicated journalist groups and websites for fact-checking news articles. Many researchers are working on computational detection approaches that can detect fake news with less effort and rather high accuracy. As shown by cases like #PizzaGate, misinformation also spreads via social media. Thus, platforms like Facebook, Twitter, and Instagram have taken measures to combat the spread of fake news. Most notably, Facebook has deployed a range of approaches such as red-flagged warnings, showing fact-checking articles as related articles, asking for confirmation before sharing, and reducing the post's size. The countermeasures first introduced by Facebook in 2016 appear to bring about the desired effect of impeding the spread of false information in the social network. Since then, engagement with fake news on Facebook has dropped by more than 50% [3]. A less examined aspect is the users' perspective on the measures taken by their social media platforms. As the main actors, their acceptance and opinion should not be left out of consideration.

Platforms are increasingly taking on measures to counter fake news, of which many are directly visible to the users and affect their experience in the social network. Their point of view is the central aspect of this work. Therefore, our two research questions are:

(1) How should approaches to counter fake news in social media be designed to be accepted by the users of such platforms?
(2) How effective are those approaches to reduce the influence of fake news?

After presenting the related work on the topic of fake news in Section 2, we apply a three-step study. Thereby, the first two steps mainly focus on the first research question. Firstly, we investigate the users' opinion on countermeasures: A survey representative to gender, age, education and income of German adults (18 to 74 years) examines their general opinion on and handling of such approaches — focusing on warnings (Section 4). Its results show that most users prefer disputed content to be tagged with a warning and want the platform to explain why. Secondly, we conduct qualitative semi-structured interviews to get more detailed insights on how possible approaches should look like. Based on existing literature, we compare five countermeasures and ask for the users' opinions on them. All elements follow different strategies to counter misinformation in social media posts: (1) a warning message stating that independent fact-checkers had disputed the article and a link for more information, (2) related articles underneath the post with headlines contradictory to the false claim, (3) reducing the size of the article's photo and headline significantly, (4) covering the article, as known from explicit contents, and (5) requiring confirmation before interacting with the post (Section 5). Based on the findings, we select four promising approaches and further examine them in the third step. Within this, we compared four study groups with different approaches and a control group to which no approach was shown. All participants were shown a total of 24 news articles in the format of a Facebook post. We selected 16 disputed article headlines from various fact-checkers (correctiv.org, a German investigative journalist group, Mimikama, an Austrian fact-checking site, and Snopes, a US site) (Section 6). Here, we also tackle our second research question and compare the effectiveness of those approaches regarding the reduction of the influence of fake news. We discuss our results in Section 7 and finally draw a conclusion in Section 8.

2 RELATED WORK

In recent years, and particularly since the 2016 US presidential election, the term ”fake news” has become a popular buzzword. Strengthened by the emergence of social media, fake news are widely spread and is getting through to many people. Although it may seem that fake news had no impact on election outcomes [2], it can have other serious consequences. Hence, there is a need for strategies to counter fake news in social networks. In general, two steps are necessary: detecting false information and taking action against it. In the following chapter, we will outline different existing approaches for the detection and action against fake news.

2.1 Detecting Fake News in Social Media

There are several methods to detect fake news in social media. For example, platforms can allow their users to report suspicious content, and professional fact-checkers can manually verify or dispute the claims. Furthermore, a large amount of research has looked into computational approaches to detect fake news automatically. Those approaches focus on the characteristics of the text content [23, 24, 45, 65], of the interacting users [54, 58, 60], or the propagation in the social network [37, 57, 63]. Other papers regard the heading's stance towards the body [8], its argumentation [56], and conflicting viewpoints regarding a topic [30]. There is a significant downside of such detection algorithms: They are black boxes that do not explain their outcome. In other machine learning contexts, the need for "interpretability, explainability, and ultimately trustworthiness" is already highlighted and discussed [12]. Explainable machine learning aims to let users build trust in the outcome so that they make use of the systems [52]. However, there are few approaches to fake news detection using such explainable machine learning, such as that by Reis et al. [46] and Yang et al. [64]. Other white box approaches focus on user education and media literacy. Hartwig and Reuter [27] developed the browser extension TrustyTweet that gives hints on which characteristics of a Twitter post might suggest untrustworthy content. Similarly, Bhuiyan et al. [6] presented a browser extension that aims to nudge Twitter users in better news credibility assessment. In the context of information overload subjective character of information quality as well as the need for "tailorability and transparency of filtering" are also being discussed [31].

Recent CSCW research, e.g. [17, 20, 29], furthermore examined the collaborative character of fake news. It outlines the relevance to ascertain whether and how individuals are affected by fake news. They understand misinformation as a social ill that spread widely on social media platforms and that people fundamentally are vulnerable to it. This justifies the necessity to develop fact-checking tools that try to use evidence to rebut misinformation. By detecting linguistic signals in user comments [29] or using (credibility) labels to impact user's news article selection [20], collaborative approaches are presented on how people's perception can be influenced. At the same time, Epstein et al. [17] examined the distortion in the ranking that occurs when search engines are selected in the results. He summarizes the strong and undetectable influence this has on users as Search Engine Manipulation Effect. They propose a browser extension that triggers distortion warnings in real-time [17]. Usually, platforms use a combination of different strategies to detect false information. Facebook offers users the option to report suspicious posts and uses algorithms to detect and prioritize false news posts, which are eventually examined by independent fact-checkers [35, 38].

2.2 Actions Against Detected Fake News

After being aware of which posts contain false claims, the next step is to take action against them. In the aftermath of the 2016 US election, Facebook started to display a warning beneath disputed articles [38] until the end of 2017, when they replaced the warnings for more subtle measures, such as diminishing attention by reducing the post's size, listing fact-checker articles in the related articles section, and ranking it lower in the news feed [33, 35]. In October 2019, Facebook stated — in context of the upcoming 2020 US election — to further invest in better misinformation-identification tools and, thus, reduce the spread of viral misinformation and fake accounts to protect the democratic process. Therefore, they aim to improve fact-checking labels and to help people to better understand the information they see online.Such credibility labels can help to reduce the sharing of fake news [53].

Furthermore, before the emergence of fake news during the US election, Bode and Vraga [7] had already examined the possibility to counter misinformation in social media by providing correcting information in the related articles section below the post. In fall 2014, they conducted two studies on approximately 1,000 participants with identical 2 × 4 between-subject design. They showed them posts containing controversial claims which contradict scientific consensus: claiming a link either between genetically modified organisms (GMOs) and health or between vaccination and autism, depending on the condition. Two headlines in the related article section either confirmed or corrected that misinformation, did both, or contained unrelated information. Beforehand, they asked participants about their initial misperception on the issue. Their results showed that related articles correcting the misinformation in the main post could reduce the misperception among those participants with an initial misperception on the issue of GMOs. They did not find such an effect on the vaccination issue. According to them, the reason is that the vaccination issue has existed way longer than the GMO issue, and the initial misperception is more established and hence harder to reduce [7]. Moreover, browser extensions, such as examined by Ennals et al. [16], can alert users when the information they read online is disputed by a source that they might trust. Users can click on the highlighted text passages and phrases that resemble known disputed claims and the application will show articles that put forward alternative points of view from trustworthy sources. In their study, they found that most of the participants are interested in having a tool like this. Nevertheless, some participants criticized the relatively low fraction of disputed claims and had difficulties to add new claims themselves [16].

Existing research has proven that warning against misinformation can reduce its perceived accuracy [10, 15, 32] but might also backfire [5, 21, 36, 39, 62]. For example, Garrett and Weeks [22] compared immediate corrections of misinformation to delayed corrections. They find that an immediate correction of misinformation has the most significant overall impact on the belief accuracy. However, when misinformation confirms users' attitudes, they find the potential for a backfire effect and delayed corrections to be more efficacious [22]. However, Wood and Porter [62] found no corrections capable of triggering backfire, despite testing precisely the kinds of polarized issues where backfire should be expected. Furthermore, they outlined that the evidence of factual backfire is far more tenuous than prior research suggests. They sum up that citizens heed factual information, even when it challenges their ideological commitments [62].

Pennycook et al. [42] showed that the *Illusory Truth Effect* — repeated exposure to misinformation increasing its perceived accuracy — also applies to fake news stories in social media. More importantly, they also found that warnings can reduce their perceived accuracy. Pennycook et al. [41] confirmed the positive effect of such warnings and further introduced the concept of an *Implied Truth Effect*. Presenting a Bayesian model, they argued that showing warning messages on known false posts not only reduced the belief in them but also increased the belief in (possibly false) posts without such a warning. In a first study, they presented a mix of true and false news headlines to 5,271 American participants. In the control condition, they presented none of the headlines with a warning. In the warning condition, half of the false news headlines were presented with a warning. They asked the participants to estimate each headline's accuracy. As expected, they found both a significant Warning Effect and a significant Implied Truth Effect. They conducted a second study with 2,991 American participants in which they could replicate those findings. Furthermore, they could eliminate the Implied Truth Effect in a third condition by — in addition to showing warnings on false posts — also showing verifications on posts checked to be true [41].

Clayton et al. [11] compared several kinds of warnings. Besides the specific warnings on some false headlines, they also tested a general warning beforehand. Moreover, they examined two different wordings for specific warnings on headlines: "disputed" and "rated false." The results of their study show only minimal effects of a general warning, but significant effects of the specific warnings — confirming the results of Pennycook et al. [41]. They found that "rated false" warnings are significantly more effective than "disputed" warnings. Both effects are more significant than those reported by Pennycook et al. [41] (using "disputed" tags). Moreover, the political congeniality seemed not to have any effect: "(...) the reduction in belief is similar regardless of whether the headline is politically congenial" [11].

Mena [36] confirm the effect of "disputed" warnings. In their study with 501 participants, they showed that a warning could reduce the participants' sharing intention of false news posts. They also found a significant effect on the perceived message credibility. In a mediation test, they could show that "message credibility mediated the relationship between the flagging of false news and false news sharing intentions" [36]. Furthermore, they found a third-person effect regarding the sharing intention.

Garrett and Poulsen [21] investigated different types of warnings in a study with 218 participants who were randomly assigned to four study groups with varying warning types: a peer-generated warning coming from other users, a warning coming from fact-checkers, and a warning saying the post was "coming from a site that characterized itself as a source of humor, parody, or hoaxes" [21]. Among others, they measured the participants' belief in the claim's accuracy, their sharing intention, and their reactance. They only found significant results for the self-identified humor flags and concluded — contrary to the results of Clayton et al. [11], Mena [36], Pennycook and Rand [43] — that "self-identified humor flagging is the only approach of the three tested to improve belief accuracy" [21].

2.3 Research Gap

Existing research has highlighted the relevance and impact of fake news. Multiple papers have shown that many people have encountered fake news stories [2, 18, 49]. Those who fall for fake news often share particular characteristics, such as low engagement in analytical thinking [9, 43]. Reuter et al. [49] revealed a generally broad agreement on counteractions on fake news. However, the opinion on specific strategies has not been subject to any scientific research so far. The social media users' acceptance and preferences regarding potential countermeasures remain still unclear (research gap 1).

Furthermore, existing research came up with a multitude of approaches to detect fake news stories [8, 23, 24, 30, 37, 45, 54, 56–58, 60, 63, 65]. Using warnings as well as correcting related articles has proven to be effective in reducing belief in misinformation [6, 7, 11, 16, 29, 41]. However, those studies refer to US participants and contexts, mainly the 2016 presidential election. The generalizability and applicability of their results to other contexts need to be confirmed. This is also true for the topics of false headlines, as Mena [36] suggests future work to test countermeasures "using topics related to other areas beyond international politics". Research papers have already examined some approaches to counter fake news in social media individually. However, there has not been any comparison of different approaches beyond warnings, and research on possible combinations of such approaches remains missing (research gap 2). That is why we put this work's focus on the users' perspective and comparison of possible approaches.

3 RESEARCH APPROACH

In this work, our objective is to learn more about approaches to counter misinformation in social media with a focus on the platforms' users, and thus, to examine its value to support the work between individuals. Our research questions are how such approaches should look like regarding the users' needs and how effective they are.

We address the research gaps in three ways:

(1) First of all, we aim to fill the gap of users' preferences and opinions by conducting a survey representative with regard to gender, age, education and income, specifically investigating warnings as a countermeasure in more detail (Section 4).
(2) Furthermore, since there seems to be a general backing of approaches like warnings, we aim to examine possible alternatives in more detail. Conducting semi-structured interviews, we strive to gain deeper qualitative insights by presenting five possible elements of such an approach to the participants and asking about their opinions. This step reveals that warnings appear to be the most popular approach among users (Section 5).
(3) Last, based on the results of those interviews and building on existing work [6, 7, 11, 16, 41], we select four approaches, which we then compare in a representative online experiment regarding their effectiveness and user preference. With these steps, we aim to find user-friendly solutions to counter misinformation in social networks effectively (Section 6).

By focusing our studies on the German population, we extend the context of existing studies and provide a basis for cross-country comparisons in future work. We conducted all three studies in Germany between July and November 2019.

Based on our research approach, we present the first part of the study on handling claims and warnings of misinformation in Section 4. Then, we describe Step 2, where we conduct 15 semi-structured interviews on various approaches (Section 5) and Step 3, where we compare some approaches based on their effectiveness (Section 6). Finally, we discuss the implications and limitations of our findings (Section 7) and draw a conclusion (Section 8).

4 STEP 1: SURVEY ON CRITICAL ENGAGEMENT TO SOCIAL MEDIA CONTENT AND WARNINGS OF MISINFORMATION

4.1 Study Design: Representative Online Survey

In the first step, we conducted an online survey representative with regard to gender, age, education and income. We asked a total of 1,012 participants from Germany two sets of questions: The first set was about their personal strategy to handle possibly false posts on social media. In the second set of questions, we asked them about their opinion on tagging fake news with a warning. The 14 questions are listed in Table 1. Those parts contained subscales of the General Decision Making Style (GDMS) and Rationale-Experiential Inventory (REI) questionnaires [40, 55], which are also interesting in the context of fake news. The GDMS is a psychometric questionnaire that measures how subjects are making decisions and comprises the following five subscales: rational, avoidant, dependent, intuitive, and spontaneous [55]. We used the rational subscale, which indicates logical and systematic decision making and verification of the information source and the dependent subscale, which shows how much a person prefers to rely on advice for making decisions [55]. The REI captures rational and experiential thinking styles [40]. We used the rational ability subscale and the rational engagement subscale, which measure the participants' ability to think logically and analytically and their enjoyment to do so, respectively. We combined those four subscales to a total score reaching from 65 to 150, which we interpret as an estimation of the level of analytical thinking [40].

We questioned a total of 1,012 participants using the panel provider Respondi. The majority of Respondi's panel lists are recruited through their affiliate network. In this way, they ensure the representativeness of the panels. Participation is voluntary and can be ended by both parties at any time. All participants are rewarded with a small allowance of about €1 for taking part and they receive bonus points for their answers. In the master data questionnaire, completion of which is obligatory on registration, around 40 basic and further 500 pieces of profile data are available on each panel list. In the sampling process, Respondi draws a random sample from the population

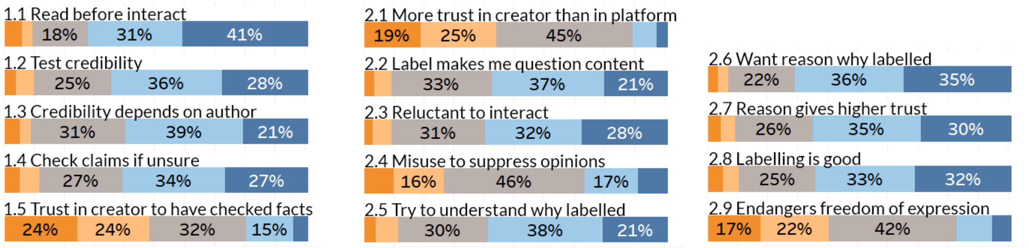
No.	Question	Mean	SD
1.1	I first read the linked article before interacting with it (like, comment, share).	3.98	1.13
1.2	I check the credibility of the post, for example by trying to understand its arguments.	3.79	1.04
1.3	Depending on who the post comes from, I consider it more or less credible.	3.67	1.02
1.4	If I am not sure whether statements made in the article are true, I try to verify them, for example by searching the internet.	3.72	1.09
1.5	I trust that the authors of the post have examined the claims made in it.	2.53	1.16
2.1	I trust the creators of the post more than the labelling of the platform.	2.53	1.00
2.2	The labelling causes me to question the content of the article.	3.66	0.98
2.3	I hesitate to interact with the post (like, comment, share).	3.76	1.03
2.4	I believe that labelling is misused to suppress uncomfortable opinions.	3.01	1.06
2.5	I try to understand why the post is labeled.	3.65	1.02
2.6	I would like a reason to be given why the post is marked.	3.96	1.01
2.7	I would rather trust the label if a reason was given.	3.82	1.05
2.8	I think it is good to label posts with false information.	3.81	1.10
2.9	Labeling of posts endangers the freedom of expression.	2.67	1.09

Table 1. The 14 survey items and their means and standard deviations. The first set of questions targets the participants' personal strategies on social media. The second set asks for their opinion on false information getting labeled by the platform.

of the online access panel based on socio-demographic data. Filter criteria can be master data, performance criteria, field data, and project data. Groups defined in this way are then ready to be used as samples for the project. In the course of drawing the sample, they also use a stratified or quota module. This allows them to set exact quotas for any number of master data criteria according to absolute or relative values. The participants for this study are representative of the German adult (18 to 74 years) population in terms of gender, age, education, and income. Our sample's composition shows only a few differences to the proportions of the German federal states. There are 48.7% female participants, whereas 51.3% were male. We gathered age in the following groups: 18 to 29 years, 30 to 39 years, 40 to 49 years, 50 to 59 years, and 60 to 74 years. Since participants of the study decide themselves whether to participate or not, the results are representative only with regard to the criteria mentioned above and basically correspond to a quota sample. We asked our participants if their monthly net income was lower than 2,000€, between 2,000€ and 4,000€, or above 4,000€. Using the χ^2-test, we tested representativity, which showed no significant differences to the German population regarding gender ($\chi^2(df = 1) = 0.32, p = .57$), age ($\chi^2(df = 4) = 3.54, p = .47$), income ($\chi^2(df = 2) = 3.35, p = .19$), and education ($\chi^2(df = 2) = 1.22, p = .54$). The results of the subscales of the GDMS and REI questionnaires show a median score of 106. The 25-percentile is at 97, and the 75-percentile is at 116; hence, half of the participants have a score between 97 and 116.

4.2 Results I: Critical Engagement to Social Media Content and Warning of Misinformation

The majority of participants report somewhat cautious handling of possibly false posts (see Figure 1a and Table 1). About three out of four (72%) state that they read the article before interacting with a post. Two out of three (65%) participants would check the article's credibility. While these numbers



(a) Questions on critical engagement to social media content

(b) Questions regarding labels/warnings to misinformation

Fig. 1. Stacked bar representation of the 14 survey items. Participants responded on a 5-level Likert scale: (1) Strongly disagree (2) Rather disagree (3) Undecided (4) Rather agree (5) Strongly agree

do not come across as very unexpected, they show that fake news is not affecting everybody but targets a rather specific subset of people which averages one-third of the participants: 28% of participants in a survey study do not read the article before interacting with a post and 35% percent of participants would not check the article's credibility. Consequently, it is interesting how many people responded contrary to the broad majority. There are 20% who state they trusted the author of the post to have verified the claims made in it. One in ten participants (10%) would not read the article before interacting with a post. Also, 10% would not check the post's credibility, and 12% would not search the internet if having doubts about its credibility. However, the results must be considered against the background of a possible social desirability of bias, which will be discussed later in the limitations.

A majority of participants state a warning would make them question the post's content and hesitate before interacting with the post (see Figure 1b and Table 1). Tagging fake posts with a warning seems to be widely accepted, as almost two in three participants state their endorsement. Even more participants wish that the platform explained why it labeled the post as fake. About one in eight participants state that they trust the post author more than the platform's label, whereas almost every other participant was undecided on this question.

4.3 Results II: Links to Gender, Age, Income, and Analytical Thinking

There is no significant difference between the responses of male and female participants. Female participants showed slightly stronger agreement on many of the items. We found that 38% of the female participants strongly agree that they wish to have an explanation of why the platform marked a post as fake. In comparison, 33% of male participants responded in the same way. Age appears to have a stronger influence on the responses. The group of 60 to 74 years old participants tended to agree more on many items. In this group, two out of three agreed that a warning would make them scrutinize the post's content while the other age groups vary between 52% and 56%. Similarly, a higher income correlates with more agreement on those items. While 66% of participants with a net income above 4,000€ state they would hesitate to interact with a marked post, just 55% of participants with an income below 2,000€ would do so.

We found a moderate but significant correlation between the analytical thinking score and the item on checking the post's veracity, i.e., by trying to understand its contentions ($r = 0.37$, $p < .0001$). While 78% of the participants with an analytical thinking score above the median agree on this item, only 52% of the participants below the median do so. Participants with a higher value in analytical thinking are more likely to scrutinize the post's content when given a warning

($r = 0.29$, $p < .0001$) and try to understand why it was marked ($r = 0.33$, $p < .0001$). We found that 70% of the participants above the median would scrutinize the post's content, while 46% of the participants below the median would do likewise. In other words, those above the median more likely desire to get an explanation of why the post was marked ($r = 0.26$, $p < .0001$, 82% agree above the median, 50% below).

4.4 Summary of Step 1

By conducting a survey in Germany, we found a broad majority who claimed to be critical with content in social media and discovered that around 10% do not check the post's credibility. These results seem to be linked to the level of analytical thinking. Our findings agree with previous work, which found that only a small group of people tend to be vulnerable to fake news, connecting them to less analytical thinking [9, 25, 29, 43]. We further found broad acceptance of warnings as an approach to counter misinformation in social media, coinciding with previous work [49]. Users also seek for an explanation of these warnings.

In this step, we showed that users widely welcome and accept measures to counter fake news. However, we still need to know how to implement such measures.

5 STEP 2: QUALITATIVELY COMPARING POTENTIAL APPROACHES TO COUNTER FAKE NEWS

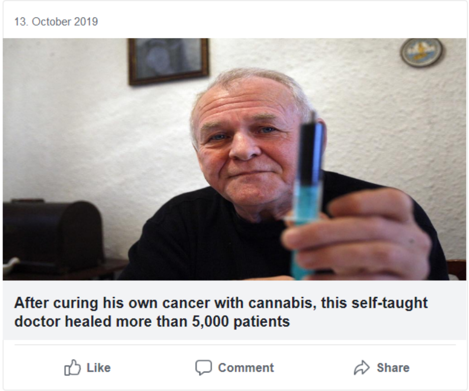
5.1 Study Design: Semi-structured Interviews

Based on the survey's findings that there is a demand for countermeasures, the second step, looked more detailed into their implementations. Several approaches were developed and evaluated in a qualitative interview study. We conducted an interview study to gain qualitative insights into the users' perspectives on various methods to counter fake news in social media. In Step 1, we found an overall agreement on the use of warnings in social networks. In this approach, we asked how such approaches should look to suit the users' needs. Furthermore, we extend our focus to include other approaches besides warnings into the comparison. Based on the corresponding work of Bhuiyan et al. [6] we developed an approach concerning the coverage of false articles. Bode and Vraga [7] developed an approach to show related articles section below the post, and Ennals et al. [16] refers to a "Disputed claim warning". Based on this corresponding literature we come up with five elements (see Figure 2). Those approaches follow different strategies to counter fake social media posts:

(1) Reducing the size of the article's photo and headline significantly,
(2) a warning message stating that independent fact-checkers had disputed the article and a link for more information [16],
(3) related articles underneath the post with headlines contradictory to the false claim [7],
(4) covering the article, as known from explicit contents [6], and
(5) requiring confirmation before interacting with the post.

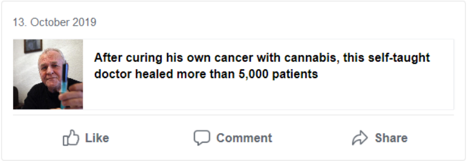
Most of these elements have or had been applied by Facebook to counter misinformation. Only the obstructing of a post is not used in the context of fake news but to protect users from violent and explicit content. Facebook uses some of the elements in combination, but we separated them for a more precise analysis.

We carried out a total of 15 semi-structured interviews by recruiting students and others who volunteered to participate. The interview-length was on average about 20-25 minutes. The group comprises eight female and seven male participants who were between 18 and 39 years old. The sampling strategy has to be seen in the context of the three-step structure of this study. Whereas the quantitative approaches of steps 1 and 3 are based on a broader demography and fulfill the


13. October 2019

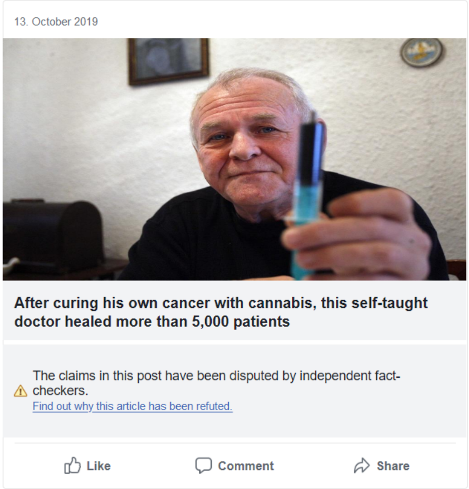
After curing his own cancer with cannabis, this self-taught doctor healed more than 5,000 patients

(a) Original post presenting a false news article


13. October 2019

After curing his own cancer with cannabis, this self-taught doctor healed more than 5,000 patients

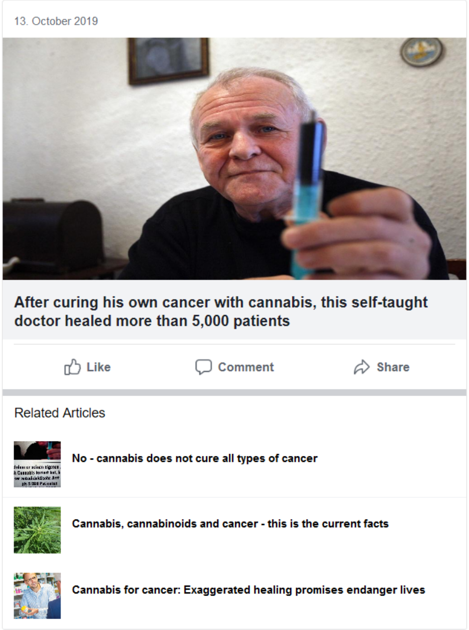
(b) Reduced post size


13. October 2019

After curing his own cancer with cannabis, this self-taught doctor healed more than 5,000 patients

⚠ The claims in this post have been disputed by independent fact-checkers.
Find out why this article has been refuted.

(c) A warning element underneath the false article [16]


13. October 2019

After curing his own cancer with cannabis, this self-taught doctor healed more than 5,000 patients

Related Articles

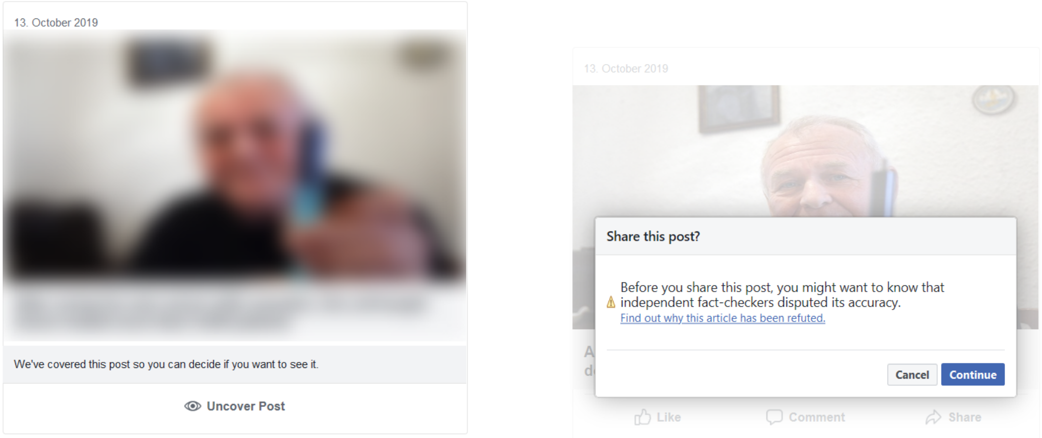
No - cannabis does not cure all types of cancer

Cannabis, cannabinoids and cancer - this is the current facts

Cannabis for cancer: Exaggerated healing promises endanger lives

(d) A related articles section below the post showing several articles with contradicting headlines [7]

representativeness according to the outlined criteria, the participants in this step were selected from the interviewer's direct environment and thus, no systematic sampling has taken place. It was primarily a matter of differentiating the individual factors, which were then, subsequently, quantitatively validated. Thus, step 2 was subsequently confirmed in step three with a broader and representative demography. All of the participants had experience with social media platforms: everybody stated that they use either Facebook, Instagram, or Twitter at least several times a month. Only one participant did not use Facebook at all; however, he stated that he used Twitter several times a week. In total, 11 participants stated that they use Facebook at least several times a month.



We've covered this post so you can decide if you want to see it.

Share this post?

⚠ Before you share this post, you might want to know that independent fact-checkers disputed its accuracy.
Find out why this article has been refuted.

(e) Covering the false article [6]

(f) A confirmation alert before interacting with the post

Fig. 2. A Facebook post with a fake news headline (2a) combined with the five approaches that we examined in the the interviews (2b – 2f). We used German graphics in the interviews.

Instagram is used by ten participants at least several times a week, whereas only four participants stated that they use Twitter several times a month or more. All participants use YouTube, at least every month.

The same interviewer conducted all interviews. After a short introduction on the topic of the study and filling the demographic questionnaire, we showed the participants an example of a social media post containing a false news headline. This post was formatted based on Facebook's post design. We told the participants that the platform would engage independent fact-checkers who would find that the shown article was false news. We also noted that, in this interview, we aimed for approaches that would somehow alter the post without removing or filtering it completely. In the following, we subsequently showed them the five elements presented in Figure 2.

On each element, we asked the participants what they found good and bad about it, and if they had ideas to improve it. Furthermore, we asked them about their opinion on the suitability and effectiveness of this element, their acceptance as users of a platform applying it, and how encountering it in their news feed would make them feel and react. Furthermore, we asked if they had ideas other than the presented approaches and which of these they liked the most and the least. While we had a prepared set of questions, we gave enough space for topics beyond the scope of these questions. Afterwards, the interviews were transcribed. By using qualitative content analysis according to Mayring [34] we developed a category system. The categories included: positive aspects, negative aspects, considerations, suggestions for improvement, suitability of the tools, acceptance, experience. By using previously defined categories, the gathered data was then analyzed.

5.2 Results

5.2.1 The Most and Least Preferred Elements. The participants liked those approaches which transparently informed about the false information. All participants favored a warning message under the post. Also, 13 out of 15 liked the confirmation before sharing or otherwise interacting with the post. The other elements were less popular among the participants. Six participants thought to reduce the size was a good choice, while just as many liked covering fake news posts. Only three participants favored related articles with contradictory headlines. When asked which approaches they liked the least, eight participants named the covered post. Seven did not like the reduced size, and six participants said they did not like the related articles. No participant stated that he or she did not like the warning or confirmation elements. Additionally, it has to be mentioned that each participant could select several countermeasures at a time and their choice was not limited to one.

5.2.2 Warning. Regarding the warning, participants liked that it was immediately apparent that the article was fake news. They also liked the link providing more information and explanations why the article had been marked. In this context, one participant stated *"You can really see at first glance that there are some false facts in it. Especially through the link, through which you can really see what's wrong with it."* (I01). Furthermore, some participants mentioned the element's transparency, honesty, and self-responsibility. However, many participants stated that the warning was too unobtrusive. As one participant noted: *"The correction is smaller than the lie"* (I05). Moreover, participants stated that *"the warnings were not noticed directly"* (I07, I09) and were *"unobtrusive"* (I12, I13). Furthermore, it was negatively noted that it *"is not immediately obvious" that it refers to a specific post"* (I11). Another participant outlined that the warning should stress *"a bit more conspicuous that this is actually part of the contribution and also who these fact-checkers are, so that it is even more credible."* (I09). Some participants noted that *"it is very important to have some degree of trust in the fact-checkers"* (I11). Especially in the context of polarization, where *"people end up in echo chambers and believe the other side is lying to them"* (I06). Almost all interviewees noted that the warnings should be made even more obvious and striking.

5.2.3 Confirmation. Besides the warning, a confirmation dialog before sharing (or possibly liking and commenting) was also popular among most participants. They liked that false information was not spreadable in one click anymore, giving a moment of reflection, and increasing the inhibition level. For example, it is mentioned that by *"directly conveying a feeling that something is wrong with the post that you should not see" an inhibition threshold arises and you consider "whether you really click on it"* (I13). Many participants believed it to be a suitable approach to curb the spread of fake news. On the other hand, they criticized that — unless combined with other approaches — the warning would only become visible to users who interact with the post. Among other things, it has been criticized *"that it is not at all recognizable whether it is fake news or not, but one has to interact first"* (I02) and that without a warning *"it would not stop the idea affiliating "* (I06).

5.2.4 Reduced Size. Reducing the post's size was seen more critically by some participants. Some mentioned the missing transparency; others stated it was, e.g., *"a bit more extreme response, maybe a bit too far"* (I05) and *"coming close to censoring content"* (I05). However, many participants also found it favorable that it would arouse less curiosity, interest, and attention. One participant made it clear that the tool illustrates fake news much less eye-catching. Since *"it is often the picture that makes you read the headline in the first place [...]"*, the article would *"perhaps be less interesting and less weight"* (I07). Some also mentioned that with this approach, fake news would take up less space in the user's feed: *"If it is spam then it is great that it uses up less space."* (I06). In general, the participants do not consider the approach to be so suitable without placing an additional warning, while a warning is considered to be more effective.

5.2.5 Covered Content. The participants liked covering the post in that it adds another obstacle before viewing fake news, and it needs curiosity and the user's action before being able to see the post. Among other things, it is positively mentioned that *"you have to do something actively to even read the headline"* (I04) and that it*"does not distract"* (I07). Thus, false information would not stick in one's memory. On the other hand, this approach might arouse even more attention and curiosity. One respondent noted that *"if it is a covered item and it appears on my timeline, it would seem a bit bizarre because it will make me more curious and hungry for information."* (I11). Moreover, the covered content could lead to incorrect conclusions, since *"it [the post] looks like [it] is just for adults. [...] Hence, it is not apparent at first glance that it is fake news"* (I07). It gives no hint on what topic the post is about, making it hard to come to an informed decision if one wants to uncover the post or not. Overall, the participants were divided into those who stated this approach would make them lose interest and scroll past the fake news post and those who would be very curious and uncover the post.

5.2.6 Related Articles. Regarding the fact-checks in related articles, participants liked that it allows them to do research on their own and *"putting [them] in control of what [they] want to choose to believe"* (I06). One participant positively points out that *"one can see that the assertion is not confirmed in any of the related articles"* (I04). Against this backdrop, one interviewee outlined that the related articles make it clear *"that there are more ways of thinking"* (I14). Its journalistic touch might encourage users to read the presented related articles. However, they also criticized that it would not tell them that the fake article was false, and it was not clear that the related articles were more credible than the main article. For instance, one of the participants is wondering who is behind the fact-check and remarks that *"this is subjective again"* (I08). Another one criticized that due to *"the fact that related articles are available [...] one rather says, okay, [this post] is not fake news, because there is more information about them"* (I02). Moreover, they noted that this element would take up lots of space in their news feed, which might become annoying. Furthermore, it might be overlooked and not read at all. Some participants suggested using a more descriptive title instead of *Related Articles*.

5.2.7 Suggestions for Further Approaches. Moreover, the participants were asked if they had suggestions for other approaches to counter fake news. Two participants suggested showing an extract of the fact checker's article or some counter-argument directly under the post. Some proposed to add some peer aspect to the warning since some users would probably rather listen to their peers. For example, when combined with a confirmation before sharing, the warning underneath the post could state that *"five of your friends decided not to share this article"* (I06). Some participants suggested more restrictive approaches, like removing the ability to share fake news posts or to block those posts completely from distributing. For example, it is proposed to circle the fake news *"again in red, i.e. work on the overall presentation. Removing it completely would also be an option, but it is always difficult with freedom of expression to delete it directly"* (I04). Several participants expressed their wish to have a custom setting to allow or disallow detected fake news posts to appear in their news feed.

5.3 Summary of Step 2

Our results reveal that most participants liked the warning and confirmation approaches. These are the two approaches which aim to inform users transparently about the false information. Interestingly, a warning — being the most popular option among the participants — had already been deployed by Facebook but was revoked later. Facebook's current approach is to use a reduced post size, related articles, and confirmation before sharing. However, the first and second elements of that approach were assessed rather critically by our participants. From a user's point of view, it is understandable to prefer transparent approaches providing them with an opportunity for informed decisions.

Our interview study exposed the potential conflict of interests between the users and the platform: Users asking for transparency while platforms adopt somewhat less transparent solutions.

6 STEP 3: WHICH APPROACHES ARE EFFECTIVE?

6.1 Study Design: Representative Online Experiment

In the second step, we found a clear tendency of users to prefer transparent approaches, like a warning. Countermeasures do not just need to be accepted and welcomed by the users. It is also crucial for those measures to be practical to counter fake news. Therefore, we used these outcomes to select and improve promising approaches and — as a third step — examined and compared their effectiveness. We also looked into the user acceptance and their preferences to extend the interview findings using a larger, representative sample.

Based on the findings in Step 2, we conducted an online experiment with 1,030 participants representative to the German adult population between 18 and 75 years. As outlined in the first step (4.1), the participants were selected by the panel provider Respondi. It draws a random sample from the population of online access panel based on socio-demographic data.

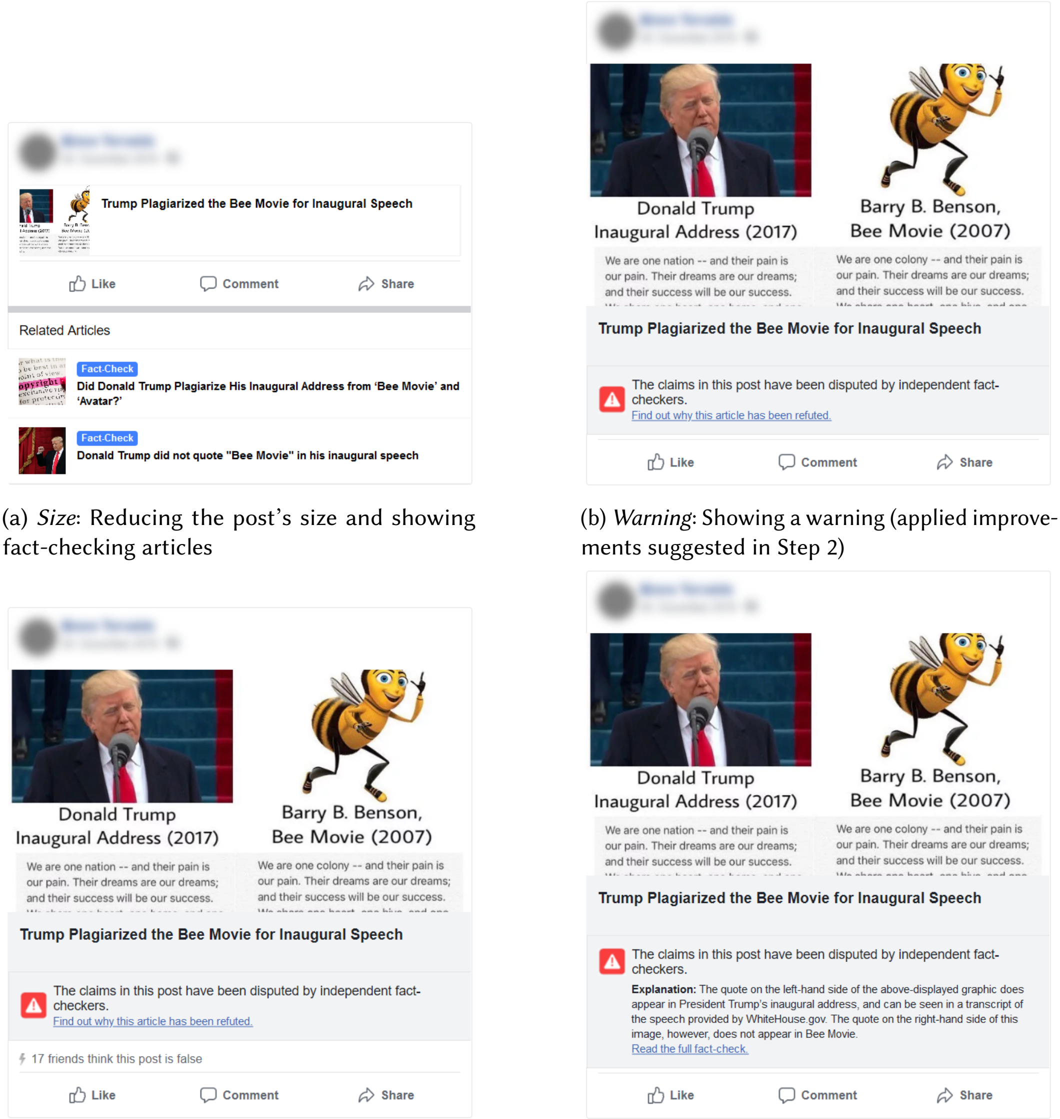
We compared four strategies to counter fake news in social media, as shown in Figure 3. From the five elements we examined in Step 2, we excluded the confirmation element and covered-content element for reasons of the method, since those approaches required user interaction. Being the most popular element in Step 2, we included the warning element as a condition (Figure 3b). Furthermore, we introduced two of the suggested approaches from Step 2 in combination with the warning element: Displaying how many friends believe the article is false (Figure 3c) and showing an explanation right beneath the warning (Figure 3d). To apply this peer element to our study and to get a realistic feeling of this approach, we used random numbers between eight and 19 assigned to the different headlines.

We included the combined elements of Facebook's current approach (except for the confirmation): reducing the post's size and showing fact-checking articles in the related articles section (Figure 3a). In the following, we refer to these four conditions as *Size*, *Warning*, *Peer*, and *Explanation* approach. The limitation to these four conditions was necessary since more would have gone beyond the scope of the paper or beyond receiving strong empirical results. Based on the results of step 2, the four selected terms were identified as the most interesting and most feasible. We veiled the post's author name and profile picture, as well as the date to minimize the influence of confounding factors.

Our study design builds on Pennycook et al. [41] and Clayton et al. [11]. Just like them, we used a between-subject experimental design. While Pennycook et al. [41] compared only one study group to a control group, and Clayton et al. [11] used a more complex 2 × 3 study design, we compared four study groups with different approaches and a control group to which no approach was shown. All participants were shown a total of 24 news articles in the format of a Facebook post. We selected 16 disputed article headlines from various fact-checkers (correctiv.org, a German investigative journalist group, Mimikama, an Austrian fact-checking site, and Snopes, a US site).

Furthermore, we selected eight accurate headlines from mainstream media or headlines verified by fact-checkers. While the number and composition of headlines differ from Pennycook et al. [41] and Clayton et al. [11], the proceeding is overall consistent. In the study groups, we presented half of the false articles using the corresponding approach; and displayed the others as regular Facebook posts — just like the accurate headlines. We asked participants to rate the perceived accuracy of those headlines using a 4-point Likert-scale with the following options: *false*, *rather false*, *rather true*, and *true*. Again, this is similar to previous work.

However, we also deviate in certain aspects. Unlike the previous work, we obtained a representative sample concerning age, education, and gender, based on the criteria of the panel provider. While the American political landscape allowed for a simple binary political composition of headlines (pro-Trump vs. anti-Trump or republican vs. democratic), we were not able to transfer this to the



Trump Plagiarized the Bee Movie for Inaugural Speech

Related Articles

Fact-Check

Did Donald Trump Plagiarize His Inaugural Address from 'Bee Movie' and 'Avatar?'

Fact-Check

Donald Trump did not quote "Bee Movie" in his inaugural speech

Donald Trump Inaugural Address (2017)

We are one nation -- and their pain is our pain. Their dreams are our dreams; and their success will be our success.

Barry B. Benson, Bee Movie (2007)

We are one colony -- and their pain is our pain. Their dreams are our dreams; and their success will be our success.

Trump Plagiarized the Bee Movie for Inaugural Speech

The claims in this post have been disputed by independent fact-checkers.

Find out why this article has been refuted.

17 friends think this post is false

Explanation: The quote on the left-hand side of the above-displayed graphic does appear in President Trump's inaugural address, and can be seen in a transcript of the speech provided by WhiteHouse.gov. The quote on the right-hand side of this image, however, does not appear in Bee Movie.

Read the full fact-check.

(a) *Size*: Reducing the post's size and showing fact-checking articles

(b) *Warning*: Showing a warning (applied improvements suggested in Step 2)

(c) *Peer*: Showing how many friends believe the article is false, in addition to a warning

(d) *Explanation*: Giving an explanation on why the post was marked with a warning

Fig. 3. The four approaches tested and compared in the third step. In the online experiment, we used German versions of the graphics.

composition of German fake news headlines. It is harder to divide the German political landscape into a few distinct categories because there are various political parties. Also, we could only find a limited number of left-wing-motivated headlines. Therefore, we omitted such a strict political categorization while still trying to collect a balanced set of headlines (see a list of all headlines in Table 2).

#	Headline	Accuracy	Source
1	Trump Plagiarized the Bee Movie for Inaugural Speech	false	Snopes
2	Man arrested at Berlin Airport with over 10 million euros in cash	false	Correctiv
3	The Icelandic government pays $ 5,000 for every man who marries an Icelandic woman	false	Correctiv
4	Terrifying study: cancer patients die faster with chemotherapy than without treatment	false	Correctiv
5	France passes law saying that children can consent to sex with adults	false	Correctiv
6	Black Friday: The term goes back to the slave trade	false	Correctiv
7	Burglary crime has increased massively in NRW — the clearance rate has stagnated	false	Correctiv
8	Elon Musk leaves Tesla and switches to financial technology	false	Correctiv
9	Trump plans to ban TV shows that promote homosexuality	false	Snopes
10	After curing his own cancer with cannabis, this self-taught doctor healed more than 5,000 patients	false	Correctiv
11	Greens: The liter of gasoline should cost at least 6–7 euros!	false	Correctiv
12	Pentagon wants to have a German warship with them off the Syrian coast	false	Correctiv
13	Croatia donates entire World Cup awards — tough letter to politicians	false	Correctiv
14	Merkel: “Freedom of expression needs strict limits”	false	Correctiv
15	CDU demands pork duty in canteens	false	Mimikama
16	Greens: heating the apartment to 15 degrees is enough	false	Correctiv
17	Hurricane killed and injured in Germany	true	Zeit
18	More asylum seekers live in NRW than in the whole of Italy	true	Correctiv
19	Criminologist: Refugees are reported more often than Germans	true	FAZ
20	Germany earns billions with Greece loans	true	Tagesspiegel
21	Germany pays significantly more child benefit abroad	true	Zeit
22	Palmer demands that violent migrants’ freedom of movement be restricted	true	Welt
23	Germany must bring back IS members	true	SZ
24	AfD subsidizes demo participation	true	Correctiv

Table 2. The 24 headlines presented to the participants. Sixteen of which are false and eight are true headlines. While we used German headlines in the study, this table shows their English translation for better intelligibility.

Before presenting the news headlines, participants had to fill demographic data and further questions. Besides demographic data, we asked the participants about their political attitude and their media behavior. Furthermore, as done by Pennycook et al. [41], we let them complete the Cognitive Reflection Test (CRT) [19]: a short test comprising three trick questions that intuitively elicit wrong answers. We can then take the number of correct answers as a measure for analytical thinking, similar to the GDMS and REI questionnaire subscales in Step 1. We used the CRT here instead of GDMS/REI to gain a better comparability to the study of Pennycook et al. [41]. Since analytical thinking is not a focal issue in this Step, a shorter questionnaire is also more suitable. Following the headlines phase, we further asked about their acceptance of the corresponding strategy to counter fake news. Eventually, we showed all participants the four examined approaches and asked them to order them by preference. This study was conducted in a joint survey in which other research questions preceded and followed our part of the study.

A total of 1,030 respondents participated in our experiment. The five groups contain between 193 and 215 subjects each. Our sample consists of 502 male and 528 female participants between 18 and

Group	N	Mean	SD
Control	213	-0.33	0.96
Size	193	-0.38	0.95
Warning	197	-0.44	0.99
Peer	215	-0.42	0.98
Explanation	212	-0.50	0.96

Table 3. Overview of the five groups and their descriptives of the estimated accuracy.

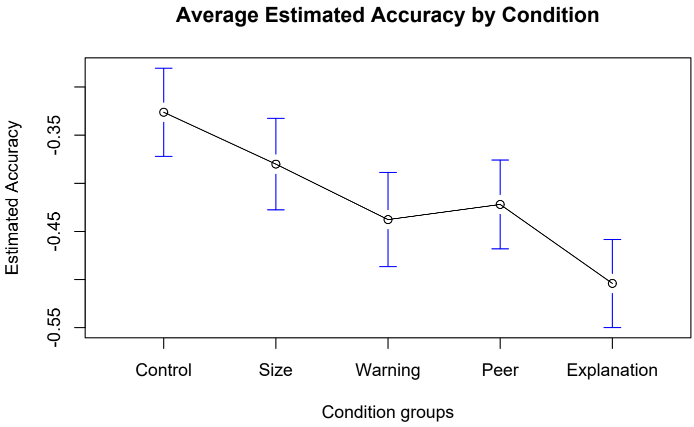


Fig. 4. Estimated accuracy means and their 95% confidence intervals by the five condition groups.

74 years with a mean age of 47 years. It is representative of the German population regarding gender ($\chi^2(df = 1) = 0.29, p = .59$), age ($\chi^2(df = 4) = 0.007, p = 1$), education ($\chi^2(df = 2) = 1.29, p = .52$), and federal state ($\chi^2(df = 15) = 8.16, p = .92$).

We found 80% of the participants stating to have used Facebook, Twitter, or Instagram before. Around two out of three use one of those platforms, at least every week. There are 39% of participants who report sharing political content on social media.

6.2 Results I: Effectiveness

In our analysis, we compared the responses to the eight headlines that were altered by each condition. We mapped the 4-point Likert-scale responses on an interval scale centered to zero with the following four steps: "false" (-1.5), "rather false" (-0.5), "rather true" (0.5), and "true" (1.5). A first look on the average estimated accuracy of the headlines shows that all four study groups have a lower average than the control group meaning their headlines were rated less accurate (see Table 3 and Figure 4).

To test if those differences are significant, we conducted an Analysis of Variances (ANOVA), including headlines and participants fixed effects. The ANOVA revealed highly significant group differences ($F(4, 7196) = 11.1, p < .0001$ ***). To find out which conditions show significant differences, we used Tukey's HSD for post hoc analysis. Here, we found three of the four conditions to differ significantly from the control condition. The results of the *Size* condition were not significant ($p = .33$). The *Warning* group ($p = .0009$ ***), the *Peer* group ($p = .006$ **), and the *Explanation*

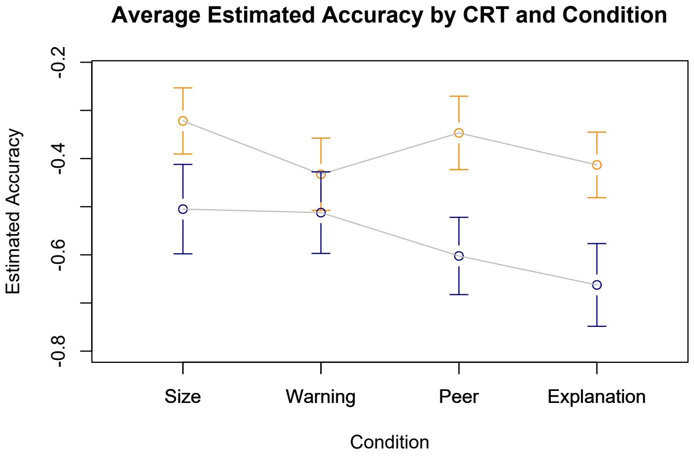


Fig. 5. Means and 95% confidence intervals of estimated accuracy. Comparison of the conditions between respondents with low and high CRT scores. The upper orange graph represents mean accuracy for respondents with a CRT score of 0; the lower blue graph represents respondents with a CRT score of 2 or 3.

group ($p < .0001$ ***) — which all three comprised a warning message — were highly significantly different from the control group. Furthermore, the *Explanation* group differed significantly from the *Size* group ($p = .0002$ ***) and the *Peer* group ($p = .03$ *). While the *Explanation* approach seems to increase this effect of a simple *Warning* (but is not significant: $p = .14$), the *Peer* aspect does not deviate much from the *Warning* condition ($p = .98$).

Moreover, we looked at possible spillover effects on other unaltered headlines. Pennycook et al. [41] report the so-called *Implied Truth Effect*, meaning a higher perceived accuracy of false headlines without a warning when some other headlines are labeled with a warning. To check for such spillover effects, we conducted another ANOVA — this time on the eight unaltered false headlines. Its results show no significant differences between these groups ($F(4, 7196) = 0.6, p = .66$). These findings align with the results of Clayton et al. [11], who was also not able to replicate such an *Implied Truth Effect* — possibly explainable by a lack of precision in our data.

6.2.1 Differences by Cognitive Reflection Test. When we look at the results of the CRT, we find the mean estimated accuracy of respondents with low CRT (CRT score = 0, $M = -0.38, SD = 0.99$) and high CRT (CRT score > 1, $M = -0.57, SD = 0.95$) to be significantly different using an independent sample t-test, $t(4790) = 6.79, p < .0001, d = 0.20$. So, a higher CRT score is linked to lower estimated accuracy of false news headlines with one of the approaches applied (see Figure 5).

Previous work found a low CRT to indicate a lack of analytical thinking and connected it to susceptibility to fake news [43]. Hence, it would be of interest to find the most effective solutions to this particular target group. However, as Figure 5 shows, the differences between the four approaches are even smaller in the group with low CRT. So, we did not find one or multiple approaches that are especially effective for this group of participants.

6.3 Results II: User Acceptance and Preference

The agreement to the use was very similar among the four approaches (see Figure 6). The amount of strong agreement for the approaches lies between 23% (*Warning*) and 31% (*Explanation*). About

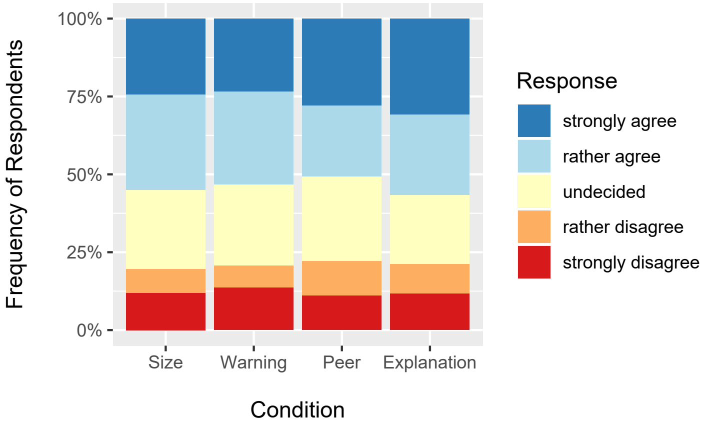


Fig. 6. Respondents' agreement to the use of the four approaches on their social media platform.

the same number of respondents rather agreed to the use of the respective approach (23%–31%). Accordingly, the overall agreement varies between 51% (*Peer*) and 57% (*Explanation*). About 20% to 22% disagreed with the use of the approaches.

In the next question, we presented all four approaches to the participants and asked them to rank them by favorability. Although users had more experience with their condition's approach compared to the others, there are no significant differences in the ranking between the study conditions (Size approach: $\chi^2(df = 12) = 15.81, p = .20$, Warning approach: $\chi^2(df = 12) = 20.65, p = .06$, Peer approach: $\chi^2(df = 12) = 9.89, p = .63$, Explanation approach: $\chi^2(df = 12) = 10.48, p = .57$). In this ranking, we find stronger differences between the approaches: As Figure 7 shows, the *Explanation* approach was ranked higher than the other approaches and has a mean rank of $M = 1.66$, followed by the *Warning* approach with a mean rank of $M = 2.46$. More than 50% of all participants ranked the *Explanation* approach in the first position. The *Size* ($M = 2.94$) and *Peer* ($M = 2.90$) approaches are similarly less favored. In a nutshell, both approaches were used: At first, we asked the participants how to evaluate the specific approach shown, and then, afterward, the other approaches were presented and ranked by them. This analysis reveals that a transparent approach that explains to the users is not only effective but is also favored by participants.

6.4 Summary of Step 3

We compared four different approaches in a between-subject online experiment. One of them was based on Facebook's current approach to counter fake news by reducing the post's *Size* and showing fact-checking articles beneath the post. The other three were similar warning-based approaches: a simple *Warning*, a warning extended by a *Peer* aspect, or extended by a short *Explanation*. We found those three warning-based approaches to have significant effects on the perceived accuracy of false news headlines. The *Explanation* approach had the lowest mean value and was shown to be significantly more effective in reducing the perceived accuracy than most other conditions. While the *Size* condition also had a lower accuracy mean than the control group, this difference was not significant. The CRT was overall negatively linked to the perceived accuracy of false headlines. We did not find significant differences between the conditions when focusing on low CRT participants only. Furthermore, we found that a majority of participants agree to the use of the respective approach in their social network.

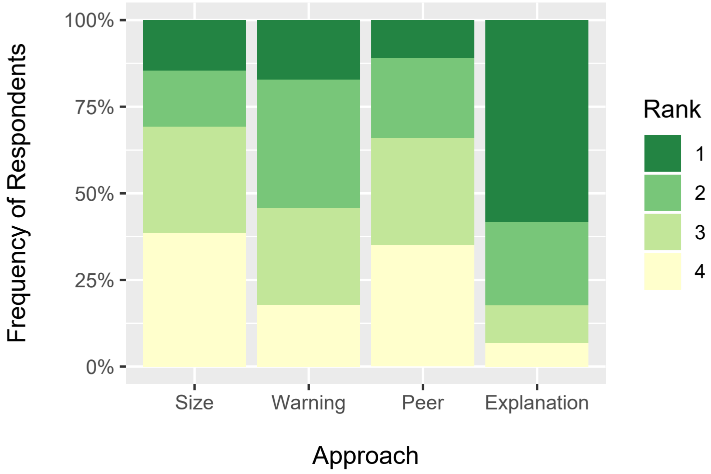


Fig. 7. Rank distribution of the four approaches.

While this question had little differences between the four study groups, it became clear that the *Explanation* approach was the most popular approach when we asked participants to compare them to each other.

7 DISCUSSION

Our two research questions were how measures to counter fake news in social networks should look like from the users' point of view, and how effective such approaches are. To answer these questions, we conducted a three-step study. As the first step, we ran an online survey to find out about the users' general handling of credibility assessment in social media and their agreement to fake news countermeasures such as warnings. In the second step, we aimed to understand more about how to design approaches to counter fake news in social media and conducted 15 qualitative interviews comparing five different approaches. In the third step, we compared promising approaches in an online experiment to gain representative knowledge about their acceptance and popularity and learn about their effectiveness to reduce the belief in fake news.

7.1 Discussion of Key Findings

We come up with the following key findings:

(1) **Users want social media platforms to warn them of fake news:** They do not want to see fake news in their news feed without being informed about it. In Step 1, we found two-third of German users welcoming a warning message to label fake news in their feed. In the second step, all interviewees favored a warning approach, and in the third step, the three warning-based approaches showed an acceptance rate between 51% and 57%. More generally, we can say that users prefer transparent over non-transparent approaches. Platforms and other parties developing countermeasures should consider how their approaches can fulfill this aspect of transparency.

(2) **Users want social media platforms to explain why some posts have a warning:** Users want to be able to understand what is inaccurate in the presented claims, and they want to make their own informed decision about the accuracy of the news article. Our results are even more striking regarding the explanation approach. In the first step, we found 71%

demanding platforms explain why they marked a post with a warning, and in the third step, a clear majority of participants ranked the *Explanation* approach in the first position. Again, this aspect is of importance to the development of countermeasures against fake news.

(3) **Users welcome Facebook's approach less than the other approaches; it is also less effective.** We covered only a part of Facebook's strategy and cannot make statements about their approach as a whole. Nevertheless, it became clear that the users were somewhat critical about reducing the post size and displaying fact-checking articles in the Related Articles section. In the second step interviews, six out of fifteen interviewees liked a reduced post size, and seven did not. Only three preferred fact-checks in the Related Articles section, while six did not like that approach. Furthermore, the results of Step 3 showed no significant effect of this approach on the estimated accuracy of headlines. Our results put this approach's suitability up for discussion.

(4) **Warning-based approaches can decrease belief in misinformation; it seems to be most effective when combined with an explanation.** In the third step, the three warning-based approaches had a significant effect on the estimated accuracy of false headlines. The *Explanation* approach was significantly more effective than most of the other approaches. Hence, our results suggest that social media platforms should examine those approaches in more depth and field tests.

7.2 Answering the Research Questions

Regarding our first research question (designs to counter fake news), all three studies showed that users prefer particularly transparent approaches to counter fake news. We repeatedly showed that users welcomed warnings and especially warnings combined with an explanation and favored them most, compared to alternatives. Users want to get the opportunity and information to make an informed decision about the contents in question. While they generally also welcome other approaches, warnings and explanations were preferred.

Concerning our second research question (effectiveness), we can say that the *Warning*, the *Peer*, and the *Explanation* approaches lead to a lower estimated accuracy of the corresponding headlines and, thus, are effective measures to reduce the belief in misinformation. Especially the *Explanation* approach showed the most significant effect, which, however, was not significantly higher than the *Warning* condition. While the *Size* approach also showed a lower average in responses, it was not significantly different from the control group.

In Step 1 and Step 3, we asked participants for their agreement of showing warnings under fake news posts and giving an explanation to it. In the first step, we found 65% agreement for labeling fake news, and 71% agreement for an explanation (see Figure 1b), while the results of the third step were slightly lower (between 51% and 57%). The concrete implementations of the approaches may explain the lower agreement rates compared to the abstract description in the first study. So when asked about their agreement to use the shown approach, they might have come up with more negative aspects than when just given a short description. Furthermore, some users might dislike some specific details of the presented implementation while still welcoming the overall concept of the approach.

7.3 Contribution and Implications

Our work builds on some previous papers that examined, e.g., warnings as an approach to counter fake news. In a similar experiment, Pennycook et al. [41] had revealed a warning effect as well as an implied truth effect regarding the perceived accuracy and willingness to share an article. While we confirmed a significant effect of warnings on the perceived accuracy, we did not find such an implied truth effect — just as Clayton et al. [11].

In their study, Clayton et al. [11] compared two differently worded warnings ("disputed" vs. "rated false"). Garrett and Poulsen [21] compared three types of warnings (peer-generated, fact-checker flags, and self-identified humor). We extended those research approaches and included potential solutions beyond warnings. Moreover, we did not only study the effects of those approaches on the perceived accuracy but also put the focus on the users' acceptance and preference.

When looking at the CRT score, we found them to be linked to the perceived accuracy, thereby confirming the results of Pennycook and Rand [43]. Similarly, Bode and Vraga [7] had found significant effects of putting fact-checks into the Related Articles section. In our work, we empirically examined a similar approach that also reduced the post size and that we based on one of Facebook's strategies. However, we could not find a significant effect of this approach, although its mean accuracy was still lower than the control group's mean. We possibly missed the necessary precision in our study. Nevertheless, Bode and Vraga [7] had also found an effect in only one of two studies. It remains unclear under which circumstances this approach is practical.

While our findings regarding the effect of warnings align with most previous work [11, 36, 41], it partially contradicts the results of Garrett and Poulsen [21] who concluded that fact-checker flags and peer-generated warnings would not affect belief accuracy. However, their sample size and number of trials are notably smaller compared to our study and their study data probably misses the necessary precision to find this comparably small effect of fact-checker warnings. Apart from that, we did not find peer information increasing the effect of a fact-checker warning. This finding aligns with the results of Garrett and Poulsen [21] and suggests that peer-based approaches might not be suited.

In our work, we emphasize the suitability of warnings as countermeasures to fake news in social media. Those approaches seem to be promising options for social media platforms. We suggest that these aspects should be taken into consideration during the development and implementation of new approaches and provide further research opportunities.

7.4 Limitations and Future Work

While we could provide valuable findings to the research field on countermeasures to fake news, our work comes with some limitations.

(1) In the context of the results of the data collection from all three surveys, it must be assumed that there is a certain social desirability bias that influences the results. For Step 1, it can be assumed that with the use of simple questionnaire items, there is a possibility of data bias due to the social desirability. One example is that three-quarters of participants claim to read news stories before sharing. Against this backdrop, it must be assumed that respondents give preference to answers that they believe are more likely to meet with social approval than the real answer, where they fear social rejection.

(2) A further limitation is the lack of representativeness of the qualitative Step 2. As outlined, the interviewees were all from the interviewer's environment and brought a certain level of education. Step 2 was conceived only as an intermediate step and was confirmed by Step 3 with a broader demographic group. Nevertheless, the lack of representativeness is a limitation. This should be confirmed again with a broader representative survey.

(3) It should also be noted that the selection of study participants for Step 1 and Step 3 was representative according to the selection criteria of the panel provider Respondi. Nevertheless, one restriction that must be taken into account is that all participants must have been registered with the Respondi panel to participate in the study. Thus, a pre-selection has taken place. Moreover, as the sample is representative only in terms of the above criteria, it is basically a quota sample.

(4) In our work, we aimed to eliminate the influence of the post author on the measured effect of perceived accuracy. However, in real-life situations, the post author (or sharer) — and especially one's trust in this source — probably play a pivotal role in one's opinion on a story. This aspect and notably its interaction with different approaches are worth more profound research in the future.
(5) In the third step, we observed diverging means for the effectiveness of the different study groups; however, those differences were statistically insignificant between some groups. To gain statistically significant insights, a larger sample size with higher precision would be necessary. Another possible reason why the *Size* approach had no significant effect might lie with the study's procedure. Our study design asked the participants to assess all headlines. This instruction deviates from the usual behavior in a social media feed where users often scroll past entries without looking at them. It is reasonable to assume that a reduced post size would facilitate overlooking disputed posts. However, our study did not capture this possible effect.
(6) As previous research has shown, it is a rather small and specific group of people susceptible to fake news in social media. We have seen a vast majority of people apply measures of credibility assessment on social media content. Although it is more relevant to consider this majority of users when building solutions to the problem of fake news, research should emphasize the group of vulnerable users as the primary target group. We attempted to examine this group by using the Cognitive Reflection Test but lacked the necessary precision in our data. Future work should put a stronger focus on this most relevant group.

8 CONCLUSION

In this work, we examined and compared measures to counter fake news in social media. Using a three-step study design, we took a closer look at their effectiveness and the user perspective. In summary, we found high acceptance of a warning approach, especially when combined with an explanation. Warning-based approaches were also effective in reducing the estimated accuracy of false news headlines. Our work provides relevant insights into the field of CSCW and social computing and in particular to the question of a collaborative character of fake news. It is one thing to detect fake news content quickly and reliably, but the question remains what to do with such posts. Social media platforms like Facebook have started to develop and deploy strategies for that. With this work, we contribute to a better understanding of those approaches that can help to improve existing efforts.

ACKNOWLEDGMENTS

This work has been funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY centre as well as by the German Federal Ministry of Education and Research and the Hessen State Ministry for Higher Education, Research and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE. We would like to thank Lilian Reichert for her valuable editorial support.

REFERENCES

[1] Gregor Aisch, Jon Huang, and Cecilia Kang. 2016. Dissecting the #PizzaGate Conspiracy Theories. *New York Times* (2016). https://www.nytimes.com/interactive/2016/12/10/business/media/pizzagate.html
[2] Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of economic perspectives* 31, 2 (2017), 211–236. https://www.aeaweb.org/articles?id=10.1257/jep.31.2.211
[3] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics* 6, 2 (2019), 8.

[4] Ahmer Arif, Leo Graiden Stewart, and Kate Starbird. 2018. Acting the Part: Examining Information Operations Within #BlackLivesMatter Discourse. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 20 (Nov. 2018), 27 pages. https://doi.org/10.1145/3274289

[5] Adam J. Berinsky. 2017. Rumors and Health Care Reform: Experiments in Political Misinformation. *British Journal of Political Science* 47, 2 (2017), 241–262. https://doi.org/10.1017/S0007123415000186

[6] Md Momen Bhuiyan, Kexin Zhang, Kelsey Vick, Michael A. Horning, and Tanushree Mitra. 2018. FeedReflect: A Tool for Nudging Users to Assess News Credibility on Twitter. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) *(CSCW '18)*. Association for Computing Machinery, New York, NY, USA, 205–208. https://doi.org/10.1145/3272973.3274056

[7] Leticia Bode and Emily K. Vraga. 2015. In Related News, That was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media. *Journal of Communication* 65, 4 (06 2015), 619–638. https://doi.org/10.1111/jcom.12166

[8] Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. From Clickbait to Fake News Detection: An Approach based on Detecting the Stance of Headlines to Articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, I. Watson Research Center (Ed.). ACL, 84–89.

[9] Michael V. Bronstein, Gordon Pennycook, Adam Bear, David G. Rand, and Tyrone D. Cannon. 2019. Belief in Fake News is Associated with Delusionality, Dogmatism, Religious Fundamentalism, and Reduced Analytic Thinking. *Journal of Applied Research in Memory and Cognition* 8, 1 (2019), 108–117. https://doi.org/10.1016/j.jarmac.2018.09.005

[10] Man-Pui Sally Chan, Christopher R. Jones, Kathleen Hall Jamieson, and Dolores Albarracín. 2017. Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation. *Psychological science* 28, 11 (2017), 1531–1546. https://doi.org/10.1177/0956797617714579

[11] Katherine Clayton, Spencer Blair, Jonathan A. Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, Morgan Sandhu, Rachel Sang, Rachel Scholz-Bright, Austin T. Welch, Andrew G. Wolff, Amanda Zhou, and Brendan Nyhan. 2019. Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media. *Political Behavior* (2019), 1–23. https://doi.org/10.1007/s11109-019-09533-0

[12] Cristina Conati, Kaska Porayska-Pomsta, and Manolis Mavrikis. 2018. AI in Education needs interpretable machine learning: Lessons from Open Learner Modelling. In *Proceeedings of 2018 ICML Workshop on Human Interpretability in MachineLearning (WHI 2018)*. https://arxiv.org/pdf/1807.00154

[13] Madeleine de Cock Buning. 2018. *A multi-dimensional approach to disinformation: Report of the independent high level group on fake news and online disinformation.* Publications Office of the European Union, Luxembourg.

[14] William H. Dutton and Laleah Fernandez. 2019. How Susceptible Are Internet Users? *Intermedia* 46, 4 (2019), 36–40. http://dx.doi.org/10.2139/ssrn.3316768

[15] Ullrich K. H. Ecker, Stephan Lewandowsky, and David T. W. Tang. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & Cognition* 38, 8 (2010), 1087–1100. https://doi.org/10.3758/MC.38.8.1087

[16] Rob Ennals, Beth Trushkowsky, and John Mark Agosta. 2010. Highlighting disputed claims on the web. In *Proceedings of the international conference on World wide web*, Michael Rappa (Ed.). ACM, 341. https://doi.org/10.1145/1772690.1772726

[17] Robert Epstein, Ronald E. Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the Search Engine Manipulation Effect (SEME). *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22. https://doi.org/10.1145/3134677

[18] Martin Flintham, Christian Karner, Khaled Bachour, Helen Creswick, Neha Gupta, and Stuart Moran. 2018. Falling for Fake News: Investigating the Consumption of News via Social Media. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, Article 376, 10 pages. https://doi.org/10.1145/3173574.3173950

[19] Shane Frederick. 2005. Cognitive Reflection and Decision Making. *Journal of economic perspectives* 19, 4 (December 2005), 25–42. https://doi.org/10.1257/089533005775196732

[20] 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. https://doi.org/10.1145/3274324

[21] R. Kelly Garrett and Shannon Poulsen. 2019. Flagging Facebook Falsehoods: Self-Identified Humor Warnings Outperform Fact Checker and Peer Warnings. *Journal of Computer-Mediated Communication* 24, 5 (2019), 240–258. https://doi.org/10.1093/jcmc/zmz012

[22] R. Kelly Garrett and Brian E. Weeks. 2013. The Promise and Peril of Real-time Corrections to Political Misperceptions. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW '13)*. ACM, 1047–1058. http://doi.acm.org/10.1145/2441776.2441895

[23] Mykhailo Granik and Volodymyr Mesyura. 2017. Fake news detection using naive Bayes classifier. In *2017 IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON)*. IEEE, 900–903. https://doi.org/10.1109/UKRCON.2017.8100379

[24] Georgios Gravanis, Athena Vakali, Konstantinos Diamantaras, and Panagiotis Karadais. 2019. Behind the cues: A benchmarking study for fake news detection. *Expert Systems with Applications* 128 (2019), 201–213. https://doi.org/10.1016/j.eswa.2019.03.036

[25] Nir Grinberg, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. Fake news on Twitter during the 2016 U.S. presidential election. *Science* 363, 6425 (2019), 374–378. https://doi.org/10.1126/science.aau2706

[26] Steffen Haesler, Stefka Schmid, and Christian Reuter. 2020. Crisis Volunteering Nerds: Three Months After COVID-19 Hackathon #WirVsVirus. In *MobileHCI '20: Proceedings of the Workshop Mobile Resilience: Designing Mobile Interactive Systems for Societal and Technical Resilience*. ACM.

[27] Katrin Hartwig and Christian Reuter. 2019. TrustyTweet: An Indicator-based Browser-Plugin to Assist Users in Dealing with Fake News on Twitter. In *Proceedings of the International Conference on Wirtschaftsinformatik (WI)*. AIS, 857–871.

[28] Y. Linlin Huang, Kate Starbird, Mania Orand, Stephanie A. Stanek, and Heather T. Pedersen. 2015. Connected Through Crisis. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, Dan Cosley (Ed.). ACM, 969–980.

[29] Shan Jiang and Christo Wilson. 2018. Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23. https://doi.org/10.1145/3274351

[30] Zhiwei Jin, Juan Cao, Yongdong Zhang, and Jiebo Luo. 2016. News Verification by Exploiting Conflicting Social Viewpoints in Microblogs. In *Proceedings AAAI Conference on Artificial Intelligence* (Phoenix, Arizona). AAAI Press, 2972–2978.

[31] Marc-André Kaufhold, Nicola Rupp, Christian Reuter, and Matthias Habdank. 2020. Mitigating Information Overload in Social Media during Conflicts and Crises: Design and Evaluation of a Cross-Platform Alerting System. *Behaviour & Information Technology (BIT)* 39, 3 (2020), 319–342. https://doi.org/10.1080/0144929X.2019.1620334

[32] Stephan Lewandowsky, Ullrich K. H. Ecker, Colleen M. Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and Its Correction: Continued Influence and Successful Debiasing. *Psychological science in the public interest : a journal of the American Psychological Society* 13, 3 (2012), 106–131.

[33] Tessa Lyons. 2017. Replacing Disputed Flags With Related Articles. (2017). https://about.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/

[34] Philipp Mayring. 2000. Qualitative Content Analysis. *Forum: Qualitative Social Research* 1, 2 (2000), 10.

[35] Michael McNally and Lauren Bose. 2018. Combating False News in the Facebook News Feed: Fighting Abuse @Scale. https://www.facebook.com/atscaleevents/videos/2078868845719542/

[36] Paul Mena. 2019. Cleaning Up Social Media: The Effect of Warning Labels on Likelihood of Sharing False News on Facebook. *Policy & Internet* 3, 1 (2019), 12. https://doi.org/10.1002/poi3.214

[37] Federico Monti, Fabrizio Frasca, Davide Eynard, Damon Mannion, and Michael M. Bronstein. 2019. Fake News Detection on Social Media using Geometric Deep Learning. , 15 pages. https://arxiv.org/pdf/1902.06673

[38] Adam Mosseri. 2016. Addressing Hoaxes and Fake News | Facebook Newsroom. (2016). https://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/

[39] Brendan Nyhan and Jason Reifler. 2010. When Corrections Fail: The Persistence of Political Misperceptions. *Political Behavior* 32, 2 (2010), 303–330. https://doi.org/10.1007/s11109-010-9112-2

[40] Rosemary Pacini and Seymour Epstein. 1999. The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of personality and social psychology* 76, 6 (1999), 972–987. https://doi.org/10.1037/0022-3514.76.6.972

[41] Gordon Pennycook, Adam Bear, Evan 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* (2020). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3035384

[42] 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), 1865–1880.

[43] 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. http://www.sciencedirect.com/science/article/pii/S001002771830163X

[44] Kenneth Rapoza. 2017. Can 'Fake News' Impact The Stock Market? *Forbes* (2017). https://www.forbes.com/sites/kenrapoza/2017/02/26/can-fake-news-impact-the-stock-market/#6a27a80c2fac

[45] Hannah Rashkin, Eunsol Choi, Jin Yea Jang, Svitlana Volkova, and Yejin Choi. 2017. Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Martha Palmer, Rebecca Hwa, and Sebastian Riedel (Eds.). ACL, 2931–2937.

[46] Julio C. S. Reis, Andre Correia, Fabricio Murai, Adriano Veloso, and Fabrıcio Benevenuto. 2019. Explainable Machine Learning for Fake News Detection. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*. ACM, New York, NY, USA, 17–26. https://doi.org/10.1145/3292522.3326027

[47] Christian Reuter. 2019. *Information Technology for Peace and Security - IT-Applications and Infrastructures in Conflicts, Crises, War, and Peace*. Springer Vieweg, Wiesbaden, Germany. 1–424 pages. https://doi.org/10.1007/978-3-658-25652-4

[48] Christian Reuter. 2020. Towards IT Peace Research: Challenges at the Intersection of Peace and Conflict Research and Computer Science. *S+F Sicherheit und Frieden / Peace and Security* 38, 1 (2020), 10–16. https://doi.org/10.5771/0175-274X-2020-1-10

[49] Christian Reuter, Katrin Hartwig, Jan Kirchner, and Noah Schlegel. 2019. Fake News Perception in Germany: A Representative Study of People's Attitudes and Approaches to Counteract Disinformation. In *Proceedings of the International Conference on Wirtschaftsinformatik (WI)*. AIS, 1069–1083.

[50] Christian Reuter and Marc-André Kaufhold. 2018. Fifteen Years of Social Media in Emergencies: A Retrospective Review and Future Directions for Crisis Informatics. *Journal of Contingencies and Crisis Management (JCCM)* 26, 1 (2018), 41–57. https://doi.org/10.1111/1468-5973.12196

[51] Christian Reuter, Marc-André Kaufhold, Thomas Spielhofer, and Anna Sophie Hahne. 2017. Social Media in Emergencies: A Representative Study on Citizens' Perception in Germany. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 90 (Dec. 2017), 19 pages. https://doi.org/10.1145/3134725

[52] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[53] Guy Rosen. 2019. Helping to Protect the 2020 US Elections. *Facebook Newsroom* (2019). https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/

[54] Natali Ruchansky, Sungyong Seo, and Yan Liu. 2017. CSI: A Hybrid Deep Model for Fake News Detection. In *CIKM 2017*, Ee-Peng Lim, Rakesh Agrawal, Yu Zheng, Carlos Castillo, Aixin Sun, Vincent S. Tseng, Chenliang Li, Marianne Winslett, Mark Sanderson, Ada Fu, Jimeng Sun, Shane Culpepper, Eric Lo, Joyce Ho, and Debora Donato (Eds.). ACM Association for Computing Machinery, 797–806.

[55] Susanne G. Scott and Reginald A. Bruce. 1995. Decision-Making Style: The Development and Assessment of a New Measure. *Educational and Psychological Measurement* 55, 5 (1995), 818–831. https://doi.org/10.1177/0013164495055005017

[56] Ricky J. Sethi. 2017. Crowdsourcing the Verification of Fake News and Alternative Facts. In *HT'17*, Peter Dolog, Peter Vojtas, Francesco Bonchi, and Denis Helic (Eds.). ACM Association for Computing Machinery, 315–316.

[57] Kai Shu, H. Russell Bernard, and Huan Liu. 2019. Studying Fake News via Network Analysis: Detection and Mitigation. In *Emerging Research Challenges and Opportunities in Computational Social Network Analysis and Mining*, Nitin Agarwal, Nima Dokoohaki, and Serpil Tokdemir (Eds.). Springer International Publishing, Cham, 43–65. https://doi.org/10.1007/978-3-319-94105-9_3

[58] Kai Shu, Suhang Wang, and Huan Liu. 2019. Beyond News Contents: The Role of Social Context for Fake News Detection. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining* (Melbourne VIC, Australia) *(WSDM '19)*. ACM, NY, USA, 312–320. https://doi.org/10.1145/3289600.3290994

[59] Kate Starbird, Ahmer Arif, and Tom Wilson. 2019. Disinformation as Collaborative Work: Surfacing the Participatory Nature of Strategic Information Operations. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 127 (Nov. 2019), 26 pages. https://doi.org/10.1145/3359229

[60] Eugenio Tacchini, Gabriele Ballarin, Marco L. Della Vedova, Stefano Moret, and Luca de Alfaro. 2017. Some Like it Hoax: Automated Fake News Detection in Social Networks. In *Proceedings of the Second Workshop on Data Science for Social Good (SoGood 2017), Skopje, Macedonia, September 18, 2017. (CEUR Workshop Proceedings)*, Ricard Gavaldà, Irena Koprinska, and Stefan Kramer (Eds.). http://ceur-ws.org/Vol-1960/paper2.pdf

[61] Wei Lim Zheng Tandor, Edson C and Richard Ling. 2018. Defining "Fake News". *Digital Journalism* 6, 2 (2018), 12. https://doi.org/10.1080/21670811.2017.1360143

[62] Thomas Wood and Ethan Porter. 2019. The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence. *Political Behavior* 41, 1 (2019), 135–163. https://doi.org/10.1007/s11109-018-9443-y

[63] Liang Wu and Huan Liu. 2018. Tracing Fake-News Footprints. In *WSDM'18*, Yi Chang, Chengxiang Zhai, Yan Liu, and Yoelle Maarek (Eds.). ACM Association for Computing Machinery, 637–645.

[64] Fan Yang, Shiva K. Pentyala, Sina Mohseni, Mengnan Du, Hao Yuan, Rhema Linder, Eric D. Ragan, Shuiwang Ji, and Xia (Ben) Hu. 2019. XFake: Explainable Fake News Detector with Visualizations. In *The World Wide Web Conference* (San Francisco, CA, USA) *(WWW '19)*. ACM, NY, USA, 3600–3604. https://doi.org/10.1145/3308558.3314119

[65] Xinyi Zhou, Atishay Jain, Vir V. Phoha, and Reza Zafarani. 2019. Fake News Early Detection: A Theory-driven Model. https://arxiv.org/pdf/1904.11679

Received January 2020; revised June 2020; accepted July 2020