Meaningful Context, a Red Flag, or Both? Preferences for Enhanced Misinformation Warnings Among US Twitter Users

FILIPO SHAREVSKI, DePaul University, United States
AMY DEVINE, DePaul University, United States
EMMA PIERONI, DePaul University, United States
PETER JACHIM, DePaul University, United States

Warning users about misinformation on social media is not a simple usability task. Soft moderation has to balance between debunking falsehoods and avoiding moderation bias while preserving the social media consumption flow. Platforms thus employ minimally distinguishable warning tags with generic text under a suspected misinformation content. This approach, evidence suggests, made users either simply ignore the warning tags or believe the misinformation content more, not less. To curtail this unfavorable outcome, we developed enhancements to the misinformation warnings where users are advised on the context of the information hazard and exposed to standard warning iconography. The purpose of these enhancements is to make the warning tags easily distinguishable while conveying comprehensible, hard-to-ignore warning text. We ran an A/B evaluation with the Twitter's original warning tags in a 337 participant usability study with users from the United States. The majority of the participants preferred the enhancements as a nudge toward recognizing and avoiding misinformation. The enhanced warning tags were most favored by the politically left-leaning and to a lesser degree moderate participants, but they also appealed to roughly a third of the right-leaning participants.

CCS Concepts: • **Security and privacy → Social aspects of security and privacy**; **Usability in security and privacy**.

Additional Key Words and Phrases: misinformation, soft moderation, warnings, Twitter

ACM Reference Format:
Filipo Sharevski, Amy Devine, Emma Pieroni, and Peter Jachim. 2022. Meaningful Context, a Red Flag, or Both? Preferences for Enhanced Misinformation Warnings Among US Twitter Users. In *2022 European Symposium on Usable Security (EuroUSEC 2022), September 29–30, 2022, Karlsruhe, Germany*. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3549015.3555671

1 INTRODUCTION

Warnings and secure user behavior seems to have a perennially fraught relationship, despite the rich mediation of usability [2, 17], interaction/visual design [20], and behavioral insights [59, 80]. It is understandable that the complexity of this problem requires patience and eventual alignment between the security literacy of the average user and the pace with which new security hazards are introduced into users' daily life [18, 28]. Usable security has made noticeable advancements of warnings that users do actually heed in conformance with the security recommendations: avoiding phishy websites and questionable attachments [57], skipping unencrypted communication [75], warming up to multi-factor authentication [40], and following up on system updates [43]. Advancements such as adaptive strategies for getting accustomed to warnings and security advice also help users transition to an acceptable secure behavior [23, 29].

What actually is a bit difficult to understand is why, despite these advancements in usable security, warnings about misinformation on social media have made little progress in fostering desirable security behavior [69]. One could argue that the nature of the security hazard differs between the two settings - traditional programmatic security is far more complex to grasp than picking up on a causal post that links the COVID-19 vaccines with infertility - and that makes designing misinformation warnings an entirely different challenge. True, the one-size-fits-all here won't work because yesterday were the elections [32, 83], today is COVID-19 and QAnon [6, 49], and who knows what alternative narratives will emerge tomorrow. Embracing this predicament as a challenge in a usable security context has been sporadic so far, with the focus largely placed on mapping the "sources of misinformation" [14].

Misinformation sources won't go away, and so long as the Internet evolves, they will too [58]. In the context of social media, these sources generate *misinformation* content such as disinformation, fake news, rumors, conspiracies, hoaxes, trolling, and spam [82]. It took some time for mainstream platforms to acknowledge that they have a serious problem on hands when misinformation started piling up [45]. They responded with warnings in conformance with their interface aesthetics of and language appearing as unbiased and non-judgmental to users with diverse perspectives [72]. But this so-called "soft moderation" was applied halfheartedly, turning the warnings into *hazards themselves*—users started believing the misinformation more, not less, when a warning was explicitly appended to it [11, 50].

The design of the warnings, thus, requires adaptation of the approach to retain their usability in various misinformation scenarios while avoiding a "backfire effect" [71] as well as ensuring users don't simply ignore them as indistinguishable from the platforms' user interface. Current warnings, according to recent usability studies [11, 25, 50], have the opposite effect of the intended one when heeded in that users believe more, not less, any content that was labeled as misinformation on social media. Such an effect is unfavorable as it enables misinformation to gain traction and potentially spread across the entire social media space [45, 62]. The platforms, however, have been slow to respond to this effect as no noticeable change in the designing of the warnings has been made so far.

To fill this gap, we developed enhancements to the misinformation warning tags used by Twitter and evaluated them with a sample of 337 users from the US. The purpose of these enhancements is to address two aspects: meaningful context of the intentionally spread misinformation [82] and sufficiently potent interruption of the regular social media consumption flow [16, 24] that is hard to ignore. Therefore, we formulated the warnings' text to fit the scenario surrounding a misinformation tweet and introduced red flag watermarks as a characteristic iconography of the visual frictions that users encounter in every aspect of their daily life [12].

The results of an A/B evaluation study with the warning tags currently employed on Twitter show that the majority of users *do* welcome the usable security enhancements. The added meaningful context was praised in helping participants avoid, ignore, and skip misinformation *"right away."* The red flag watermarks were lauded for their *"attention-grabbing"* effect. There were groups of users that expressed their protest against a perceived forceful opposition-opinion-forming by Twitter as a self-appointed truth authority. Therefore, we analyzed the sentiment the warning tags incited and found that the enhancements did tilt the overall sentiment toward more positive from the *status quo* of the original Twitter warnings. Sentiment often reflects users' political leanings and is shaped by the structure of their demographic identity [38, 74]. Our results suggest that the left-leaning participants overwhelmingly welcomed the meaningful context, and the moderates and right-leaning joined them in lauding both the context and the red flag watermarks.

Following this introduction, we delve into the current state of misinformation warnings on social media in Section 2. We then elaborate on our usable security enhancement approach in Section 4. Section 5 provides the results of our A/B evaluation study and sentiment analysis. We discuss the results in Section 6 and provide our recommendations for the future of soft moderation before we conclude the paper in Section 7.

2 MISINFORMATION WARNINGS

Warnings on social media usually come in two main forms: (i) *interstitial covers* which obscure the misinformation and require users to click through to see the information [36]; or (ii) *trustworthiness tags* which appear under the content and do not interrupt the user or compel action, as shown in Figure 1. The former are more suitable for sensitive content where the exposure to the hazards should be avoided in the first place and the latter are usually applied to content with questionable factual provenance where the decision whether it is or not of misinformation nature is left to the user. But the COVID-19 "infodemic" demanded all hands on deck for soft moderation, and so mainstream social media platforms applied both warning variants to warn users of misleading and harmful COVID-19 information [45, 62].

Fig. 1. Generic Content Indicators on Twitter serving as Trustworthiness Tags.

Evidence suggest that only the interstitial covers, but not the trustworthiness tags, make the users heed the warnings of misinformation [64, 65, 69, 83]. It is tempting to simply discard the trustworthiness tags and only use interstitial covers, however. The interstitial covers do require additional clicks to get to the content in question, which could make the users avoid the content, but leave with a feeling that the social media platform is overtly imposing, "biased," "punitive" or "restrictive of free speech" [33, 64]. The trustworthiness tags might be more usable and mitigate the overt intrusion by blending with the user interface aesthetics (e.g. same colors, fonts, and obscure text), but they do run into other problems. Next to the "backfiring effect" [11, 16, 48, 74], the tags could desensitize users to soft moderation ("illusory truth effect" [51]) or might confirm misinformation as accurate ("implied truth effect" [50]).

Other factors also contribute to these negative effects, for example users' political identities. When trustworthiness tags directly challenged political falsehoods, they had the intended effect on Democrats but the opposite effect (e.g. they "backfired) on Republicans [74]. In the context of the COVID-19 pandemic, the tags resulted in a "belief echo," manifested as skepticism of adequate COVID-19 immunization particularly among Republicans and Independents [69]. Another factor is the asymmetrical nature of soft moderation—the mere exposure to misinformation often generates a strong and automatic affective response, but the warning itself may not generate a response of an equal and opposite magnitude [24]. This is because the trustworthiness tags often lack meaning, have ambiguous wording, or ask users to find context themselves which is cognitively demanding and time consuming [16]. Therefore, a natural step toward minimizing the said negative effects, is enhancing the trustworthiness tags to counter this asymmetry while keeping the appeal relevant for users of all ages, analytical prowess, and political leanings [38].

3 ENHANCED MISINFORMATION WARNINGS

The trustworthiness tags applied by Twitter make an interesting case of usable security interventions. Appended under suspected misinformation, this brand of tags warns *after* a user is exposed to the potentially harmful content [62]. Choosing to warn a user after-the-fact goes somewhat against the practice of using warning screens in browsers

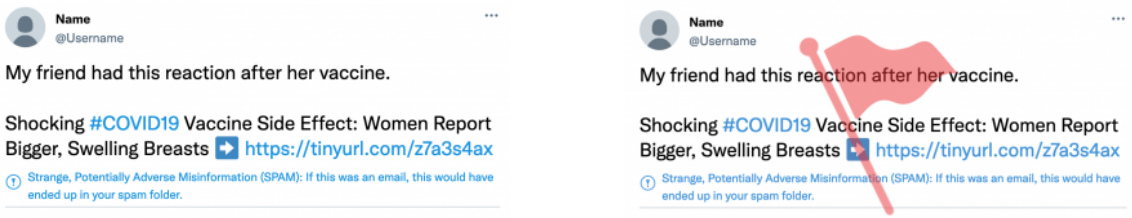
that come *before* a user gets a chance to visit a questionable website [20] (this effect is achieved with the interstitial covers, but they are verbose and disruptive of the natural social media consumption flow [7]). One could argue that the after-the-fact notification is chosen to counter "habituation", or the diminished response with repetitions of the same warning screens like these, or perhaps break the effect of "generalization" that might occur when habituation to these screens carries over to novel security interventions that look like the warning tags [78]. Camouflaged amongst the existing interface features, the warning tags are *blue* and not *red* in color, they do not obscure the suspected misinformation tweet, nor do they occur predictably like the warning screens every time an Internet browser cannot verify the visiting website's certificate (the tweets in question have to be fact checked, if not automatically flagged [32]).

Twitter's warning tags might compare to the lock icons at the beginning of an URL bar in a browser indicating a "secure" connection [56]. Besides the habituation and generalization, the lock icons are confusing and don't convey the threat to the users in the first place so proposals have been made to pair the usable security iconography with words when possible [21]. Thus, it seems reasonable to pair an exclamation mark with a generic short text for warning users about misinformation tweets. But both the icon and the text are colored in the specific Twitter blue and fail to provide contrast to attract user's attention like the lock icons do with either red for "insecure" or green for "secure" browsing (alternatively a display of a locked/broken golden lock or strike-through the word "HTTPS"). Deliberately avoiding contrast makes it easier for users to overlook, ignore, or simply mistrust the warning tags as honest security aids [40].

Perhaps pairing the generic warning text with a link to a Twitter-curated page or external trusted source containing additional information on the claims made within a suspected tweet could compensate for the lack of contrast. Often with a one-liner, users are offered to "get the facts on the COVID-19 vaccine," or "learn how the voting by mail is safe and secure" [62]. The Fear of Missing Out (FOMO) aside [4], the warning text in fact *advises* users to contextualize the tweet themselves on the particular (mis)information topic. Users, unfortunately, rarely heed this advice and largely refuse to investigate any (mis)information further [25].

Security advice is not entirely anathema to users, particularly when it comes to their online security hygiene [55]. So it is not unreasonable to expect that users might heed the suggestion brought forth, brandished in a warning tag, if the advice itself provides a *meaningful context* for a particular topic of contention on Twitter without asking users to follow a link (which conflicts Twitter's own idea of curating "more accounts, and less links" in user's feeds [7]). Balancing for comprehensibility, we developed enhanced warning tags that provide meaningful context in regards (1) fabricated facts; and (2) improbable interpretations of facts. The enhancement choice follows the misinformation front put forth by Twitter and allowed us to conduct an A/B usability evaluation with the current warning tags applied to misinformation hazard. The enhanced warnings, in their tag-only variant, incorporate catchy acronyms as frictions indented to grab users attention in the absence of contrast [12].

We paired the text-only warning tags with the hereto ignored usable security intervention when it comes to misinformation: red flags as watermarks over suspected misinformation tweets. The tag-and-watermark variant provided option for us to also test users' receptivity to warnings that incorporate contrast (red), gestalt iconography for general warnings (flag), and actionable advice for inspection (watermark). The choice of red flag was made after an extensive deliberation concerning warning design [13, 81], warning cognition (automatic or System I; deliberate or System II) [44], and user experience design [20]. We decided against a smaller red flag as smaller labeling symbols were ignored on social media, e.g. Facebook used a small red box on the left with an exclamation mark and was either ignored or users believe the flagged post more, not less [61]. We decided to use red and not other color flags because a "red flag" is a common signal of oncoming danger and requires users to switch from System I to System II of cognition.



(a) Text-only SPAM Warning Tag

(b) Text-and-flag SPAM Warning Tag

Fig. 2. Warning Tags Contextualizing Fabricated Facts

Green usually signals "no danger" while orange or yellow signal "caution" but are often processed by System I cognition [60]. Red also has the highest "perceived hazardousness" on the color palette [81].

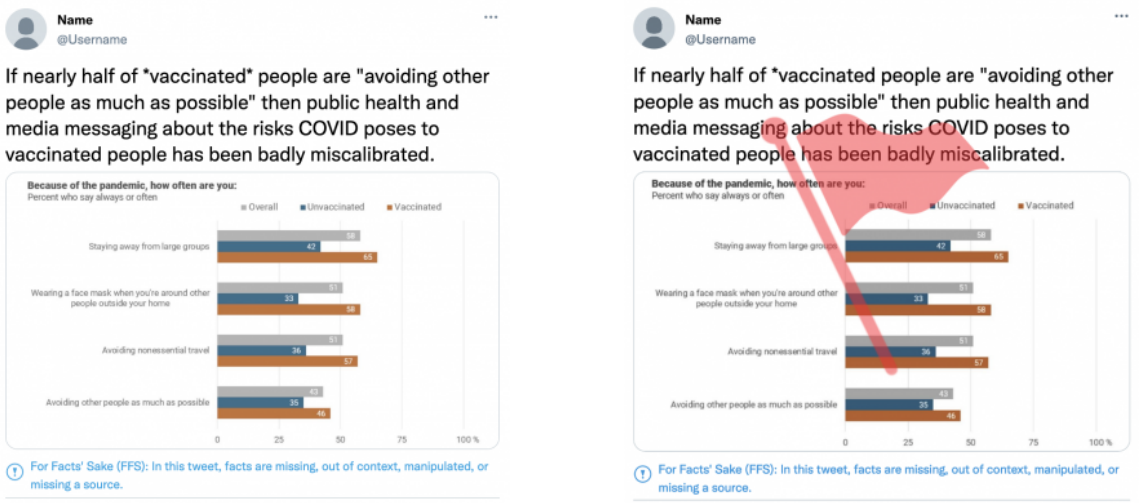
3.1 Fabricated Facts (SPAM)

The first text-only warning tag is shown in Figure 2a. We crafted a tweet, based on [37], and tagged for fabricated facts and presented it under a generic name, username, and without a profile image to avoid any threat to the validity of our A/B evaluation. Instead of advising the users to "get the facts about the COVID-19 vaccine"[62], we coined a catchy, yet familiar acronym: **SPAM** or **S**trange, **P**otentially **A**dverse **M**isinformation. With **SPAM** we wanted to see if we can contextualize the tweet's content, with an analogy to an already meaningful aspect of spam email, something most Twitter users have experience with [9]. We did break the one-liner rule for the warning text, but we opted for an increased attention and warning adherence behavior. Our warning text following the **SPAM** acronym read: "**If this was an email, this would have ended up in your spam folder**." The overarching idea with the **SPAM** warning was to harness the "availability" and "recognition" heuristics characteristic captured in a Twitter flow [1, 47]. Misinformation and fabricated facts are not always spam or vice versa, but anyhow align on the actionable outcome: ignore, delete, or take it with a grain of salt [52], which we argue is preferable compared to the "backfiring effect" of the original tags.

The upgraded **SPAM** warning tag with a 50% transparency red flag watermark over the entire tweet is shown in Figure 2b. The "upgrade" bolsters the warning tag context along the same lines of "availability" and "recognition" heuristics by invoking the well-known analogy between red flags and calls for attention. We opted for a watermark and not a replacement of the exclamation point inline the warning tag to avoid confusion with the red flag emoji frequently used on social media. The watermarking, centered in a ratio over the entire tweet area, follows the paradigm for misinformation flagging proposed in [71] with a midpoint transparency to create a non-negligible design friction for anyone attempting to read the tweet. By this choice, we wanted to stretch the overall text-and-flag warning *throughout* the suspected misinformation tweet and not only *after* it.

3.2 Improbable Interpretations of Facts (FFS)

The second set of warning tags is shown in Figure 3a and Figure 3b for the text-only and text-and-flag variants, respectively. Here we crafted a tweet, based on [46], containing an improbable interpretation of facts, keeping the engagement and posting structure in the similar order. In this case, we chose to provide a meaningful choice of context when tweets attempt to "spin" facts as a refined way of promulgating misinformation [19]. This practice, for example, earned Representative Marjorie Taylor Greene a permanent ban from Twitter [3]. Since we want to draw users' attention to such practices, we decided to ask whether they consider such tweets for **F**or **F**acts' **S**ake or **FFS**, if not for anything



(a) Text-only FFS Warning Tag

(b) Text-and-flag FFS Warning Tag

Fig. 3. Warning Tags Contextualizing Improbable Interpretation of Facts

else. We deliberately selected the acronym **FFS** to blend with the characteristic communication on Twitter that utilizes "compact language" due to the tweets' length restriction [85].

The **FFS** warning tag intended to provoke a pause in "recognition" heuristics since there are multiple meanings associated this acronym. We were aware that this might cause brief confusion, but nonetheless proceeded, since we wanted to explore if a resolution by contextual advice would suffice in refraining from taking the improbable interpretations of facts at face value. We utilized the growing evidence of "design frictions" purposefully created to disrupt mindless automatic interactions, prompting moments of reflection [12]. The brief confusion, promptly, is resolved by the following warning text advising users that "**In this tweet, facts are missing, out of context, manipulated, or missing a source.**" To gauge the limits of the warnings-as-friction, the red flag watermark provides another stimulus to capitalize on by seeing what works as a resolution against the questionable content: incomplete factual presentation [79], lack of contextual consistency [34], overt factual manipulation [68], or obscure factual provenance [30].

4 EVALUATION STUDY

4.1 Research Questions

The evaluation of the enhanced warning tags was intended to gauge a preferential approach to soft moderation as well as understand the underpinning reasoning for it's acceptance (or lack thereof). A/B testing is a regular practice in usable security studies that informs the design of interface affordances, cues, and frictions [26, 63, 67]. Building on the exposure to contextual warning tags, a qualitative inquiry of how they fare in the misinformation front is important because the soft moderation employed by social media in general, and Twitter in particular, so far has yielded far from desirable results [39]. Users' often materialize their identity and political personas within social media and Twitter discourse [27, 70], therefore we also investigated how this materialization shapes the preferences for our proposed soft moderation nudges. Based on this argumentation, the resulting research questions were:

- **RQ1:** What are the preferences of Twitter users for the **SPAM** and the **FFS** enhanced misinformation warning tags in both the text-only and text-and-flag variants?
- **RQ2:** How effective are the **SPAM** and the **FFS**, enhancements in dispelling fabricated facts and improbable interpretations of facts?
- **RQ3:** What is the relationship between the Twitter users' preferences for the enhanced misinformation warning tags and users' political leanings?

4.2 Recruitment

Our study was approved by our Institutional Review Board (IRB) before any research activities began. Subsequently we set to sample a population that was 18 years or above old, regular Twitter users from the United States through the Amazon Mechanical Turk. Reputation and attention checks were included to prevent input from bots and poor responses. The survey took 20 minutes and participants were compensated with the standard participation rate ($18 per hour). Participants were randomly assigned to either the A/B evaluation of the text-only or text-and-flag enhanced warning variants. We opted for this evaluation strategy with two separate participant groups to avoid both a generalization and habituation bias in our results by exposing participants to all of the enhancements as similar stimuli [78]. We also randomized the order of each of the **SPAM** and **FFS** text-only and text-and-flag segments for each participant.

We selected the content of the tweets to be of relevance to the participants so they could meaningfully engage with the tweet's content and see a clear relationship between the tweet and the warning tag (i.e. to prevent arbitrary and irrelevant responses). The two COVID-19 related tweets represent the main target of soft moderation front by Twitter during the execution of the study [November 2021 - January 2022] [76]. We selected one misleading tweet by Nate Silver [46], and wrote a second one based on a common piece of vaccine misinformation [37]. To account for accessibility, we provided alternative text describing each of the tweets and interventions we used to avoid visual misinterpretation. We assumed participants understood the Twitter interface, the tweets, and the warning tags. The survey was anonymous and allowed users to skip any question they were uncomfortable answering.

4.3 Process and Measures

Participants first indicated the reasons they usually come to Twitter for. Next, they were asked to indicate if they encountered warning tags and what were the content and the warnings about. We were aware that not every participant might have been exposed to warning tags so we included a small training segment where we created exposure to the concept of soft moderation with generic warning tags. A pre-exposure training was used to ensure a baseline understanding of content moderation among the participants, i.e. that Twitter uses content indicators for various types of contents (misinformation, sensitive content, graphic content, etc.) as the one shown in Figure 1.

Participants then were asked to evaluate each of the enhancements in comparison to the original tag ("Get the facts about COVID-19") [62]. This evaluation consisted of two segments of information: (i) *preference*, expressed as the choice of the option the participants preferred, either Option A (referring to the original tag) or Option B (referring to the enhancements, either SPAM or FFS in Figure 2 and Figure 3); and (ii) *justification*, expressed as a verbal argumentation in an open ended question by participants' own words. Each participant was also given the option to select neither Option A nor Option B as a preference. Participants were asked to first select a preference and then to write a textual justification, if they wanted to (it was not a required answer). Next, they asked if seeing an enhanced warning tag would influence their dismissal of the tweet or tweets on the same contested topic as misinformation (yes/no and verbal

justification). Finally we collected participants' political leanings as demographic factor regularly considered respective to studies of misinformation on social media [51].

4.4 Data Analysis

The qualitative responses from the open ended questions were coded and categorized in respect the preference and the accompanying justification and were used to answer both the **RQ1** and **RQ2**. A chi-square statistical analysis $\chi(n)$ of the relationships between the preferences and participants' political leanings was performed to uncover any statistically significant relationships that answer the **RQ3**. We performed an exploratory analysis of the preferences and justification to learn where the enhanced warning tags fair well (or vice versa) as a usable security nudges against misinformation.

For each of the justifications in the open-ended questions we performed a sentiment analysis using the Valence Aware Dictionary for Sentiment Reasoning (VADER) [15, 31, 35]. VADER yields a compound score between -1 for a very negative piece of text, and 1 for a very positive one. We also used a Linguistic Inquiry and Word Count (LIWC) analysis to qualify the sentiment expressions in the responses respective to *clout* and *tone* [73]. Each one ranges between 0 and 100 with scores close to 0 indicates less confidence and weak argumentation (clout) or negative emotions (tone).

5 RESULTS

After the consolidation and consistency checks, a total of 337 participants have completed the study, with 176 in the text-only and 161 in the text-and-flag warning tag groups, respectively. Users indicated that communication was the most frequent factor for coming to Twitter (85.4%), followed by entertainment/cultural awareness (71.8%), news (63.5%), politics (46.5%) and health (26.7%). Around every third participant (32.9%) has encountered some form of a warning tag as part of Twitter's soft moderation effort in general. The distribution of participants per their self-reported political leanings was: 147 (43.6%) left-learning, 96 (28.5%) moderate, 61(18.1%) right-leaning, and 33 (9.8%) apolitical.

To ensure consistency in the analysis and validity of the results, each of the open-ended responses in the survey was coded independently by three researchers. The codebook was simple and included a coding on the preference expressed for the A/B evaluation as well as codes for the preference justification quotes from the participants. The Fleiss's kappa κ, as a measure of inter-coder agreement, was 0.960 on average with a 0.878 lower bound for the 95% confidence, which indicates an "excellent" inter-coder agreement overall.

5.1 Fabricated Facts (SPAM)

5.1.1 A/B Evaluation. The breakdown of preferences for both the text-only and text-and-flag variants of the **SPAM** warning tag is given in Table 1, respectively with a statistical significance. The category **Original (A)** refers to the original Twitter tag as Option A in the A/B testing, **SPAM (B)** refers to the **SPAM** enhancement as Option B in the A/B testing from Figure 2 (for the text-only variant, Figure 2a; for the text-and-flag variant, Figure 2b). There is also a **Neither** category, where participants indicated they don't prefer neither Option A nor Option B. In the text-only variant, more than a half of the participants who preferred the original warning tags explicitly echoed a protest against Twitter's intrusion in contested matters such as COVID-19 vaccination. Verbosity and confusion was cited by roughly one out of five participants as a preference against the **SPAM**. The same number of participants didn't provided any justification. A small number of participants judged the **SPAM** tag as misaligned with Twitter's aesthetic and therefore, illegitimate. Neither of the text-only warning tags was the choice of 12.6% of the participants.

The **SPAM** text-only warning tag (Figure 2b) received the highest preference (46.3%). The meaningful context provided by the extended security advice was welcomed by 43.2% of them indicating that *"The SPAM explanation is a*

Table 1. **SPAM**: Preferences

Option	Pct.	Justifications	Representative Quotes
Text-only Warning Tags			
Original (A) (41.1%)	55.6%	Twitter intrusion	Telling me something is spam is an opinion concerning this topic and feels intrusive to trying to control my opinions.
	19.4%	Verbosity/confusion	Because it's simple and straight to the point. The SPAM is confusing and too wordy.
	19.4%	No justification	Warning Tag A.
	5.6%	Legitimate	Get the facts seems more legit to me.
SPAM (B) (46.3%)	43.2%	Meaningful Context	Letting me know something is SPAM and dangerous is more useful than telling me where to find facts;
	36.8%	On-point warning	It tells me right away why Twitter marked it as misinformation so I don't have to wonder the reason on my own. I can also easily decide if I agree and move on or research further outside of Twitter. I love this and would be happy to see this on posts.
	10%	Attention Grabbing	"B" is better at catching the attention of the reader. "A" could just be a service announcement - it just isn't strong enough.
	6%	Link reluctance	It's more detailed and explains why its there without having to click on anything
	4%	No justification	Warning Tag B.
Neither (12.6%)		Neutral	I wouldn't utilize either.
Text-and-Flag Warning Tags			
Original (A) (37.9%)	44.2%	Distracting Flag	Seeing the red flag almost makes the tweet look like it is harmful or not true at all. It stands out too much.
	42.6%	Twitter intrusion	I would prefer the original warning tag. "B" is too opinionated and biased.
	11.6%	No justification	Warning Tag A
SPAM (B) (46.0%)	41.9%	Attention Grabbing	I prefer the red flag, as it is impossible to miss. I often read Twitter on my phone, while I take the dog out and such, so I find myself thinking I should look something up after reading a tweet, but then I get busy doing other things and don't follow up.
	25.7%	Meaningful Context	I like the red flag for sure - and the warning tag beneath gives a better commentary on why there was a red flag.
	22.9%	On-point warning	The red flag makes it very obvious that the material is potentially false and can't be trusted
	9.5%	No justification	Warning Tag B
Neither (16.1%)		Neutral	I wouldn't prefer any of them.

valid one, and makes sense in the context of the tweet's content." The on-point warning of questionable content was cited by 36.8% in preference of *"a direct misinformation label right there without having to dig further into it."* One tenth of the pro-**SPAM** participants found the acronym and the text *catchy*, *cheeky*, and positively attention-grabbing. Reluctance to follow the links in the original tag variant was cited by 6% of the participants. Only 4% didn't provide any justification.

The pairing of the red flag with the **SPAM** warning tag was either too distracting or an indicator of Twitter's intrusion into the way content should be consumed. The preference against the text-and-flag **SPAM** tag was expressed in terms of *"visual clutter that makes the tweet more difficult to read"*, *"Doomsday level of importance"*, or *"symbol of political hate"*. The pro text-and-flag **SPAM** tag participants welcomed the attention grabbing of the red flag suggesting that *"the flag gets your attention; the text tells you it is misinformation - I tend skim when reading twitter posts and the other one is not as noticeable."*. The enhanced context and the on-point warning for misinformation was preferred because *" the flag reinforces the positive information that the tweet is spam"*.

5.1.2 Sentiment Analysis. The sentiment analysis of the preferences for **SPAM** warning tags is shown in Figure 4. The violin plots show a multimodal distribution of sentiments where the original warning tag received an equal number of positive sentiments for being *"simple and straightforward"* as well as negative sentiments that *"rather not see Twitter's judgement on whether something is misinformation or not"*. The justifications showed low confidence (clout = 26.15) but positive emotions (tone = 60.65). The text-only **SPAM** positive sentiment outweighs the negative one that captures justifications indicating that *"'B' does a better job letting you know that the tweet's information is bad"*, with a bit more confidence (clout = 32.59) and on par with the positive emotions (tone = 62.74).

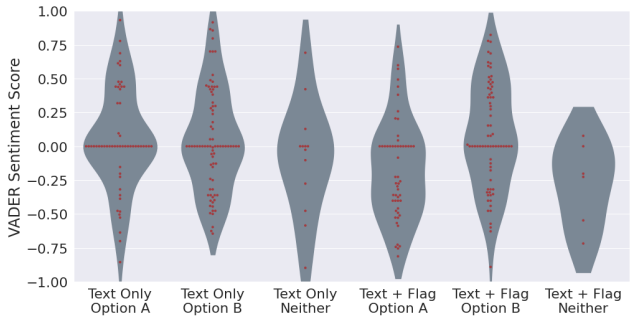


Fig. 4. Sentiments: **SPAM** Warning Tags. The violin plots show a multimodal distribution of sentiments (number of responses) over the VADER sentiment score ranging from -1 (negative) to +1 (positive) sentiment.

The introduction of the red flag in the **SPAM** warning tag apparently induced more negative sentiment when justifying the choice for the original warning tag. The justifications were a bit more convincing (clout = 30.91) but the emotions were highly negative (tone = 7.61). The red flag increased the positive sentiment for the text-and-flag **SPAM** warning tag with the most confidence of all justifications (clout = 34.11) and positive emotions (tone = 55.52). While the participants that were neither "A" or "B" were evenly distributed in the text-only variant, the negative sentiment was dominant in the text-and-flag variant. Both being very low on confidence and high on negative emotions, the introduction of the flag might have exacerbated the feelings against the soft moderation for some of these participants.

5.1.3 Dispelling Fabricated Facts. The A/B evaluation only obtained the preference for the **SPAM** warning tags without explicitly asking the participants to consider the security advice as applied to the tweets containing fabricated facts. To see if the **SPAM** warning tags actually work, we ask the participants to indicate if the tags helped them dispel fabricated facts in the example tweet. The results shown in Figure 5 indicate that the **SPAM** warning tags doesn't have to be users' best choice in order to work.

Roughly half of the ones that preferred the original warning tag commented that the text-only *"helped them understand the meaning of the tweet in a broader context."* In the text-and-flag variant participants found the warning tags helpful too rationalizing that *"Twitter should just remove the whole post in general if it comes to a big red flag watermark."* Even some of the neutral participants noted that the warning tag was reassuring on the inaccuracy of the content. Overall, 62% of the participants indicated that the **SPAM** warning tags worked for them with the desired effect of dispelling the fabricated effects of the COVID-19 vaccines.

5.1.4 Preferences and Political Leanings. The COVID-19 pandemic didn't escape deep politicization and that naturally was reflected around the soft moderation effort of Twitter following the ban of President Donald Trump [49]. We

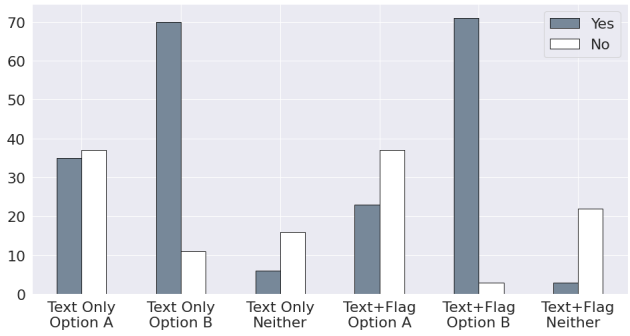


Fig. 5. Dispells: **SPAM** Warning Tags

were interested, therefore, to see if participants' preferences are affected by their political leanings. For both **SPAM** warning tags variants, as indicated in Table 4, the Pearson's Chi-Square test yielded a statistically significant relationship between their choices and where they stand on the political spectrum: $\chi(3) = 16.251$, $p = .003^*$ and $\chi(3) = 25.864$, $p = .001^*$, respectively. The original tags are appealing to left-learning with a 1:1 ratio to the moderate and 2:1 ratio to the right-leaning participants. The text-only **SPAM** variant increased these ratios to a 4.5:1. Here, the neither "A" or "B" participants are uniformly distributed.

Table 2. **SPAM** vs Political Leanings: Frequency Distribution

	Political Leanings		
Option	**Left**	**Moderate**	**Right**
Text-only Warning Tags			
Original (A)	27	27	14
Spam (B)	54	12	13
Neither	6	5	5
2 cells (22.2%) have expected count less than 5. The minimum expected count is 3.14			
Text-and-Flag Warning Tags			
Original (A)	18	23	14
Spam (B)	39	25	8
Neither	2	4	7
2 cells (22.2%) have expected count less than 5. The minimum expected count is 2.6			

The introduction of the flag tipped the left-leaning with a 1:1.3 ratio to the moderate and with a 1.3:1 ratio to the right-leaning ones that preferred the original warning tag. Left-leaning preferences for the text-and-flag **SPAM** warning tag were 1.56:1 with the moderates, but 4.875:1 with the right-leaning participants. The moderate and right-leaning were the most present for the neither "A" or "B" in the text-and-flag variant. Overall, the context is useful for the left-leaning and moderate participants the most, with a considerable portion of the moderates and right-leaning preferring a minimum intervention and distraction from Twitter.

5.2 Improbable Interpretation of Facts (FFS)

5.2.1 A/B Evaluation. The breakdown of preferences for both the text-only and text-and-flag variants of the **FFS** warning tag is given in Table 3, respectively with a statistical significance. The category **Original (A)** refers to the

original Twitter tag as Option A in the A/B testing, **FFS** refers to the **FFS** enhancement as Option B in the A/B testing from Figure 3 (for the text-only variant, Figure 3a; for the text-and-flag variant, Figure 3b). In the text-only variant, only one third preferred the original tag and more than half of the participants choose the **FFS** text-only tag. Verbosity and confusion was the reason for almost two thirds of the participants to dislike the text-only **FFS** warning-tag. Roughly one third disliked it because of an anti-soft-moderation stance and one tenth provided no justification.

Table 3. **FFS**: Preferences

Option	Pct.	Justifications	Representative Quotes
Text-only Warning Tags			
Original (A) (33.5%)	60%	Verbosity/confusion	Because the other is just too many words. It just needs to be simple to understand.
	31.6%	Twitter intrusion	Twitter is bad enough when it tries to manipulate and control their own agendas. I don't want to see more additional information.
	9.4%	No justification	Warning Tag A.
FFS (B) (50.8%)	69.2%	Meaningful Context	I would rather see context. It would bother me that some facts are missing and that I don't have the whole story. Vaccinations are too important of a topic to be misconstrued.
	10%	On-point warning	Because it explains right off the bat that this content is manipulated or missing a source.
	6.6%	Attention Grabbing	"B" is engaging with the funny acronym.
	14.2%	No justification	Warning Tag B.
Neither (15.7%)		Neutral	I wouldn't utilize either.
Text-and-Flag Warning Tags			
Original (A) (38.4%)	54.1%	Distracting Flag	It doesn't have the big red watermark that might make people feel like victims.
	34.4%	Twitter intrusion	Human beings are simple creatures, and they do not respond well to being patronized. The latter is patronizing.
	11.5%	No justification	Warning Tag A
FFS (B) (46.0%)	47.9%	Attention Grabbing	Option B really draws your attention and is impossible to miss or misunderstand.
	34.2%	Meaningful Context	It's important people really pick up on the fact this information might be misleading.
	11%	On-point warning	The flag big and bold and it will tell me easily what to avoid and what is problematic.
	6.9%	No justification	Warning Tag B.
Neither (15.6%)		Neutral	I wouldn't prefer any of them.

The meaningful context provided by the **FFS** text-only warning tag (Figure 3b) was welcomed by almost 70% of the participants *"because it doesn't just say that the tweet is disputed, it mentions the various ways that the tweet is incorrect."*. One out of ten participants liked that the **FFS** text-only warning tag because of the *"assertive statement as opposed to just one word 'disputed' in 'A'. 'B' is more specific."*. A small number deemed the acronym as *"funny/witty"* and 14.5% simply just liked the **FFS** security advice. The preference against the text-and-flag **FFS** tag was again expressed in terms of destruction by more than a half of the participants preferring the original tag. A third of them cited the contempt for Twitter's decision to patronize users about how to interpret facts. A bit more than one tenth of the pro-original warning tags didn't provide justification. The participants pro the text-and-flag **FFS** liked the attention grabbing effect of the red flag noting that they *"like that the red flag is big; You can see right away there is a problem with the tweet."* in roughly half of the cases. The context (34.2%) and the on-point warning that the tweet is a form of misinformation (11%) was preferred because *"knowing that something is missing context is more informative than knowing it's disputed; Everything is disputed by someone."*. Only 6.9% didn't provide justification pro the text-and-flag **FFS** warning tag.

5.2.2 Sentiment Analysis. The sentiment analysis of the preferences for **FFS** warning tags is shown in Figure 6. As the violin plots demonstrate, the original warning tag received roughly an equal number of positive sentiments for the *"simple and straightforward and it doesn't try to make a judgment of the tweet"* as well as negative sentiments that *"the red watermarking is overkill regardless of placement and size."*. The justifications showed again showed low confidence (clout = 21.15) but positive emotions (tone = 67.72). The text-only **FFS** positive sentiment further outweighs the negative one praising the tag's way of *"explaining why the facts are probably being used in a misleading way."*. The praises show twice as more confidence as the ones for the original warning tag (clout = 51.93) and more positive emotions (tone = 72.32).

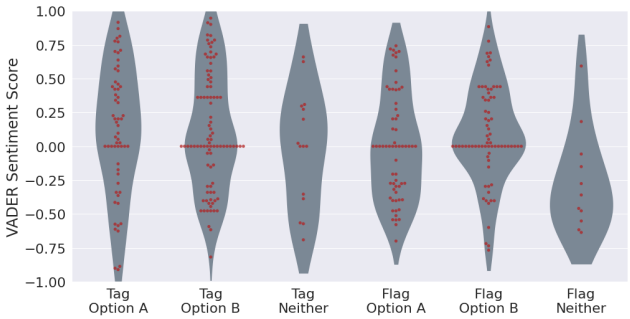


Fig. 6. Sentiments: **FFS** Warning Tags. The violin plots show a multimodal distribution of sentiments (number of responses) over the VADER sentiment score ranging from -1 (negative) to +1 (positive) sentiment.

The red flag in the **FFS** warning tag again caused a shift toward more negative sentiment as was cast as *"condescending"* and *"too distracting"*. The confidence plummeted in response to the flag-and-text variant (clout = 19.3) with the emotions remaining negative (tone = 31.61). The positive sentiment is prevalent with the pro **FFS** text-and-flag tag participants, which wielded a tad better justifications (clout = 30.95) and expressed more positive emotions (tone = 60.52). The red flag again tilted the balanced sentiment of the neutral participants in the text-only variant toward a more negative one in the text-and-flag variant.

5.2.3 Dispelling Fabricated Facts. Figure 7 shows that the detailed context provided through the **FFS** security advice is even more potent in dispelling improbable interpretation of facts. Overall, 68% of the participants indicated that the **FFS** warning tags worked for them, which is a 6% increase from the dispelling rate for the **SPAM** warning tags. Roughly half of the participants preferring the original tag conceded that the **FFS** warning tags in both variants are helpful in discrediting the manipulative tweet. A small but noticeable increase in the dispelling effect is also present for the neither "A" nor "B" participants compared to the **SPAM** warning tags. Similarly, the participants preferring both **FFS** tags were slightly more assertive of the desired effect compared to their responses for the **SPAM** tag.

5.2.4 Preferences and Political Leanings. The preferences for both **FFS** warning tags variants, as indicated in Table 4, were related with a statistical significance to participants' political leanings: $\chi(3) = 27.732$, $p = .000^*$ and $\chi(3) = 36.483$, $p = .000^*$, respectively. The original tags are appealing to left-learning participants with a 1:1 ratio to the moderate ones and with a 2.5:1 ratio to the right-leaning participants. The text-only **FFS** variant has these ratios increased to a 3.6:1 between the left-leaning and the moderate participants and 4.8:1 between the left-learning and right-leaning participants. Unlike the **SPAM** variants, here, the neither "A" or "B" participants are dominantly right-leaning with a 2:1 ratio to both the left-leaning and moderate participants.

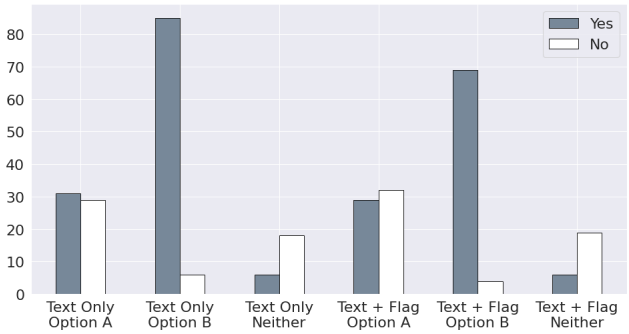


Fig. 7. Dispels: **FFS** Warning Tags

Table 4. **FFS** vs Political Leanings: Frequency Distribution

	Political Leanings		
Option	**Left**	**Moderate**	**Right**
Text-only Warning Tags			
Original (A)	25	23	10
Spam (B)	58	16	12
Neither	4	5	10
2 cells (22.2%) have expected count less than 5. The minimum expected count is 3.73			
Text-and-Flag Warning Tags			
Original (A)	20	26	12
Spam (B)	38	23	7
Neither	1	3	10
2 cells (22.2%) have expected count less than 5. The minimum expected count is 2.90			

The introduction of the flag again kept the balance between the left-learning and moderate participants, but increased the ratio to almost 2:1 to the right-leaning ones that preferred the original warning tag. The left-leaning preferences for the text-and-flag **FFS** warning tag were in a 1.65:1 ratio with the moderates, but in an overwhelming 5.42:1 ratio with the right-leaning participants. The right-leaning again dominate in the neither "A" or "B" preferences for the **FFS** text-and-flag variant. Compared to the **SPAM** case, the extended **FFS** context is even more useful for the left-leaning participants. The moderates are roughly evenly split, but the right-leaning participants show a more salient anti-soft-moderation preference when exposed to the **FFS** warning tags.

6 DISCUSSION

In this study we were motivated to bring soft moderation closer to users' everyday experiences while minimizing imposition, which as witnessed, often backfire [71]. We distinguished between a need for context when the hazard comes from the fabrication of facts and when the hazard comes from the interpretation of facts in a rather improbable way. In the first case, we were careful to avoid the perception trap of "correction of feelings, not falshoods" [41] and used an analogy with spam emails. We did so because users, by now, can recognize spam when they see it [10] and accept that spam filtering, performed by email providers, works well [53]. Understanding this, we wanted to regain the trust in the platform and signal absence of bias or judgment in their action [42].

With this in mind, the **SPAM** warning tag shows a promising step toward unified interpretation and increased trust in soft moderation (only related to COVID-19 misinformation, for now). If support from left-leaning participants was already hinted at from previous studies, it was nonetheless reinforced in both the text-only (*"...it tells participants, rather quickly, that the tweet is garbage"* and text-and-flag variants (*"the red flag will alert me before I even read any of it"*. Moderates were evenly split, expectedly, but reassured that the text-only variant *"really tells you more of what is going on"* while the text-and-flag variant *"gives more specifics and is thus tougher to refute"*. In significant numbers, right-leaning participants made it clear that the text-only variant *seems more appropriate because it's far more specific; the original tag feels more like an ad and nothing that I didn't already know."* and praised the text-and-flag variant as *"a large visual cue that's hard to ignore and will bring attention to the idea that something is going on with this information."*.

In the second case, we wanted to avoid authoritative imposition and thus worded the warning not to personify senior public health experts, usually responsible for interpretation of facts [77]. We also opted for a "bold" acronym choice to lure users' attention to the text of the warning tag. Once "hooked," the cost to read the warning text was less then avoiding it as the derivation of new meaning to acronyms is a pragmatic way of conveying context on social media, e.g. the hashtags on Twitter [66]. The text wasn't asking the user to "get facts" or "learn more," but instead, it gave several convincing options for users themselves to pick why the context is fitting to the possibly misinformation tweet [54].

The **FFS** tag did just that and succeeded. Left-learning participants liked that the text-only variant *"gives real reasons why this tweet is suspect"* and moderates seconded that the **FFS**'s context *"goes more in depth and makes you more alert to the tweet"*. Right-leaning participants confirmed our idea to avoid any relationship to an imposing authority: *"The context in 'B' is better because Facebook came out saying that most if not all of their fact checkers don't check for facts, they just do it on opinion base. I'm sure Twitter does the same"*. The consensus across the political spectrum was that the *"red flag watermark was really draws more attention"*, lead by the left-leaning ones, supports the potency of the **FFS** acronym as the "hook" entirely absent in the current soft moderation on Twitter.

6.1 Ethical Considerations

Ethical concerns do arise when dealing with misinformation, or allegedly harmful information, within the pluralistic social media population. The tension between impartiality, profitability, and social responsibility of the platforms might not always ensure that misinformation is dealt with using consistent soft moderation criteria. With the honest, yet inevitable false-positives/negatives, the proposed enhancements - if applied - might be seen as unfair at best or simply harmful at worst. We therefore are open for democratic participation in the design that allows for remediation of concerns in such instances. Soft moderation, at least in our view, is a form of honest communicative action rather than an authoritative and absolute determination of truth, and as such, beneficial to all Twitter users without discrimination [5]. We are aware that facts change, become irrelevant, or are refuted over time so a retroactive application should also be considered to enable versatile soft moderation to the best of our (and Twitter's) abilities.

6.2 Limitations

We note several limitations of our study to be addressed in future. The enhancements in misinformation warnings rest on the assumption that an "enhacement" is needed in the first place when warning users about potentially false information on Twitter. While we arrived at this assumption informed by studies suggesting the current Twitter warnings are not working as intended (users believed the false information more, not less [11, 50]), the lack of so-called "debunking" effect does not directly posit that SPAM, FFS, or any an "enhacement" will in and of itself hypothetically remedy this issue. We acknowledge that future studies going beyond the preferences of the users and towards studying the user

behaviour surrounding misinformation warnings must test such assumptions in the first place to clearly identify all the confounding factors that might operate before the misinformation label can take its effect.

For example, an implicit assumption that we have not tested in our hypothesising (withing the formulation of the research questions) about enhancements is that the users trusted Twitter to be as much as possible impartial, faithful to known facts, and unbiased in the process of labeling. This might not hold true in future studies, especially knowing that fact-checking services not always enjoy a full transparency and integrity reputations [8], and should be further tested in what capacity such assumption underpins findings about misinformation warning preferences and user behavior. To the last point, our study focused on exploring usability preferences but not *per se* the behavior of users in the United States. This distinction is important to made because the assumption that the usability preferences are a proxy for believing or identifying misinformation might not hold true in all instances concerning misinformation. A particular design or wording of a misinformation label enhancement, be that SPAM or FFS, might be well suited for a user and applied to all the related false information, yet the user might still fall for it or unwittingly propagate it. Studying the actual behaviour entails much larger resources and longitudinal research design, a future investigation which we are already committed to, based on the exploratory significance of the results in this study. It also entails a stronger theoretical background, which, at the particular time when we conducted our study, was limited in availability and applicability to the exploratory scope of our work.

Regarding the research design, the size of the sample could be enlarged to obtain an as varied as possible Twitter population. We used only two examples of misinformation on COVID-19, which is a limitation steaming both from restricted financial resources and limited attention span of participants [36]. An extended, or perhaps a longitudinal study that incorporates more COVID-19 misinformation instances over a time could not just help generalize our findings, but reveal important behavioral patterns in dealing with soft moderation. Also, it could help with an A/B evaluation for warning tags pertaining other contested topics such as elections [83]. Participants were exposed to generic formatting of the tweets emphasizing the content and the warning tags. In reality, misinformation could come from individual accounts, influencers, or accounts controlled by nefarious actors [84].

Misinformation is often amplified by social bots, and appears in users' feeds with variable degree of visual interference [22]. Controlling for this interference requires a study executed in partnership with Twitter where the enhancements are tested with selected users on the live platform. Such a test could not just capture the preferences of the regular Twitter users but help observe the "backfiring," "implied truth," and "illusory truth" effects. We didn't explicitly test for these in our study. Our A/B evaluation is limited by the current formatting of the original warning tags on Twitter [62]. If Twitter chooses to reformat the tags, eliminate the links, or place them elsewhere, the enhancements also should change and the results might not hold for these new conditions.

7 CONCLUSION

This paper conveys the first extensive A/B evaluation of enhancements for misinformation warnings on Twitter. Providing users a meaningful context and attention-grabbing iconography, our results suggest, does help users recognize and contain COVID-19 misinformation. We weren't poised to solve the predicament of soft moderation in one shot; rather, the goal was to utilize the usable security body of knowledge to trace a path toward "inoculation" against information hazards on social media.

REFERENCES

[1] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. 2017. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. 50, 3, Article 44 (aug 2017), 41 pages. https://doi.org/10.1145/3054926

[2] Devdatta Akhawe and Adrienne Porter Felt. 2013. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness. In *22nd USENIX Security Symposium (USENIX Security 13)*. USENIX Association, Washington, D.C., 257–272. https://www.usenix.org/conference/usenixsecurity13/technical-sessions/presentation/akhawe

[3] Davie Alba. 2022. Twitter Permanently Suspends Marjorie Taylor Greene's Account. https://www.nytimes.com/2022/01/02/technology/marjorie-taylor-greene-twitter.html

[4] Aarif Alutaybi, Emily Arden-Close, John McAlaney, Angelos Stefanidis, Keith Phalp, and Raian Ali. 2019. How Can Social Networks Design Trigger Fear of Missing Out?. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. 3758–3765. https://doi.org/10.1109/SMC.2019.8914672

[5] Jack Andersen and Sille Obelitz Søe. 2020. Communicative actions we live by: The problem with fact-checking, tagging or flagging fake news – the case of Facebook. *European Journal of Communication* 35, 2 (2020), 126–139. https://doi.org/10.1177/0267323119894489

[6] Ahmed Anwar, Haider Ilyas, Ussama Yaqub, and Salma Zaman. 2021. Analyzing QAnon on Twitter in Context of US Elections 2020: Analysis of User Messages and Profiles Using VADER and BERT Topic Modeling. In *DG.O2021: The 22nd Annual International Conference on Digital Government Research* (Omaha, NE, USA) *(DG.O'21)*. Association for Computing Machinery, New York, NY, USA, 82–88. https://doi.org/10.1145/3463677.3463718

[7] Jack Bandy and Nicholas Diakopoulos. 2021. More Accounts, Fewer Links: How Algorithmic Curation Impacts Media Exposure in Twitter Timelines. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 78 (apr 2021), 28 pages. https://doi.org/10.1145/3449152

[8] Petter Bae Brandtzaeg and Asbjørn Følstad. 2017. Trust and Distrust in Online Fact-Checking Services. *Commun. ACM* 60, 9 (aug 2017), 65–71. https://doi.org/10.1145/3122803

[9] Fin Brunton. 2013. *Spam: A Shadow History of the Internet.* MIT Press, Cambridge, MA.

[10] Casey Inez Canfield, Baruch Fischhoff, and Alex Davis. 2016. Quantifying Phishing Susceptibility for Detection and Behavior Decisions. *Human Factors* 58, 8 (2016), 1158–1172. https://doi.org/10.1177/0018720816665025

[11] Katherine Clayton, Spencer Blair, Jonathan A Busam, Samuel Forstner, John Glance, Guy Green, Anna Kawata, Akhila Kovvuri, Jonathan Martin, Evan Morgan, et al. 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.

[12] Anna L. Cox, Sandy J.J. Gould, Marta E. Cecchinato, Ioanna Iacovides, and Ian Renfree. 2016. Design Frictions for Mindful Interactions: The Case for Microboundaries. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 1389–1397. https://doi.org/10.1145/2851581.2892410

[13] S. David Leonard. 1999. Does color of warnings affect risk perception? *International Journal of Industrial Ergonomics* 23, 5 (1999), 499–504.

[14] Marc Dupuis, Kelly Chhor, and Nhu Ly. 2021. Misinformation and Disinformation in the Era of COVID-19: The Role of Primary Information Sources and the Development of Attitudes Toward Vaccination. In *Proceedings of the 22st Annual Conference on Information Technology Education* (SnowBird, UT, USA) *(SIGITE '21)*. Association for Computing Machinery, New York, NY, USA, 105–110. https://doi.org/10.1145/3450329.3476866

[15] Upasana Dutta, Rhett Hanscom, Jason Shuo Zhang, Richard Han, Tamara Lehman, Qin Lv, and Shivakant Mishra. 2021. Analyzing Twitter Users' Behavior Before and After Contact by the Russia's Internet Research Agency. 5, CSCW1, Article 90 (apr 2021), 24 pages. https://doi.org/10.1145/3449164

[16] Ullrich KH Ecker, Stephan Lewandowsky, and David TW Tang. 2010. Explicit warnings reduce but do not eliminate the continued influence of misinformation. *Memory & cognition* 38, 8 (2010), 1087–1100.

[17] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy) *(CHI '08)*. Association for Computing Machinery, New York, NY, USA, 1065–1074. https://doi.org/10.1145/1357054.1357219

[18] Michael Fagan and Mohammad Maifi Hasan Khan. 2016. Why Do They Do What They Do?: A Study of What Motivates Users to (Not) Follow Computer Security Advice. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 59–75. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/fagan

[19] Don Fallis. 2014. A functional analysis of disinformation. *iConference 2014 Proceedings* (2014).

[20] Adrienne Porter Felt, Alex Ainslie, Robert W. Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. 2015. Improving SSL Warnings: Comprehension and Adherence. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 2893–2902. https://doi.org/10.1145/2702123.2702442

[21] Adrienne Porter Felt, Robert W. Reeder, Alex Ainslie, Helen Harris, Max Walker, Christopher Thompson, Mustafa Embre Acer, Elisabeth Morant, and Sunny Consolvo. 2016. Rethinking Connection Security Indicators. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 1–14. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/porter-felt

[22] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. 2016. The Rise of Social Bots. *Commun. ACM* 59, 7 (jun 2016), 96–104. https://doi.org/10.1145/2818717

[23] Alain Forget, Sarah Pearman, Jeremy Thomas, Alessandro Acquisti, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, Marian Harbach, and Rahul Telang. 2016. Do or Do Not, There Is No Try: User Engagement May Not Improve Security Outcomes. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 97–111. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/forget

[24] Bertram Gawronski, Roland Deutsch, Sawsan Mbirkou, Beate Seibt, and Fritz Strack. 2008. When "Just Say No" is not enough: Affirmation versus negation training and the reduction of automatic stereotype activation. *Journal of Experimental Social Psychology* 44, 2 (2008), 370–377. https://doi.org/10.1016/j.jesp.2006.12.004

[25] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake News on Facebook and Twitter: Investigating How People (Don't) Investigate. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376784

[26] Sanam Ghorbani Lyastani, Michael Schilling, Michaela Neumayr, Michael Backes, and Sven Bugiel. 2020. Is FIDO2 the Kingslayer of User Authentication? A Comparative Usability Study of FIDO2 Passwordless Authentication. In *2020 IEEE Symposium on Security and Privacy (SP)*. 268–285. https://doi.org/10.1109/SP40000.2020.00047

[27] Jennifer Golbeck and Derek Hansen. 2011. Computing Political Preference among Twitter Followers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 1105–1108. https://doi.org/10.1145/1978942.1979106

[28] Cormac Herley. 2009. So Long, and No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In *Proceedings of the 2009 Workshop on New Security Paradigms Workshop* (Oxford, United Kingdom) *(NSPW '09)*. Association for Computing Machinery, New York, NY, USA, 133–144. https://doi.org/10.1145/1719030.1719050

[29] Jonas Hielscher, Annette Kluge, Uta Menges, and M. Angela Sasse. 2021. "Taking out the Trash": Why Security Behavior Change Requires Intentional Forgetting. In *New Security Paradigms Workshop* (Virtual Event, USA) *(NSPW '21)*. Association for Computing Machinery, New York, NY, USA, 108–122. https://doi.org/10.1145/3498891.3498902

[30] Austin Hounsel, Jordan Holland, Ben Kaiser, Kevin Borgolte, Nick Feamster, and Jonathan Mayer. 2020. Identifying Disinformation Websites Using Infrastructure Features. In *10th USENIX Workshop on Free and Open Communications on the Internet (FOCI 20)*. USENIX Association. https://www.usenix.org/conference/foci20/presentation/hounsel

[31] C.J. Hutto and Eric Gilbert. 2015. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.*

[32] Peter Jachim, Filipo Sharevski, and Emma Pieroni. 2021. TrollHunter2020: Real-Time Detection of Trolling Narratives on Twitter During the 2020 U.S. Elections. In *Proceedings of the 2021 ACM Workshop on Security and Privacy Analytics* (Virtual Event, USA) *(IWSPA '21)*. Association for Computing Machinery, New York, NY, USA, 55–65. https://doi.org/10.1145/3445970.3451158

[33] Shagun Jhaver, Christian Boylston, Diyi Yang, and Amy Bruckman. 2021. Evaluating the Effectiveness of Deplatforming as a Moderation Strategy on Twitter. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 381 (oct 2021), 30 pages. https://doi.org/10.1145/3479525

[34] Shan Jiang and Christo Wilson. 2018. Linguistic Signals under Misinformation and Fact-Checking: Evidence from User Comments on Social Media. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 82 (nov 2018), 23 pages. https://doi.org/10.1145/3274351

[35] Kenneth Joseph, Wei Wei, and Kathleen M. Carley. 2017. Girls Rule, Boys Drool: Extracting Semantic and Affective Stereotypes from Twitter. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) *(CSCW '17)*. Association for Computing Machinery, New York, NY, USA, 1362–1374. https://doi.org/10.1145/2998181.2998187

[36] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J. Nathan Matias, and Jonathan Mayer. 2021. Adapting Security Warnings to Counter Online Disinformation. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 1163–1180. https://www.usenix.org/conference/usenixsecurity21/presentation/kaiser

[37] Kara Corvus. 2021. The COVID vaccine totally makes your boobs bigger and grows your pp AT LEAST 3 inches. WHO CAN CONFIRM? We have to spread awareness and the truth about these vaccines! https://mobile.twitter.com/karacorvus/status/1421498443802021897

[38] 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 (oct 2020), 27 pages. https://doi.org/10.1145/3415211

[39] 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 (oct 2020), 27 pages. https://doi.org/10.1145/3415211

[40] Katharina Krombholz, Karoline Busse, Katharina Pfeffer, Matthew Smith, and Emanuel von Zezschwitz. 2019. "If HTTPS Were Secure, I Wouldn't Need 2FA" - End User and Administrator Mental Models of HTTPS. In *2019 IEEE Symposium on Security and Privacy (SP)*. 246–263. https://doi.org/10.1109/SP.2019.00060

[41] Stephan Lewandowsky. 2020. The 'post-truth'world, misinformation, and information literacy: A perspective from cognitive science. *Informed societies—Why information literacy matters for citizenship, participation and democracy* (2020), 69–88.

[42] Cameron Martel, Mohsen Mosleh, and David Gertler Rand. 2021. You're definitely wrong, maybe: Correction style has minimal effect on corrections of misinformation online. (2021).

[43] Arunesh Mathur, Josefine Engel, Sonam Sobti, Victoria Chang, and Marshini Chetty. 2016. "They Keep Coming Back Like Zombies": Improving Software Updating Interfaces. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 43–58. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/mathur

[44] Patricia L Moravec, Antino Kim, and Alan R Dennis. 2020. Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research* 31, 3 (2020), 987–1006.

[45] Adam Mosseri. 2016. Addressing Hoaxes and Fake News. *Facebook* (2016). https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.

[46] Nate Silver. 2021. If nearly half of *vaccinated* people are "avoiding other people as much as possible" then public health and media messaging about the risks COVID poses to vaccinated people has been badly miscalibrated. https://apnorc.org/projects/majorities-support-vaccine-mandates-for-some-activities-amidst-delta-surge/ https://t.co/U24wJ5f5hq. https://twitter.com/natesilver538/status/1428771069146537984

[47] Jakob Nielsen and Rolf Molich. 1990. Heuristic Evaluation of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Seattle, Washington, USA) *(CHI '90)*. Association for Computing Machinery, New York, NY, USA, 249–256. https://doi.org/10.1145/97243.97281

[48] Brendan Nyhan and Jason Reifler. 2010. When corrections fail: The persistence of political misperceptions. *Political Behavior* 32, 2 (2010), 303–330.

[49] Emma Peironi, Peter Jachim, Nathaniel Jachim, and Filipo Sharevski. 2021. Parlermonium: A Data-Driven UX Design Evaluation of the Parler Platform. In *Critical Thinking in the Age of Misinformation CHI 2021*.

[50] Gordon Pennycook, Adam Bear, Evan T Collins, and David G Rand. 2020. The implied truth effect: Attaching warnings to a subset of fake news headlines increases perceived accuracy of headlines without warnings. *Management Science* (2020).

[51] 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.

[52] Elissa M. Redmiles, Neha Chachra, and Brian Waismeyer. 2018. *Examining the Demand for Spam: Who Clicks?* Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3173786

[53] Elissa M. Redmiles, Neha Chachra, and Brian Waismeyer. 2018. *Examining the Demand for Spam: Who Clicks?* Association for Computing Machinery, New York, NY, USA, 1–10. https://doi.org/10.1145/3173574.3173786

[54] Elissa M. Redmiles, Amelia R. Malone, and Michelle L. Mazurek. 2016. I Think They're Trying to Tell Me Something: Advice Sources and Selection for Digital Security. In *2016 IEEE Symposium on Security and Privacy (SP)*. 272–288. https://doi.org/10.1109/SP.2016.24

[55] Elissa M. Redmiles, Noel Warford, Amritha Jayanti, Aravind Koneru, Sean Kross, Miraida Morales, Rock Stevens, and Michelle L. Mazurek. 2020. A Comprehensive Quality Evaluation of Security and Privacy Advice on the Web. In *29th USENIX Security Symposium (USENIX Security 20)*. USENIX Association, 89–108. https://www.usenix.org/conference/usenixsecurity20/presentation/redmiles

[56] Robert W. Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. 2018. *An Experience Sampling Study of User Reactions to Browser Warnings in the Field*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3173574.3174086

[57] Benjamin Reinheimer, Lukas Aldag, Peter Mayer, Mattia Mossano, Reyhan Duezguen, Bettina Lofthouse, Tatiana von Landesberger, and Melanie Volkamer. 2020. An investigation of phishing awareness and education over time: When and how to best remind users. In *Sixteenth Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association, 259–284. https://www.usenix.org/conference/soups2020/presentation/reinheimer

[58] Thomas Rid. 2020. *Active measures: The secret history of disinformation and political warfare*. Farrar, Straus and Giroux.

[59] Richard Roberts, Daniela Lulli, Abolee Raut, Kelsey R Fulton, and Dave Levin. 2020. Mental models of domain names and urls. In *Sixteen Symposium on Usable Privacy and Security (SOUPS 2020)*. USENIX Association.

[60] Wendy A. Rogers, Nina Lamson, and Gabriel K. Rousseau. 2000. Warning Research: An Integrative Perspective. *Human Factors* 42, 1 (2000), 102–139.

[61] Jon Roozenbeek and Sander van der Linden. 2019. Fake news game confers psychological resistance against online misinformation. *Palgrave Communications* 5, 1 (2019), 65.

[62] Yoel Roth and Nick Pickles. 2020. Updating our approach to misleading information. *Twitter* (2020). https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html.

[63] Scott Ruoti, Jeff Andersen, Tyler Monson, Daniel Zappala, and Kent Seamons. 2018. A Comparative Usability Study of Key Management in Secure Email. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. USENIX Association, Baltimore, MD, 375–394. https://www.usenix.org/conference/soups2018/presentation/ruoti

[64] Emily Saltz, Claire R Leibowicz, and Claire Wardle. 2021. *Encounters with Visual Misinformation and Labels Across Platforms: An Interview and Diary Study to Inform Ecosystem Approaches to Misinformation Interventions*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411763.3451807

[65] Zeve Sanderson, Megan A Brown, Richard Bonneau, Jonathan Nagler, and Joshua A Tucker. 2021. Twitter flagged Donald Trump's tweets with election misinformation: They continued to spread both on and off the platform. *Harvard Kennedy School Misinformation Review* (2021).

[66] Kate Scott. 2015. The pragmatics of hashtags: Inference and conversational style on Twitter. *Journal of Pragmatics* 81 (2015), 8–20. https://doi.org/10.1016/j.pragma.2015.03.015

[67] Sunyoung Seiler-Hwang, Patricia Arias-Cabarcos, Andrés Marín, Florina Almenares, Daniel Díaz-Sánchez, and Christian Becker. 2019. "I Don't See Why I Would Ever Want to Use It": Analyzing the Usability of Popular Smartphone Password Managers. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London, United Kingdom) *(CCS '19)*. Association for Computing Machinery, New York, NY, USA, 1937–1953. https://doi.org/10.1145/3319535.3354192

[68] Gautam Kishore Shahi, Anne Dirkson, and Tim A. Majchrzak. 2021. An exploratory study of COVID-19 misinformation on Twitter. *Online Social Networks and Media* 22 (2021), 100104. https://doi.org/10.1016/j.osnem.2020.100104

[69] Filipo Sharevski, Raniem Alsaadi, Peter Jachim, and Emma Pieroni. 2022. Misinformation warnings: Twitter's soft moderation effects on COVID-19 vaccine belief echoes. *Computers & Security* 114 (2022), 102577. https://doi.org/10.1016/j.cose.2021.102577

[70] Filipo Sharevski, Allice Huff, Peter Jachim, and Emma Pieroni. 2021. (Mis)perceptions and Engagement on Twitter: COVID-19 Vaccine Rumors on Efficacy and Mass Immunization Effort. arXiv:2111.05815 [cs.SI]

[71] Filipo Sharevski, Peter Jachim, Emma Pieroni, and Nate Jachim. 2021. VoxPop: An Experimental Social Media Platform for Calibrated (Mis)Information Discourse. In *New Security Paradigms Workshop* (Virtual Event, USA) *(NSPW '21)*. Association for Computing Machinery, New York, NY, USA, 88–107. https://doi.org/10.1145/3498891.3498893

[72] Jeff Smith. 2017. Designing Against Misinformation. *Medium* (2017). https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2.

[73] Y. R. Tausczik and J. W. Pennebaker. 2010. The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology* 29, 1 (2010), 24–54.

[74] Emily Thorson. 2016. Belief echoes: The persistent effects of corrected misinformation. *Political Communication* 33, 3 (2016), 460–480.

[75] Christian Tiefenau, Emanuel von Zezschwitz, Maximilian Häring, Katharina Krombholz, and Matthew Smith. 2019. A Usability Evaluation of Let's Encrypt and Certbot: Usable Security Done Right. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London, United Kingdom) *(CCS '19)*. Association for Computing Machinery, New York, NY, USA, 1971–1988. https://doi.org/10.1145/3319535.3363220

[76] The New York Times. 2022. Tracking Viral Misinformation. https://www.nytimes.com/live/2020/2020-election-misinformation-distortions

[77] Samuel P Trethewey. 2020. Strategies to combat medical misinformation on social media. , 4–6 pages.

[78] Anthony Vance, David Eargle, Jeffrey L. Jenkins, C. Brock Kirwan, and Bonnie Brinton Anderson. 2019. The Fog of Warnings: How Non-essential Notifications Blur with Security Warnings. In *Fifteenth Symposium on Usable Privacy and Security (SOUPS 2019)*. USENIX Association, Santa Clara, CA. https://www.usenix.org/conference/soups2019/presentation/vance

[79] Nathan Walter, Jonathan Cohen, R. Lance Holbert, and Yasmin Morag. 2020. Fact-Checking: A Meta-Analysis of What Works and for Whom. *Political Communication* 37, 3 (2020), 350–375. https://doi.org/10.1080/10584609.2019.1668894

[80] Rick Wash. 2010. Folk Models of Home Computer Security. In *Proceedings of the Sixth Symposium on Usable Privacy and Security* (Redmond, Washington, USA) *(SOUPS '10)*. Association for Computing Machinery, New York, NY, USA, Article 11, 16 pages. https://doi.org/10.1145/1837110.1837125

[81] Michael S Wogalter, Vincent C Conzola, and Tonya L Smith-Jackson. 2002. Research-based guidelines for warning design and evaluation. *Applied Ergonomics* 33, 3 (2002), 219–230.

[82] Liang Wu, Fred Morstatter, Kathleen M. Carley, and Huan Liu. 2019. Misinformation in Social Media: Definition, Manipulation, and Detection. *SIGKDD Explor. Newsl.* 21, 2 (nov 2019), 80–90. https://doi.org/10.1145/3373464.3373475

[83] Savvas Zannettou. 2021. "I Won the Election!":An Empirical Analysis of Soft Moderation Interventions on Twitter. *arXiv* 2101.07183v1 (18 January 2021). https://arxiv.org/pdf/2101.07183.pdf.

[84] Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. The Web of False Information: Rumors, Fake News, Hoaxes, Clickbait, and Various Other Shenanigans. *J. Data and Information Quality* 11, 3, Article 10 (may 2019), 37 pages. https://doi.org/10.1145/3309699

[85] David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018. The State-of-the-Art in Twitter Sentiment Analysis: A Review and Benchmark Evaluation. *ACM Trans. Manage. Inf. Syst.* 9, 2, Article 5 (aug 2018), 29 pages. https://doi.org/10.1145/3185045