You Know It's Fake, Right? How Habituation May Assist Misinformation Mitigation on TikTok

Early Stage Paper

Chengqi (John) Guo
James Madison University
guocx@jmu.edu

Chen Guo
James Madison University
guo4cx@jmu.edu

Nan Zheng
James Madison University
zhengnx@jmu.edu

Xin (Robert) Luo
University of New Mexico
xinluo@unm.edu

ABSTRACT

Misinformation is running rampant on short video social media platforms. Meanwhile, subscribers and viewers who binge TikTok videos or alike continuously ignore security warnings due to habituation, which research has considered a threat to security. The present study offers a nuanced view of habituation; it argues that decreased attention to security warnings may benefit misinformation mitigation efforts in a unique but subtle way, complementing the existing belief that habituation is a serious threat to the effectiveness of security warnings. A series of interrelated experiments was used to reveal and investigate the "ignore the warning but remain vigilant" behavioral response that, to our knowledge, no detailed information systems (IS) research has examined. We obtained early-stage empirical findings via eye-tracking, mouse cursor tracking, and think-out-loud interviews that measured habituation. The preliminary findings suggested that the efficacious role of memory/comprehension in promoting habituation can positively influence security warning effectiveness, paving the way for future inquiries into the interplay between stimulus and habituation in IS security research.

Keywords

Habituation, Security Warnings, Misinformation Mitigation, Social Media, Behavioral Experiment.

INTRODUCTION

Despite immense progress in cybersecurity warning effectiveness in the last several years, misinformation remains a persistent threat to netizens engaged in social media environments (Kaiser et al., 2021; Reeder et al., 2018). Moreover, video fake news is believed and shared more than text and audio versions (Sundar et al., 2021), calling for the need to investigate and test various warning designs using behavior-oriented theories, experimental methods, and diverse analytic lenses. At present, researchers are in pressing need to decipher why users fail to notice misinformation warnings while binging short videos disseminated by phenomenal social media entities. Recent studies have asserted that habituation—an individual-level nonassociative learning behavior where the response to a stimulus (e.g., a misinformation warning on TikTok) decreases—is responsible for the attenuated user attention to these warnings and constitutes a serious threat to security warning effectiveness (Vance et al., 2018).

However, the onset of habituation is a direct manifestation of *response malleability* to repeated stimulation (Groves & Thompson, 1970), suggesting that a response to an external stimulus becomes weaker due to *learning*, not fatigue or sensory adaptation (Çevik, 2014; Rankin et al., 2009). Drawing upon the Communication–Human Information Processing (C–HIP) model (Wogalter, 2018), we assert that once the memory of a security warning emerges, attention to the warning is reduced, and memory-based learning may help users conduct fact-checking of disputable short video content more efficiently. This proposition paved the way for our assessment of an interesting behavioral response in which the viewer exhibits the habituation of response but comprehends or remembers the warning message and proceeds to watch the video regardless. Our research shows that this "ignore the warning but remain vigilant" behavior is independent of the general behavioral decrement (e.g., warning adherence diminution) investigated by Vance et al. (2018).

Our study does not intend to reject the findings of previous works on habituation in the field of information systems (IS) security research (e.g., Anderson et al., 2016a; Vance et al., 2018). On the contrary, these works inspired us to study the effects of fact-checking through the lens of habituation theory. Contextual misinformation warnings, which do not interrupt users, nor do they require users to stop their current tasks, have been widely adopted by social media platforms to combat misinformation (Kaiser et al.

2021; Ling et al. 2022). Previous research (Anderson et al. 2016a, 2016c) found polymorphic warning design that repeatedly changes its appearance substantially reduced habituation of attention. However, it is unclear whether contextual polymorphic security warnings (CPSWs)—that is, repeatedly updating the appearance of contextual warnings—are still effective in countering misinformation on short-video-sharing platforms. Finally, this paper sheds light on the favorable efficacy of CPSWs over contextual static security warnings (CSSWs) in misinformation mitigation endeavors made by leading social media entities, such as TikTok. We aimed to address the following broad research questions:

RQ1: What behavioral manifestations does habituation produce when repeating CPSWs are presented to short-video viewers?

RQ2: How does habituation influence users' perceptions of the believability of videos on short-video-sharing platforms?

THEORY & HYPOTHESES

Our literature review assessed impactful and corresponding streams of research: digital stimulus design from the field of human–computer interaction (HCI), habituation contextualized in IS security, and cognitive processes from psychology, communication, and political science. Misinformation mitigation in social media is a complex task whose implications and insights rely on collaborative research endeavors from multiple fields of study to yield generalizable, multifaceted solutions.

First, the HCI community creates and evaluates stimulus designs (e.g., polymorphic security warnings) that influence users' behaviors to protect their security stances on the Internet where misinformation is proliferating (Distler et al., 2020; Kaiser et al., 2021). On the one hand, modern security warnings are very effective in thwarting security threats around 75–90% of the time (Akhawe & Felt, 2013). On the other hand, users remain vulnerable to adversaries who deliberately try to exploit victims' judgment errors (Liang et al., 2019).

Second, IS security researchers have regularly drawn upon the behavioral learning effects of habituation as a critical lens for elucidating why users frequently fail to heed security warnings when these messages are repeatedly presented to them, suggesting that habituation poses severe threats to the efficacy of security warnings (Vance et al., 2018). However, research on habituation in IS security is a developing stream. Much remains unknown with respect to the interplay between a repeated warning stimulus and a user's learning effect (Anderson et al., 2016a), calling for the need for more behavior-focused studies to unveil nuanced insights that can yield significant improvements in our understanding of the status quo.

Third, just as information security researchers care about the effectiveness of security warning messages that may facilitate misinformation mitigation, political science, communication, and psychology scholars have found similar importance in understanding the effectiveness of fact-checking messages to correct people's misconceptions that would reduce their competence in a democracy.

Our study extends on prior research by revealing a new behavioral manifestation (informed reduced behavior) that used to be categorized under the umbrella of overall reduced behavior as a direct outcome of the habituation of responses (e.g., Vance et al., 2018, p. 360). In other words, security warnings might be able to inform users' behaviors and guide them toward cybersecurity well-being, but these positive behavioral effects might result from, at least partially, the habituation of responses itself. Our focus is on short-video-sharing social media platforms on which it would be easy to spread misinformation, so we also investigate the extent to which a user believes the content in videos shared by other users. Figure 1 presents the premise of our study.

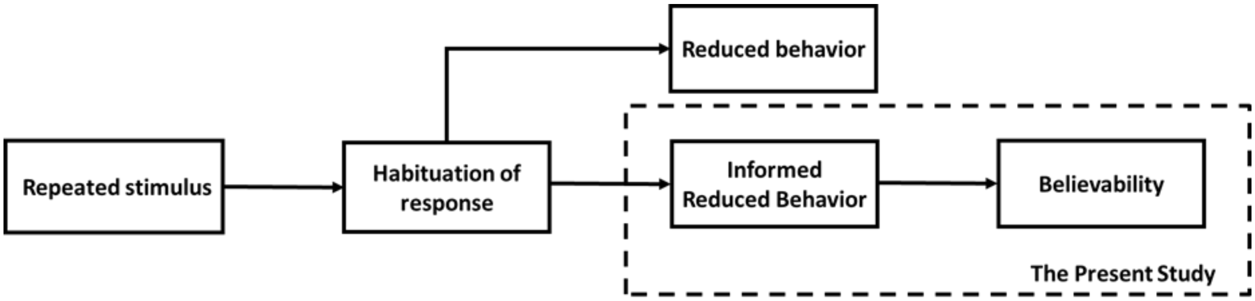

Figure 1. Extending Prior Research of Habituation in IS Security

Security Warning Stimulus Design

Friction, a design mechanism used to make changes to an interaction to elicit reflection, is often added to security warning designs to elicit increased responses from users (Gould et al., 2021). Security warning studies have found that actively interrupting people's workflows is more effective than using passive indicators to alert users (e.g., Egelman et al., 2008; Wu et al., 2006). Based on the level of friction, Kaiser et al. (2021) categorized misinformation warnings into two forms: the interstitial cover that requires interaction (i.e., click through) before the user can see the misinformation and contextual tags that only indicate that the news article contains misinformation but does not interrupt the users or require action to see misleading content. When the two formats were put through empirical tests, interstitial warnings produced more behavioral adherence and reduced the perceived accuracy of misinformation than contextual tags (Kaiser et al., 2021; Sharevski et al., 2022).

However, the interstitial design approach may not be the most appropriate for combating misinformation. This is because the misinformation warning is different from a security warning that aims to prevent people from the individualized and irremediable risks that are difficult for users to identify by themselves (Kaiser et al., 2021). Misinformation warnings can deal with a collective threat to democracy; therefore, one must also consider the broader legal and policy implications stemming from any direct intervention imposed by the government or platforms, as these warnings could prevent users from seeing content, which raises concerns about censorship (Lazer et al., 2018). Thus, it is difficult to justify the excessive use of interstitial warnings that have the potential to limit the freedom of speech.

This study focuses on contextual warnings' ability to label misinformation and provide contextual information without requiring users to take action to see the misinformation content (Kaiser et al., 2021). Previous studies that examined contextual misinformation warnings on social media (e.g., Twitter and Facebook) found that the use of specific label words or in other limited conditions had a moderate effect on reducing people's misconceptions (e.g., Clayton et al., 2020; Moravec et al., 2020; Pennycook et al., 2018). A recent study on TikTok videos found that the most commonly adopted contextual warning labels attached to videos with a COVID-19 hashtag contained only text-based general information (e.g., "learn more about COVID-19 vaccines" or "learn the facts about COVID-19") (Ling et al., 2022). In addition to the CSSWs commonly seen on TikTok, we also introduce the CPSWs, which contain more graphic elements that may increase the level of friction when encountered by users.

Habituation Theory

Habituation is a form of nonassociative learning that does not require an association between stimuli, such as sound and food in Pavlov's dog study (Groves & Thompson, 1970). Habituation is defined as a behavioral response decrement that results from repeated stimulation and does not involve sensory adaptation/fatigue or motor fatigue (Rankin et al., 2009, p. 2). In this sense, sensory adaptation, which also causes decreased reactivity to a stimulus due to repeated presentations of that stimulus, includes physiological effects (e.g., the sense of smell), while habituation is exhibited as behavioral learning effects (e.g., muscle contraction) (Rankin et al., 2009). Moreover, it is important to distinguish habituation from fatigue and ensure that the response decrement is not caused by tiredness. Therefore, researchers may choose to run a test of dishabituation, which is the presentation of another (usually strong) stimulus that results in the recovery of the habituated response (Groves & Thompson, 1970; Wagner, 1979). If the response reappears within a habituation session, the effects of fatigue can be ruled out because the person would still have been tired during the dishabituation test.

Groves and Thompson's (1970) seminal work proposed the idea of sensitization, an independent process that operates concurrently with habituation. Sensitization refers to the increasing strength of the reaction to a stimulus (Davis, 1974). Therefore, sensitization and habituation together can be presented as a dual-process theory that explains how people react to external stimuli via behavioral learning. Habituation and sensitization share the same underlying processes (Barbas et al., 2003); they both introduce a change in an individual's behavior due to repeated exposure to an external stimulus (e.g., warning signs), and they occur simultaneously when a person encounters a stimulus (Glanzman, 2009). However, sensitization is the opposite of habituation (Çevik, 2014), and it is a different concept from dishabituation. First, sensitization is a concurrent process that acts alongside habituation but has an opposite learning effect, whereas dishabituation is a separate, subsequent intervention that takes place after the initial habituation effect manifests. The purpose of dishabituation is to recover a response that has undergone habituation (Rankin et al., 2009). Understanding the implications of these concepts lays the foundation for our subsequent operationalization of the research design for habituation studies.

Another critical characteristic of habituation theory is stimulus specificity/generalization (Groves & Thompson, 1970; Rankin et al., 2009). We draw upon Characteristic 7 from Rankin et al. (2009, p. 4) to run a stimulus generalization test (the carryover of habituation from one stimulus to another novel stimulus), which is commonly mislabeled as the dishabituation test (the recovery of a response by encountering another strong stimulus) (Rankin et al., 2009; Vance et al., 2018), to compare the changes in responses to the habituated versus novel stimuli. This test allowed us to assess the strength of the habituated response resulting from repeated stimuli with different designs (e.g., contextual polymorphic vs. contextual static), enabling further evaluations of the efficacy of the warnings afforded by these designs.

H1: Users are more likely to habituate to CPSWs during an experimental period.

H2: Users habituate less to CPSWs that have more stimulus specificity than to CSSWs during an experimental period.

Habituation and Reduced Behavior

An important stream of IS security research is the interplay between security warning stimuli and user formation of habituation (and its ensuing reduced behaviors) (e.g., Anderson et al., 2016a, 2016b, 2016c; Sharevski et al., 2022). In their neurophysiological study, Vance et al. (2018) presented the full extent of the problem of habituation by systematically examining two characteristics of habituation: response attenuation and spontaneous recovery (Kim & Wogalter, 2009).

Following their steps, we study a specific behavioral learning effect resulting from habituation that occurs when users encounter repeated misinformation warning stimuli while browsing short videos on leading social media platforms (e.g., TikTok). This effect is manifested in some TikTok users who exhibit reduced eye gaze fixations on misinformation warnings and continue to watch the video knowing its potential for misinformation. More interestingly, these users seemed to have memorized and fully understood the warning content and to draw upon it while watching the disputable video, making them critical viewers. This effect echoes the work of Conzola and Wogalter (2011), who recognized that comprehending a warning's meaning is a crucial stage in the C–HIP model, a model security researchers typically apply to decipher user behavior (Kaiser et al., 2021). To our knowledge, no IS research has investigated this effect. Therefore, our inquiries in this work are based on these distinct but complementary theoretical pillars: memory/comprehension-based information processing and cognitive processes associated with fact-checking.

Vance et al. (2018) defined habituation as a general decline in participants' attention to warnings over time, although such attention can recover at least partially between sessions without exposure to warnings. While attention is being maintained on the warning, despite its attenuation over repeat presentations of a stimulus, other processes can occur concurrently, including memory formation and comprehension (Wogalter, 2018). After a memory is formed, the individual will not hold or maintain attention to the stimulus material (i.e., habituation; Vance et al., 2018; Wogalter, 2018), which leads to a mechanism that reduces user behavioral response (Vance et al., 2018). Nonetheless, a nuanced distinction emerges between behavioral response attenuation with memory and that without memory. With the information stored in the memory that accompanies the reduced behavior, individuals may still exhibit habituation. However, their successive decreased warning adherence behaviors (e.g., seemingly ignoring the warning message and proceeding to watch a video that is marked as suspicious) are *informed*, potentially equipping them with sensible decision-making capabilities that protect their security postures and ward off adversaries. Given this nuanced distinction, we refer to this form of reduced behavioral response as *informed reduced behavior*. Thus, we hypothesize that information obtained from the habituation of responses due to previous repeated exposures to warnings may be acquired to form new memories, especially when the language used in a warning is the language the individual understands or the substance of a warning is not foreign to the user.

H3: Users are likely to develop informed reduced behavior from the habituation of response.

Fact-Checking Effects: Gaps in the Literature

Fact-checking can be conceptualized as "the practice of systematically publishing assessments of the validity of claims made by public officials and institutions with an explicit attempt to identify whether a claim is factual" (Walter et al., 2020, p. 351). In operationalization terms, fact-checking can be examined as a message's ability to reduce the believability of misinformation (Walter et al., 2020) and the persistence of misconceptions in people's memory (Chan et al., 2017). Although the past decade has seen a growth in the number of professional fact-checkers and the reinvigoration of the news industry's claim that it provides explanatory and analytical reporting (Graves, 2016), empirical research has yielded very divided findings regarding fact-checking's effectiveness. By reviewing this body of literature, we highlight the gaps in the theoretical and methodological approaches that have been used and suggest that studies on habituation could close the gaps by exploring the uncharted dimensions of fact-checking's effects and addressing the challenges of fact-checking on short-video-sharing platforms.

Political science studies have suggested that a fact-checking message only has a limited effect in correcting people's inaccurate beliefs because people tend to engage in goal-directed information processing that reinforces their existing ideological views (e.g., Garrett et al., 2013; Nyhan et al., 2013; Nyhan & Reifler, 2010). Messages correcting misinformation that counter people's ideological views can even backfire and strengthen their false beliefs (Nyhan & Reifler, 2010). The weak effectiveness of fact-checking has been found to be particularly salient in the political context compared to the marketing and health contexts (Walter & Murphy, 2018). These cognitive studies assumed that people actively engage in a reasoning process informed by the new information provided by the fact-checking message and motivated by their existing views when they evaluate the truthfulness of media content. Yet, the lack of such cognitive effort has been pointed out by researchers who found that people temporarily improved their belief accuracy with the corrective information but did no extra work to come up with counterarguments, as the backfire hypothesis indicated even in the context of controversial topics (Wood & Porter, 2019). This cognitive process can be intentionally induced by experimental procedures that ask participants to generate explanations in line with or counter misinformation, which could subsequently reduce or enhance the acceptance of the fact-checking message (Chan et al., 2017). Therefore, it is doubtful to what extent participants will engage in the elaborative cognitive process laid out by these studies in their everyday encounters with fact-checking information. In reality, people's attention to a warning message is constantly competing against other ongoing tasks and the internal processing of information not based on stimuli (Wogalter, 2018). On short-video-sharing platforms, when people's attention is preoccupied with processing the video content and other tasks (e.g., liking, sharing, or commenting), it is unreasonable to expect that full-fledged cognitive processes will constantly take place when evaluating the truthfulness of content.

The characteristics of the fact-checking message can also predict its success in reducing misconceptions. Studies have found that the truth scale (e.g., Amazeen et al., 2018), messages containing more contextual details (e.g., Swire et al., 2017), simple messages (as opposed to linguistically complex messages), and messages that refute an entire statement (rather than correcting part of the statement) can yield more desirable results to correct misinformation (Lewandowsky et al., 2012). The correction message was presented as a short paragraph inserted in a newspaper article in previous studies (e.g., Nyhan & Reifler, 2010; Thorson, 2016; Wood & Porter, 2019). In a lab session, when participants were instructed to read the message carefully, these correction messages were still too subtle to yield a result (Weeks, 2015). In other studies, a full-length correction message written by third-party fact-checkers was examined for its effectiveness in reducing misconceptions (e.g., Amazeen et al., 2018). However, users were not forced to read the full-length fact-checking message after seeing the simple warning label displayed on the social media platform; instead, they needed to take an extra step by clicking on the link to access the third-party fact-checker's articles. When fact-checking labels were presented in formats that were representative of their real-world analogues on social media, their effects were only measured after exposure of one time (e.g., Moravec et al., 2020; Sharevski et al., 2022). Meanwhile, repeated exposure (i.e., one prior exposure) to misinformation headlines on Facebook was found to increase its perceived accuracy, which could not be eliminated or reduced by the presence of the fact-checking label (Pennycook et al., 2018). Thus, it is also important to examine whether repeated exposure to fact-checking labels would generate a compounding effect beyond the specific piece of content being labeled on a short-video-sharing platform.

The big data approach was used to examine fact-checking's effects on the macro-level diffusion of misinformation in a social network (Friggeri et al., 2014; Shin et al., 2017) and the specific social contingency that affects how users accept or reject the correction (Margolin et al., 2018). By capturing the unmediated consequences of fact-checking via behavioral indicators (e.g., shares and likes) and user-generated content (e.g., comments), these studies overcame the biases associated with lab-induced cognitive processes. However, this is not to say that they perfectly reflected reality. In fact, observations of social media users' engagement (e.g., liking and resharing) were insufficient to evaluate fact-checking's effects. A recent study found that Tweets with warning labels received more engagements than other Tweets posted by the same user (Zannettou, 2021). Users' interactions with labeled content also exhibited large variations, from debunking false claims and mocking the author or content to reinforcing or resharing false claims. In addition, when the results of fact-checking were directly manifested in the social media content (e.g., replies and comments), the study could not capture the average response from the majority of users who never explicitly express their acceptance or rejection of the written content.

The gaps identified in the existing literature are not intended to reject prior findings. On the contrary, they inspired us to study fact-checking's effects from an unexplored dimension through the lens of habituation theory, which has the potential to fill the gaps in the literature and provide more suitable models to study short-video-sharing platforms. First, the habituation approach indicates that once the memory of a warning is formed, the warning will not be able to maintain an individual's attention (Wogalter, 2018). Therefore, it helps to examine the effects of fact-checking on short-video-sharing platforms, which may not cause users to go through the extensive analytical thinking process suggested by motivated reasoning studies. Second, habituation reveals the effects generated by repeated exposure to warnings. On short-video-sharing platforms, the contextual details of fact-checking messages are limited by the available space and the user's potential selective exposure to it. We could still examine the effectiveness of a simple and repeated warning label's compounding effect on people's perceptions of misinformation. Lastly, habituation can be observed through the neuropsychological measurement of users' attention and behavioral adherence to a warning, which can capture the understated psychological processes that are not manifested in users' engagement with or content on short-video-sharing platforms. If users who have been habituated to the warnings indeed engage in the informed reduced behavior, it is important to examine how people's trust in the content is influenced. As such, Figure 2 shows the conceptual model and all the above-described hypotheses.

H4: User's informed reduced behavior reduces the believability of video content.

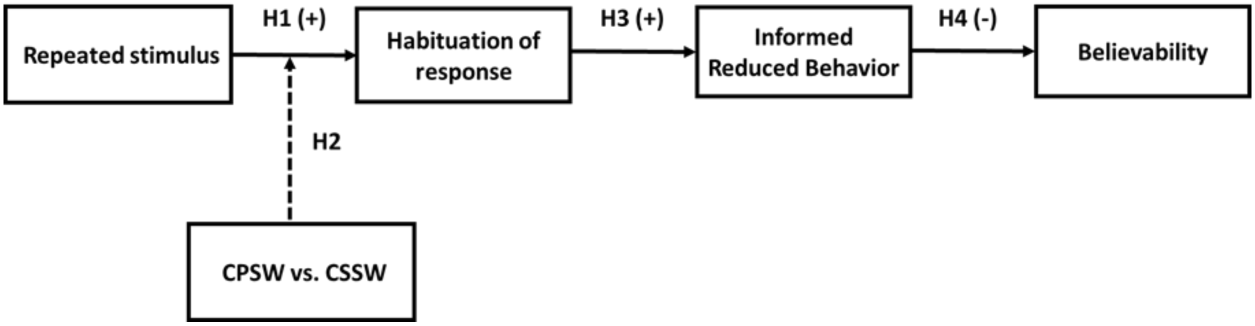

Figure 2. Conceptual Model and Hypotheses

METHOD

Pre-study

We first began with semi-structured interviews to examine how participants would process and react to contextual misinformation warnings and interstitial misinformation warnings on short-video sharing platforms, as well as how warning designs affect the perceived believability of videos. The length of the interview was on average about one hour. The study was approved by our institution's IRB.

Recruitment and Participants

We recruited 28 participants who frequently used short-video sharing platforms via convenience and snowball sampling. All the participants stated that they used TikTok and/or Instagram Reels for at least 30 minutes every day. 20 female (71%) and eight male (29%) participants who were between 18 and 29 years old (mean 22) took part in the study. The group consisted of 20 undergraduate students and one graduate student with a mix of majors and seven professionals. 19 participants (50%) were Democrats, two (14%) were Republicans, and seven (36%) were independent. On average, the sample broadly shares similar socio-economic backgrounds and was more educated than the general population in the United States.

Materials

To enhance research validity and observe participants interacting with different types of warnings in short-form videos, we built a short-video sharing app with similar TikTok features using React Native (Facebook and Community, 2015) and Expo (Expo Team, 2013). After opening the app, participants immediately saw curated featured videos on the home page, which mirrored the sort of video feed users might encounter every day while using their TikTok accounts (Figure 3).

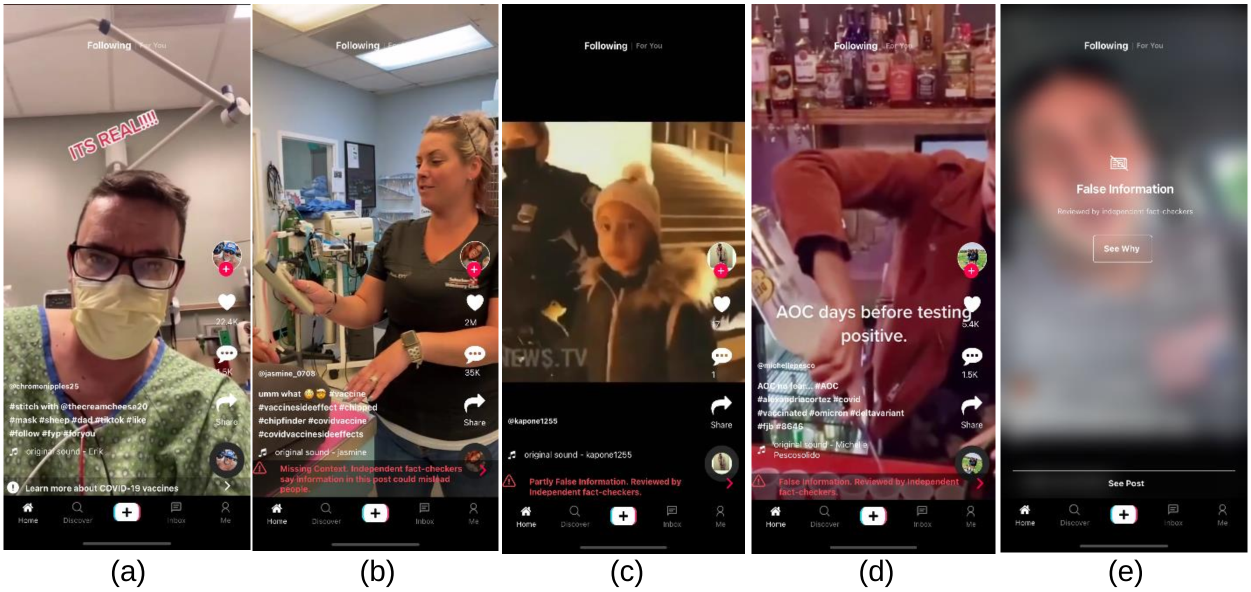
Following | For You
ITS REAL!!!!
Learn more about COVID-19 vaccines
Missing Context. Independent fact-checkers say information in this post could mislead people.
Partly False Information. Reviewed by independent fact-checkers.
AOC days before testing positive.
False information. Reviewed by independent fact-checkers.
False Information
Reviewed by independent fact-checkers
See Why
See Post
Home
Discover
Inbox
Me
Share
(a)
(b)
(c)
(d)
(e)

Figure 3: Videos are shown to participants with different label stimuli applied: (a) a general contextual warning: “learn more about COVID-19 vaccines”. (b) a specific contextual warning: “Missing Context. Independent fact-checkers say information in this post could mislead people”. (c) a specific contextual warning: “Partly False Information. Reviewed by independent fact-checkers”. (d) a specific contextual warning: “False Information. Reviewed by independent fact-checkers”. (e) an interstitial warning: “False Information. Reviewed by independent fact-checkers”. Participants can click “See Why” or “See Post”.

Eight videos were manually selected from among those debunked by PolitiFact (2007) and those reported in the screener survey as stories encountered by participants on TikTok and/or Instagram Reels. We focused on COVID-19 videos because participants reported they have frequently come across COVID-19 related misinformation on TikTok and/or Instagram Reels in the screener survey and those platforms have applied a broad or more specific mechanism for moderating videos with warning labels (Geeng et al. 2020). Based on the current approaches that have been applied by short-video sharing platforms to counter misinformation, we come up with three warning designs in our study (see Figure 3):

(1) General contextual warning: “Learn more about COVID-19 Vaccines” (Figure 3a).

(2) Specific contextual warning: “Missing Context” (Figure 3b), “Party false information” (Figure 3c), or “False information” (Figure 3d). The warning labels appear as a banner at the bottom of the content with bright red texts and click through options to see why “Independent fact-checkers say information in this post could mislead people”, “Independent fact-checkers say this information has no basis in fact”, or “Independent fact-checkers say this information has some factual inaccuracies”.

(3) Blurred interstitial warning: block and blur a misinformation video before people can see it (Figure 3e). There are two click through options on the blurred interstitial warning: tap “See Why” to view the fact-checking organization and why they identified the post as false information, and tap ”See Post” to view the video.

The videos we chose provide a mixture of real and fake videos. We carefully selected videos with different lengths ranging from 12 seconds (s) to 2:21 minutes (min), different difficulty levels from obviously mundane to the unusual news, and different interest levels from low to high. Researchers in this study ranked difficulty levels and interest levels individually and then reached a consensus in teams.

Procedure

Each participant saw six videos: one fake video with a general contextual warning, one fake video with a specific contextual warning, one fake video with a blurred interstitial warning, one fake video without any warnings, and two real videos without any warnings. We randomized video examples to control for ordering effects. Three specific contextual warning labels—"Missing Context", "False information", and "Party false information"—are randomized in our app too. When watching each video, participants were encouraged to verbalize their initial thoughts on each video and any feelings and comments that they might be thinking in their minds. Once completing the think-aloud activity, each participant was invited to take part in a further semi-structured interview, asking follow-up questions on the credibility rating of each video, how to determine the video's credibility, and any related issue that cropped up during the activity.

Each session was audio recorded and subsequently transcribed, with a screen recording capturing scrolling and mouse movements, and page navigation. We followed an iterative and open coding process (Strauss & Corbin, 1997) to analyze interview data and think-aloud videos.

Results

Habituation in contextual warnings: Almost half of the participants reported they were aware of the general contextual warning but didn't click through it when watching the video. According to an in-depth qualitative study performed by Ling et al. (2022), TikTok broadly applies warning labels on videos that include #coronavirus. The fact that all COVID-19 related videos include warning labels may cause users to ignore such warning labels and pay less attention to them.For the "Missing Context" warning and "Partly False" warning, six out of 10 participants (60%) and four out of nine participants (44%) chose to ignore the warning and didn't click through them, whereas only one participant (11%) didn't click through the "False" warning. Similar to TikTok findings, we observed participants habituated to warnings if they frequently encountered videos flagged with warning labels. Habituation may decrease users' attention to warnings and thus reduce the warning effects. This is inline with findings from Amer and Maris (2007), Brustoloni and Villamarin-Salomon (2007), and Vance et al. (2017, 2018) which consistently pointed to habituation as an important reason why users ignored security warnings. In our main study, we are interested in the interplay between warning stimulus and user habituation formation, in particular, how habituation manifests in behavioral responses given the presence of different contextual warnings.

Use warnings to aid credibility evaluation: The majority of participants considered using the specific contextual warnings and blurred interstitial warnings to access the credibility of video content and thought those warnings were very effective. However, participants barely mentioned using the general contextual warning to aid judgment. We observed that although participants didn't click through warnings, their exposure to the warning labels made them disbelieve the post initially, as one participant stated "*which made me immediately think that it was going to be fake*". A few participants pointed out that they would believe the video was real if it didn't have a warning. Since participants only encountered each warning once, we are interested to find out how users' responses change over time as they repeatedly encounter fact check warnings. Furthermore, we want to investigate how different types of warning designs affect users' believability of video content.

User acceptance and preference of warnings: The majority of participants thought the general contextual warning—"Learn more about COVID-19 Vaccines"—was the least effective approach to debunk misinformation. Almost half of them chose interstitial warning to be the most effective approach and the other half chose contextual warning. Of the group who favored blurred interstitial warning, they explained "*it just catches your eye more and stops you even before you watch the video*". However, some participants also highlighted that although the interstitial warning was very effective and informative, they didn't like how it interrupted their watch flow. That's also the main reason why the other half favored contextual warnings: "*it can let the user know that there's misinformation without disrupting the video watching aspect of TikTok*". Given that the interstitial warning was assessed rather critically by our participants and contextual warnings have already been deployed by short-video sharing platforms (Morrow et al., 2022), we chose to focus on contextual warnings in the next stage.

Our interview study exposed the potential interplay among contextual warning, habituation, and believability. Based on the pre-study's findings, the main study will look more detailed into their relationship and test our hypotheses in Section 2.

Measures to be Used in the Main Study

We will measure habituation from both neurophysiological and behavioral perspectives. Eye-tracking and mouse cursor tracking data will be collected and analyzed to explore whether users paid greater attention to CPSWs over time compared to CSSWs. Eye-tracking has been widely used to measure human visual activities and the mental process of habituation to security warnings (Anderson et al. 2016a, 2016c; Vance et al. 2018). As one of the neurophysiological manifestations of habituation, Eye Movement-based Memory (EMM) effect is evident in fewer eye-gaze fixations and less visual sampling of the regions of interest within the visual stimulus (Heisz and Shore 2008; Ryan et al. 2000). Mouse cursor tracking, as another important unobtrusive instrument, has been used to measure users' attention, especially changes in attention (Navalpakkam and Churchill, 2012; Rodden et al. 2008). Numerous studies have chosen mouse cursor tracking to measure habituation in response to repeated security warnings (Anderson et al. 2016c; Vance et al. 2019). In this study, we will use both eye-tracking (i.e., eye movements) and mouse cursor tracking (i.e., how a person responds to warnings using their phone) to provide a more holistic view of habituation to contextual warnings.

Our pre-study found that when encountering a specific contextual warning that provided explicit information about credibility, people tend to use the warning to verify the post whereas diminished attention or decreased response may be observed. In other words, people may develop informed reduced behavior (less eye-gaze fixations and less clickthrough) to contextual warnings. To measure informed reduced behavior, we will use indirect measures including self-reported and subjective opinions on the perceived effectiveness of a warning. Participants will be asked to say out loud how they make credibility judgments when watching a video. Researchers will avoid prompting participants to think about warnings during the activity. If a participant pays less attention to a warning but still mentions the use of warning to verify content while thinking aloud, we will categorize it as informed reduced behavior.

Believability will be measured by three self-reported items on a seven-point Likert scale adapted from Beltramini (1988): How believable do you find this video? How truthful do you find this video? How credible do you find this video? During the debrief session, participants will be directed to rate the perceived believability of each video.

Experiment Design in the Main Study

The empirical study will follow a within-subject design, encompassing seven CPSWs and one CSSW. Participants will be randomly assigned to view either CPSWs or the CSSW for a workweek of five days. Our pre-study reveals that a specific "false information" warning is not only effective but also favored by participants. Therefore, we use it as a CPSW to control habituation. Through thoroughly reviewing the polymorphic warning literature (Anderson et al. 2015, 2016a, 2016c; Vance et al. 2018), we developed seven graphical variations of fact check warnings by adding animation, changing the background color and text sizes, and adding different symbols. All participants will be shown a total of 20 videos in the TikTok format. We will select five videos with accurate news stories from mainstream media or verified by fact-checkers and 15 videos with inaccurate news stories that have been debunked by fact-checkers. The video content covers different categories (e.g., entertainment, education, food, travel, etc). Of 15 Videos with inaccurate news stories, 10 videos will be randomly selected and evenly assigned between the CPSW and CSSW treatments, and the others will be displayed as regular TikTok posts. During the experiment, both CPSWs and the CSSW will be repeated 5 times. For the CPSWs treatment, we will randomize contextual polymorphic variations. It may take approximately 30 minutes for each participant to watch all the videos every day.

During the experiment, we will use an eye tracker to track participants' eye movements while watching each video. By analyzing the number of fixations over the entire warning, we will be able to gauge whether the EMM effect occurs during participants' visual processing of warnings, thus, in turn, demonstrating the extent to which participants habituate to warnings over repeated exposures. We will use React Native (Facebook and Community, 2015) and Expo (2013) to develop a similar TikTok environment, which can also collect the x, y coordinates and timestamp of each mouse movement and save mouse cursor movement data in our database for further analysis. Our app will record where participants are hovering, clicking, scrolling, and pausing through each page. Participants will be tasked to find news stories that are not fully accurate and think aloud about how they make credibility judgments. This will help us determine whether or not informed reduced behavior occurs during the experiment. To ensure realism and research validity, participants can control how long they want to watch each video and what they want to explore on each page.

After watching all the videos and viewing all the warnings, participants will take part in a debrief session to rate the credibility of each video they watch and elaborate on their judgment. We will ask participants to rate the perceived believability of those videos using a seven-point Likert scale. Since participants will recall their judgments and for which reasons, believability ratings may be affected by other factors such as memory capacity. We give participants an option to re-watch videos if they couldn't recall what has been viewed. In addition to self-report measures, we will also use think-aloud data to complement and validate believability ratings.

DISCUSSION

Consistent with prior IS security research on warning stimulus and habituation (Anderson et al., 2015; Anderson et al., 2016a, 2016b, 2016c; Kaiser et al., 2021; Sharevski et al., 2022; Vance et al., 2018), stimulus, habituation of response, and reduced behavior are found to manifest on short-video-sharing platforms that prefer contextual to interstitial warnings. More importantly, we found that reduced behavior, as a concept, might be too broad to explain the detailed and subtle user responses to contextual warnings on platforms such as TikTok. This finding suggests that when users notice the existence of security warnings, they tend to develop memory/comprehension to facilitate a cognitive process that copes with repeated exposure to warnings. This process allows them to quickly draw upon the warning information stored in their memories, resulting in attenuated attention to warnings. However, they still cautiously handle suspicious videos and know that they are at least partially untrue. Our research showed that this "ignore the warning but remain vigilant" behavioral manifestation is independent of general behavioral decrement (e.g., warning adherence diminution) demonstrated by Vance et al. (2018).

This study contributes to the evolving research on habituation in IS security by (1) theorizing and testing the notion that habituation of response resulting from repeated presentations of a stimulus may give rise to a learning effect associated with memory/comprehension formation, among other effects, such as reduced behavior that is not associated with memory/comprehension formation; (2) examining the possibility that this memory-driven learning effect in the form of informed reduced behavior lowers the credibility or believability of suspicious short videos, enabling and intensifying voluntary protection at the individual level; (3) deepening knowledge of the behavioral learning underpinnings of ineffectual warning adherence, with a focus on habituation of response; and (4) using contextual warnings, a low-friction stimulus that affords distinct behavioral responses compared to high-friction interstitial warnings adopted in prior research, to expound on the interplay between stimulus and habituation in the context of video-sharing media platforms.

To fulfill these contributions, we built on HCI design principles and habituation theories, particularly those by Groves and Thompson (1970) and Rankin et al. (2009), and adopted an eye tracking-based view of habituation to test our hypotheses (Vance et al., 2018). The initial experimental data showed patterns that support our assertions. More detailed empirical findings will be offered in future studies.

Another noticeable stream of research our study relates to is the recent development of fake news terminology that has increasingly drawn attention from IS scholars, such as Khan et al. (2022) and Kapantai et al. (2022) who have questioned the rigor of the terminology and identified three problems with the concept of fake news, namely, not a clear-cut concept with distinguishing characteristics and dimensions. Although this study does not offer a remedy to the terminology predicament, it adds to the discourse that calls for a taxonomy of definitions to guide subsequent research.

CONCLUSION

In this early-stage study, we proposed to investigate the positive effects of informed reduced behavior on security warning efficacy in the context of short-video-sharing social media, which is plagued by misinformation despite organizations' resilient efforts to curb it. However, a broader challenge for service providers lies in the tension between IS security and user experience—how to secure virtual communities without undermining the user experience, or what is the best way to guide users' behaviors and deter the spread of misinformation without stomping on freedom of speech. In this sense, informed reduced behavior extends the conventional belief that habituation threatens security warning efficacy by presenting a nuanced view of habituation, a universal learning behavior that, once fully understood, can empower netizens across the board with heightened judiciousness that protects them and, at the same time, electrifies their virtual experiences.

REFERENCES

Akhawe, D., & Felt, A. P. (2013). Alice in wonderland: A large-scale field study of browser security warning effectiveness. *22nd USENIX Security Symposium, USENIX Security*, 13, 257-272. https://dl.acm.org/doi/10.5555/2534766.2534789

Amazeen, M. A., Thorson, E., Muddiman, A., & Graves, L. (2018). Correcting political and consumer misperceptions: The effectiveness and effects of rating scale versus contextual correction formats. *Journalism & Mass Communication Quarterly*, *95*, 28–48. https://doi.org/10.1177/1077699016678186

Amer, T. S., & Maris, J. M. B. (2007). Signal words and signal icons in application control and information technology exception messages—Hazard matching and habituation effects. *Journal of Information Systems*, 21(2), 1-25. https://doi.org/10.2308/jis.2007.21.2.1

Anderson, B. B., Kirwan, C. B., Jenkins, J. L., Eargle, D., Howard, S., & Vance, A. (2015, April). How polymorphic warnings reduce habituation in the brain: Insights from an fMRI study. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2883-2892. https://doi.org/10.1145/2702123.2702322

Anderson, B. B., Jenkins, J. L., Vance, A., Kirwan, C. B., & Eargle, D. (2016a). Your memory is working against you: How eye tracking and memory explain habituation to security warnings. *Decision Support Systems*, *92*, 3-13. https://doi.org/10.1016/j.dss.2016.09.010

Anderson, B.B., Vance, A., Kirwan, Eargle, D., & Jenkins, J. L. (2016b). How users perceive and respond to security messages: a NeuroIS research agenda and empirical study, *European Journal of Information Systems*, *25*(4), 364-390, https://doi.org/10.1057/ejis.2015.21

Anderson, B. B., Vance, A., Kirwan, C. B., Jenkins, J. L., & Eargle, D. (2016c). From warning to wallpaper: Why the brain habituates to security warnings and what can be done about it. *Journal of Management Information Systems*, *33*(3), 713-743. https://doi.org/10.1080/07421222.2016.1243947

Barbas, D., DesGroseillers, L., Castellucci, V. F., Carew, T. J., & Marinesco, S. (2003). Multiple serotonergic mechanisms contributing to sensitization in Aplysia: Evidence of diverse serotonin receptor subtypes. *Learning & Memory*, *10*(5), 373-386. https://doi.org/10.1101/lm.66103

Beltramini R. F. (1988) Perceived believability of warning label information presented in cigarette advertising. *Journal of Advertising*, *17*(2):26–32. https://doi.org/10.1080/00913367.1988.10673110

Brustoloni, J. C., & Villamarín-Salomón, R. (2007, July). Improving security decisions with polymorphic and audited dialogs. *Proceedings of the 3rd Symposium on Usable Privacy and Security*, 76-85. https://doi.org/10.1145/1280680.1280691

Çevik, M. Ö. (2014). Habituation, sensitization, and Pavlovian conditioning. *Frontiers in Integrative Neuroscience*, *8*, 13. https://doi.org/10.3389/fnint.2014.00013

Chan, M. P. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*, 1531–1546. https://doi.org/10.1177/0956797617714579.

Clayton, K., Blair, S., Busam, J. A., et al. (2020). Real solutions for fake news? Measuring the effectiveness of general warnings and fact-check tags in reducing belief in false stories on social media. *Political Behavior, 42*(4), 1073-1095. https://doi.org/10.1007/s11109-019-09533-0

Conzola, V. C. and Wogalter, M. S. (2011). A Communication–Human Information Processing (C–HIP) Approach to Warning Effectiveness in the Workplace. *Journal of Risk Research*, 4 (4 Apr. 15, 2011). DOI: 10.1080/13669870110062712.

Davis, M. (1974). Sensitization of the rat startle response by noise. *Journal of Comparative and Physiological Psychology*, *87*(3), 571-581. https://doi.org/10.1037/h0036985

Distler, V., Lenzini, G., Lallemand, C., & Koenig, V. (2020, October). The framework of security-enhancing friction: How UX can help users behave more securely. *New security paradigms workshop*, 45-58.

Egelman, S., Cranor, L. F., & Hong, J. (2008, April). You've been warned: an empirical study of the effectiveness of web browser phishing warnings. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1065-1074. https://doi.org/10.1145/1357054.1357219

Expo Team. (2013). *Expo* (Version 45.0.0). https://expo.dev/

Facebook and Community. (2015). *React Native* (Version 0.68). Meta Platforms, Inc. https://reactnative.dev/

Friggeri, A., Adamic, L. A., Eckles, D., & Cheng, J. (2014). Rumor cascades. *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 101–110. Retrieved http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/download/8122/8110

Garrett, R. K., Nisbet, E. C., & Lynch, E. K. (2013). Undermining the corrective effects of media-based political fact checking? The role of contextual cues and naïve theory. *Journal of Communication*, *63*, 617–637. https://doi.org/10.1111/jcom.12038

Geeng, C., Francisco, T., West, J., & Roesner, F. (2020). Social media COVID-19 misinformation interventions viewed positively, but have limited impact. *arXiv preprint*. https://doi.org/10.48550/arXiv.2012.11055

Glanzman, D. L. (2009). Habituation in Aplysia: the Cheshire cat of neurobiology. *Neurobiology of Learning and Memory*, *92*(2), 147-154. https://doi.org/10.1016/j.nlm.2009.03.005

Gould, S. J., Chuang, L. L., Iacovides, I., Garaialde, D., Cecchinato, M. E., Cowan, B. R., & Cox, A. L. (2021, May). A special interest group on designed and engineered friction in interaction. *Extended Abstracts of the 2021 Conference on Human Factors in Computing Systems*, 1-4.https://doi.org/10.1145/3411763.3450404

Graves, L. (2016). *Deciding what's true: The rise of political fact*-checking in American journalism*.* New York, NY: Columbia University Press.

Groves, P. M., and Thompson, R. F. (1970). Habituation: A dual-process theory, *Psychological Review* (77), 419-450 .https://doi.org/10.1037/h0029810

Heisz, J. J., and Shore, D. I. (2008). More efficient scanning for familiar faces, *Journal of Vision*, 8(1), 9 https://doi.org/10.1167/8.1.9

Kaiser, B., Wei, J., Lucherini, E., Lee, K., Matias, J. N., & Mayer, J. (2021). Adapting security warnings to counter online disinformation. *30th USENIX Security Symposium*, USENIX Security 21, 1163-1180. Retrieved https://www.usenix.org/system/files/sec21-kaiser.pdf

Kapantai, E., Christopoulou, A., Berberidis, C., & Peristeras, V. (2021). A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society*, *23*(5), 1301-1326.

Khan, A., Brohman, K., & Addas, S. (2022). The anatomy of 'fake news': Studying false messages as digital objects. *Journal of Information Technology*, *37*(2), 122-143.

Kim, S., & Wogalter, M. S. (2009). Habituation, Dishabituation, and Recovery Effects in Visual Warnings. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *53*(20), 1612–1616. https://doi.org/10.1177/154193120905302015

Lazer, D. M., Baum, M. A., Benkler, Y., et al (2018). The science of fake news. *Science*, *359*(6380), 1094-1096.

Lewandowsky, S., Ecker, U. K. H., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*,106–131. http://dx.doi.org/10.1177/1529100612451018

Liang, H., Xue, Y., Pinsonneault, A., & Wu, Y. A. (2019). What users do besides problem-focused coping when facing IT security threats: An emotion-focused coping perspective. *MIS Quarterly*, *43*(2), 373-394. https://doi.org/10.25300/MISQ/2019/14360

Ling, C., Gummadi, K. P., & Zannettou, S. (2022). Learn the facts about COVID-19: Analyzing the use of warning labels on TikTok videos. *arXiv preprint*. https://doi.org/10.48550/arXiv.2201.07726

Margolin, D. B., Hannak, A., & Weber, I. (2018). Political fact-checking on Twitter: When do corrections have an effect? *Political Communication*, *35*(2), 196-219. https://doi.org/10.1080/10584609.2017.1334018

Moravec, P. L., Kim, A., & Dennis, A. R. (2020). Appealing to sense and sensibility: System 1 and system 2 interventions for fake news on social media. *Information Systems Research*, *31*(3), 987-1006. https://doi.org/10.1287/isre.2020.0927

Morrow, G., Swire-Thompson, B., Polny, J. M., Kopec, M., & Wihbey, J. P. (2022). The emerging science of content labeling: Contextualizing social media content moderation. *Journal of the Association for Information Science and Technology*. https://doi.org/10.1002/asi.24637

Navalpakkam, V., & Churchill, E. (2012, May). Mouse tracking: measuring and predicting users' experience of web-based content. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2963-2972. https://doi.org/10.1145/2207676.2208705

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*, 303–330. https://doi.org/10.1007/s11109-010-9112-2

Nyhan, B., Reifler, J., & Ubel, P. A. (2013). The hazards of correcting myths about health care reform. *Medical Care*, *51*(2), 127–132. https://doi.org/10.1097/MLR.0b013e318279486b

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*(12), 1865–1880. https://doi.org/10.1037/xge0000465

Poynter Institute. (2007). *PolitiFact*. https://www.politifact.com/

Rankin, C. H., Abrams, T., Barry, R. J., Bhatnagar, S., Clayton, D. F., Colombo, J., Coppola, G., Geyer, M. A., Glanzman, D. L., Marsland, S., McSweeney, F. K., Wilson, D. A., Wu, C. F., and Thompson, R. F. (2009). Habituation revisited: An updated and revised description of the behavioral characteristics of habituation. *Neurobiology of Learning and Memory*, *92*(2), 135-138.

Reeder, R. W., Felt, A. P., Consolvo, S., Malkin, N., Thompson, C., & Egelman, S. (2018, April). An experience sampling study of user reactions to browser warnings in the field. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1-13. https://doi.org/10.1145/3173574.3174086

Rodden, K., Fu, X., Aula, A., & Spiro, I. (2008). Eye-mouse coordination patterns on web search results pages. *CHI'08 Extended Abstracts on Human Factors in Computing Systems*, 2997-3002. https://doi.org/10.1145/1358628.1358797

Ryan, J. D., Althoff, R. R., Whitlow, S., & Cohen, N. J. (2000). Amnesia is a deficit in relational memory. *Psychological Science*, *11*(6), 454-461. https://doi.org/10.1111/1467-9280.00288

Sharevski, F., Alsaadi, R., Jachim, P., & Pieroni, E. (2022). Misinformation warnings: Twitter's soft moderation effects on covid-19 vaccine belief echoes. *Computers & Security*, 114, https://doi.org/10.1016/j.cose.2021.102577

Shin, J., Jian, L., Driscoll, K., & Bar, F. (2017). Political rumoring on Twitter during the 2012 US presidential election: Rumor diffusion and correction. *New media & society, 19*(8), 1214-1235. https://doi.org/10.1177/1461444816634054

Strauss, A., & Corbin, J. M. (1997). *Grounded theory in practice*. Sage.

Sundar, S. S., Molina, M. D., & Cho, E. (2021). Seeing is believing: Is video modality more powerful in spreading fake news via online messaging apps? *Journal of Computer-Mediated Communication*, *26*(6), 301–319. https://doi.org/10.1093/jcmc/zmab010

Swire, B., Ecker, U. K. H., & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 43*(12), 1948–1961. https://doi.org/10.1037/xlm0000422

Thorson, E. (2016). Belief echoes: The persistent effects of corrected misinformation. *Political Communication, 33*, 460–480. https://doi.org/10.1080/10584609.2015.1102187

Vance, A., Kirwan, B., Bjornn, D., Jenkins, J., & Anderson, B. B. (2017, May). What do we really know about how habituation to warnings occurs over time? A longitudinal fMRI study of habituation and polymorphic warnings. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems,* 2215-2227. https://doi.org/10.1145/3025453.3025896

Vance, A., Jenkins, J. L., Anderson, B. B., Bjornn, D. K., & Kirwan, C. B. (2018). Tuning out security warnings: A longitudinal examination of habituation through fMRI, eye tracking, and field experiments. *MIS Quarterly, 42*(2), 355-380. https://doi.org/10.25300/MISQ/2018/14124

Vance, A., Eargle, D., Jenkins, J. L., Kirwan, C. B., & Anderson, B. B. (2019). The fog of warnings: how non-essential notifications blur with security warnings. *Fifteenth Symposium on Usable Privacy and Security,* 407-420.

Wagner, A.R.(1979). Habituation and memory. In A. Dickinson & R. A.Boakes (Eds.), *Mechanisms of learning and motivation: A memorial volume for Jerry Konorski.* Lawrence Erlbaum Associates.

Walter, N., Cohen, J., Holbert, R. L., & Morag, Y. (2020). Fact-checking: A meta-analysis of what works and for whom. *Political Communication*, 37(3), 350-375. https://doi.org/10.1080/10584609.2019.1668894

Walter, N., & Murphy, S. T. (2018). How to unring the bell: A meta-analytic approach to correction of misinformation. *Communication Monographs*, *85*(3), 423-441. https://doi.org/10.1080/03637751.2018.1467564

Weeks, B. E. (2015). Emotions, partisanship, and misperceptions: How anger and anxiety moderate the effect of partisan bias on susceptibility to political misinformation. *Journal of Communication*, *65*(4), 699–719. https://doi.org/10.1111/jcom.12164

Wogalter, M.S. (2018) Communication-human information processing (C-HIP) model. In M.S. Wogalter (Eds.), *Forensic Human Factors and Ergonomics* (pp. 33–49). CRC Press. https://doi.org/10.1201/9780429462269-3

Wood, T., & Porter, E. (2019). The elusive backfire effect: Mass attitudes' steadfast factual adherence. *Political Behavior*, *41*(1), 135-163. https://doi.org/10.1007/s11109-018-9443-y

Wu, M., Miller, R. C., & Garfinkel, S. L. (2006, April). Do security toolbars actually prevent phishing attacks? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 601-610 .https://doi.org/10.1145/1124772.1124863

Zannettou, S. (2021, January). "I Won the Election!": An empirical analysis of soft moderation interventions on Twitter. *Proceedings of the Fifteenth International AAAI Conference on Web and Social Media*, 865-876.