Adapting Security Warnings to Counter Online Disinformation

Ben Kaiser
Princeton University

Jerry Wei
Princeton University

Elena Lucherini
Princeton University

Kevin Lee
Princeton University

J. Nathan Matias
Cornell University

Jonathan Mayer
Princeton University

arXiv:2008.10772v1 [cs.HC] 25 Aug 2020

Abstract

Online platforms are using warning messages to counter disinformation, but current approaches are not evidence-based and appear ineffective. We designed and empirically evaluated new disinformation warnings by drawing from the research that led to effective security warnings. In a laboratory study, we found that contextual warnings are easily ignored, but interstitial warnings are highly effective at inducing subjects to visit alternative websites. We then investigated how comprehension and risk perception moderate warning effects by comparing eight interstitial warning designs. This second study validated that interstitial warnings have a strong effect and found that while warning design impacts comprehension and risk perception, neither attribute resulted in a significant behavioral difference. Our work provides the first empirical evidence that disinformation warnings can have a strong effect on users' information-seeking behaviors, shows a path forward for effective warnings, and contributes scalable, repeatable methods for establishing evidence on the effects of disinformation warnings.

1 Introduction

Online platforms are being exploited by disinformation campaigns, which intentionally spread misleading information for financial, political, and social purposes [1, 2]. One countermeasure that platforms have adopted is attaching warning messages to identified disinformation to help inform users' decisions and beliefs. Facebook began this practice as early as December 2016 [3], and Google [4], Bing [5], and Twitter [6] followed with similar warnings. There has been substantial public debate about disinformation warnings—especially after Twitter attached a warning to a May 2020 tweet by U.S. President Donald Trump [7]—but there is little evidence available about how these warnings actually affect users. Without this evidence, it is difficult to say whether disinformation warnings are a promising endeavor or a futile one.

One reason to believe that disinformation warnings can be made effective is the evolution of security warnings. At first, warnings for threats such as malware and phishing websites broadly failed to protect users [8, 9], but after a decade of iterative, multi-method research, these warnings became highly effective [10–21]. Modern security warnings reliably inform users' security decisions and help users distinguish harmful and inauthentic content online [10, 17]. Designing warnings that help users distinguish, contextualize, or avoid disinformation poses analogous challenges to those solved by security warnings researchers. In this work, we consider how to adapt findings and methods from security warnings research to the domain of disinformation.

One key finding from security research that we adapt to disinformation is that *contextual* warnings, which do not interrupt the user or compel them to act, are far less effective at changing users' behaviors than *interstitial* warnings, which interrupt the user and require interaction [8, 9, 17]. Contextual disinformation warnings have been widely studied and deployed, but no prior work has evaluated interstitial disinformation warnings.

Another relevant contribution from security researchers is a set of rigorous qualitative and quantitative methods for evaluating warnings, including structured models, realistic guided tasks, user interviews, and field studies (e.g., [11, 13, 15–18]). Our work adapted these methods to empirically examine contextual and interstitial disinformation warnings.

Across two studies, we used qualitative approaches (think-aloud exercises, interviews, and inductive coding) to understand user perceptions of disinformation warnings, as well as quantitative measures of the warnings' effects on user behavior. We considered the following research questions:

1. After encountering contextual and interstitial disinformation warnings, how often do users change their behavior by opting for alternative sources of information?
2. Why do some users choose not to change their behaviors after encountering contextual and interstitial disinformation warnings?
3. Can interstitial warnings that are highly informative effectively change users' behaviors?
4. Can interstitial warnings that are highly threatening effectively change users' behaviors?

We first conducted a laboratory experiment ($n = 40$) in which participants searched for specific facts on Google and encountered an interstitial or contextual disinformation warning. The interstitial warning was substantially more effective at changing user behavior than the contextual warning, in large part because users did not notice or comprehend the more subtle contextual warning. In post-task interviews, we found two reasons for the interstitial warning's strong behavioral effect: the *informativeness* of the warning's messaging, and the warning's threatening design, which participants found to convey a

risk of harm. In a second study conducted on a population of crowdworkers ($n = 236$), we confirmed the strong behavioral effect of interstitial disinformation warnings. We also found that the effect was not moderated by the informativeness or the risk of harm conveyed by the warning and hypothesize that concerningly, the friction introduced by the warning may be largely responsible for the change in user behavior.

Our results constitute the first evidence that interstitial disinformation warnings can effectively cause users to opt for alternative sources of information, and that such warnings can reliably inform users about the risks posed by disinformation. We also demonstrate scalable, repeatable methods for measuring warning effectiveness and testing theories about warnings' pathways of effect. Motivated by the successful trajectory of security warnings research, we contend that continued research iterating on the warnings and methods we contribute will make platforms' ongoing efforts to design and deploy disinformation warnings more productive.

2 Background and Related Work

Research on disinformation is dispersed across numerous scientific fields. In computer science, there is mature work on automatic detection [22–38] and real-world measurement [39–58], but work on responses and countermeasures is comparatively thin (see Section 2.3).

In this section, we discuss the role of websites in disinformation campaigns, existing research on the effects of disinformation, and prior work on disinformation countermeasures (including warnings).

2.1 Disinformation Websites

Disinformation campaigns are typically multimodal, exploiting many different social and media channels at once [59]. These campaigns use websites as an important tool: to host content for distribution across platforms, facilitate user tracking, and generate ad revenue [42, 45, 60–63]. Disinformation websites are frequently designed to conceal their provenance and deceive users into believing that they are legitimate news or opinion outlets.[1] Our work examined whether warnings can counter this deception and help users distinguish, contextualize, or avoid disinformation websites.

2.2 Effects of Disinformation

Disinformation campaigns hijack the heuristics that users rely on to make accurate judgments about the truthfulness of information [66]. For example, disinformation campaigns often mimic credibility indicators from real news sources [67] and use bots to promote disinformation and create the appearance of popularity or consensus [46].

Misperceptions that individuals hold after consuming disinformation are difficult to dispel [66]. Collectively, a misinformed populace may make social and political decisions that are not in the society's best interests [68] (e.g., failing to mitigate climate change [69]). Influencing policy in this way—by shaping public perception and creating division—is a goal of many disinformation campaigns, especially those by state-level actors [70].

There are several mechanisms by which encountering a warning before exposure to disinformation can protect users from harmful effects. Warnings can induce skepticism in users so they are less likely to take disinformation at face value [71]. A warning can also make a user more susceptible to later corrections [66, 72]. Finally, a warning may cause a user to choose not to read the disinformation at all.

2.3 Responding to Disinformation

There are three main types of responses that platforms and researchers have considered [73]. The first is deranking disinformation by changing recommendation algorithms [74]. Academics have studied this approach in simulated models of social networks [75–80], although not in realistic settings or with real users. Second, platforms have repeatedly removed disinformation content and banned accounts found to be promoting disinformation [81–83]. Neither platforms nor researchers have established evidence on the effects of these takedowns. Finally, platforms have added warnings, labels, and other forms of context to posts [4–6, 84, 85].

In contrast with deranking or removing content, which *restrict* users, warnings can *inform* users' behaviors [86]. Thus deranking and removal raise greater concerns about censorship and speech restrictions [87, 88], while warnings offer a lighter-touch approach—as long as the warnings function in a way that informs rather than coerces or misguides users.

"Disputed" Warnings The most well-studied disinformation warnings are contextual labels indicating a story has been "disputed" or "rated false" by fact checkers [89–94]. These labels constituted Facebook's first major effort to counter disinformation [84], and Google [4], Bing [5], and Twitter [6] have taken similar approaches. Facebook eventually discontinued use of disputed warnings after determining based on internal studies that the warnings were of limited utility [84]. More recently Facebook, Instagram, and Twitter all deployed new warning formats, including interstitials [95, 96].

Some studies of disputed warnings reported no significant effects on participants' perceptions of disinformation [90, 91], while others found effects under certain conditions [89, 93, 94]. Pennycook et al. found that warnings caused participants to rate disinformation as less accurate after repeated exposure to the warnings, but not with only a single exposure [89]. Another study by Pennycook identified a counterproductive *implied truth effect*: the presence of the warning on some headlines caused other headlines to be perceived as more accurate [93]. Seo et al. found that the warnings had a negative effect on accuracy judgments, but the effect did not persist when participants encountered the stories again later [94].

[1] Because this deception involves mimicking the conventions and visual designs of news websites, these outlets are sometimes called fake news (e.g., [64]) or junk news (e.g., [60, 65]).

Related Links Bode and Vraga examined the effects of providing related links to alternative, credible sources of information alongside disinformation, finding that when the related links debunked misinformation, participants' perceived accuracy of the misinformation decreased [97]. Facebook, Google Search, and Bing all currently use this approach.

Highlighting Falsehoods Garrett et al. tested a two-part warning, where participants were first informed that a fact-checker had identified factual errors in a story, then those errors were highlighted in the body of the story [98]. Among users already predisposed to reject the misinformation, this treatment significantly increased accuracy of beliefs, but it had no effect among users inclined to believe the misinformation.

Methods of Prior Work In all of these studies, laboratory participants looked at screenshots of Facebook-style posts, then answered survey questions about how truthful they thought the posts were and whether they would consider sharing the posts on social media. This method can establish evidence about the behavior of laboratory participants, which can form the basis for theories about how users will behave in real settings. As we discuss below, however, security researchers found that in order to measure realistic responses to warnings, it is important to design experimental tasks that cause users to perceive realistic risk [11, 99, 100]. In laboratory studies of security warnings, participants complete role-playing exercises involving sequences of realistic tasks using the real systems under measurement [9, 11, 99]; the scenarios used in disinformation warnings studies are comparatively artificial. The role-playing tasks also permit researchers to empirically observe how participants behave, rather than asking participants to self-assess their likely behavior.

2.4 Security Warnings

Effective warnings are essential to security because human error is a major cause of breaches [101]. Early studies of security warnings found that the warning formats in practice generally failed to protect users from online risks [8, 99]. In contrast, modern warnings are extremely effective: a recent study of over 25 million impressions of browser warnings showed that they protected users from malware and phishing websites around 75-90% of the time [10]. This advancement is due to numerous, rigorous studies over the course of a decade that tested varied warning designs using diverse experimental methods and analytic lenses.

The primary methods of early security warnings studies were laboratory experiments involving supervised tasks, user interviews, and surveys [8, 9, 11, 18, 20, 99, 100]. These studies examined users' beliefs and decision-making processes, in part by using structured models from warnings science literature to identify reasons that warnings failed to change users' behaviors. Security researchers typically used the Communication–Human Information Processing (C-HIP) model, which describes five information processing stages that users undergo when receiving a warning. Users must *notice* the warning, *comprehend* the warning's meaning, *believe* the warning, be *motivated* to heed the warning, and finally, *behave* in the manner intended by the warning issuer [102]. By determining the stage of the C-HIP model at which information processing was failing, researchers learned how to modify warning designs to increase the strength of the desired effect [9, 11].

Limitations It can be difficult to cause users to perceive realistic risk in a laboratory, requiring the use of deception and clever experimental design [99, 100]. It is also important that laboratory studies address the challenge that participants may be more likely to disregard warnings that hinder task completion because they want to complete the task and please the researchers [11, 100]. More recently, researchers have overcome these limitations by employing field studies, which measure users' reactions to warnings at scale in realistic environments [10, 13–15, 17, 19].

This body of work has definitively established that interrupting a user's workflow with a warning is more effective at preventing risky behaviors than passively showing the warning. Wu et al. compared popup warnings to toolbar icons and found that the popups caused users to behave with significantly more caution [8]; Egelman et al. observed that 79% of participants chose not to click through interstitial warnings compared to 13% for contextual warnings [9]. As a result, interstitials and other forms of active security warnings have become standard in all major browsers [17].

Several studies compared multiple warning designs and found that clear messages and use of visual cues can improve comprehension and adherence [11–14]. Personalizing messages based on users' psychological factors has not shown a significant effect on adherence, however [16].

3 Laboratory Study

We began with a laboratory study designed to examine how participants would process and react to contextual and interstitial disinformation warnings when searching for information. The search engine context is conducive to studying behavioral effects because participants have a concrete goal (finding a piece of information) and multiple pathways to achieve that goal (different search results and information sources).

We posed three research questions:

RQ1: In encounters with contextual and interstitial disinformation warnings, do users notice the warnings? Prior studies of contextual warnings note that one reason effect sizes are low or insignificant may be that participants fail to notice the warnings. Because users will never comply with warnings that they do not notice, effective warnings must attract users' attention through conspicuous design or prominent placement.

RQ2: When users notice contextual and interstitial warnings, do they understand that the warnings have to do with disinformation? If a user misunderstands a warning, they may drop it from further cognitive consideration or

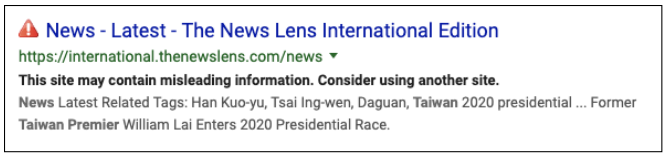


Figure 1: The contextual warning was displayed on search results pages.

respond in unintended ways that could increase risk.

RQ3: When users encounter and comprehend contextual and interstitial disinformation warnings, do they change their behaviors in the intended way by opting for alternative sources of information? This is an important outcome of warning exposure, which we aim to measure as described in Section 3.4.

3.1 Warning Designs

We adapted contextual and interstitial disinformation warnings from modern security warnings used by Google. Google's warnings are well-studied [10, 14, 15, 17, 19, 21, 103] and widely deployed, making them a useful template to design warnings that participants will believe are real.

We developed our contextual warning (Figure 1) based on a warning for compromised websites that Google displays in search results. We changed the color of the text from hyperlink blue to bold black to indicate that the text could not be clicked and also added a red and white exclamation icon next to the search result to make the warning more noticeable.

We adapted our interstitial warning (Figure 2) from Google Chrome's warning page for malware. We modified the text to reference disinformation and changed the "Details" button to "Learn more." Clicking "Learn more" revealed a message explaining that an automated system had flagged the site as disinformation and a "Continue" button that allowed the user to bypass the warning and continue to their selected page.

3.2 Study Design

In a laboratory setting, each participant completed a think-aloud role-playing task followed by an interview. By observing the participant during the role-playing task, we could tell if they noticed the warnings (RQ1) and altered their behavior in response (RQ3). Using the follow-up interviews, we could confirm whether the participant noticed the warnings, ask whether they comprehended the warnings (RQ2), and seek additional insights into how the participant processed the warnings.

Role-Playing Task The participant assumed the persona of an academic researcher trying to find answers to four questions using Google search. As our subjects were all students, we believed this persona would be comfortable and aid with task immersion. For each question, we provided multiple sources of information and attached warnings to just one of the sources so that the participant did not have to click through

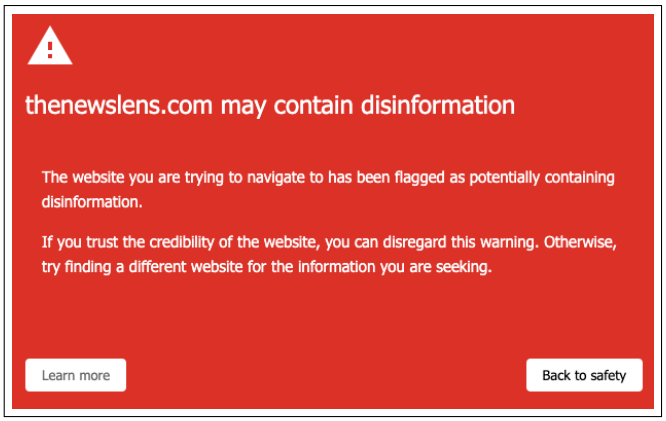


Figure 2: The interstitial warning was displayed after the participant clicked a search result.

the warning to complete the task.[2] Unknown to the participant, two questions were control rounds with no warnings and two were treatment rounds where warnings were activated. We assigned participants in equal numbers to receive either contextual or interstitial warnings in both treatment rounds.

Search Tasks We designed the search tasks (shown in Table 1) to cover simple facts that could be easily found with a single Google search. For the treatment tasks, we selected facts specific to non-American countries and covered by little-known, non-American news sources in order to satisfy three additional design goals. First, the facts were related to current events due to the study's focus on news disinformation. Second, so that participants could choose between multiple sources, each fact was publicly reported by at least two English-language news websites. Third, we aimed to select facts and websites that participants were not likely to be familiar with so as to avoid participants having preconceived biases about the information or the credibility of the sources.

3.3 Study Procedures

We explained the task and role to each participant at the beginning of their session. We asked the participant to behave as if they were using their own computer and to narrate their thoughts and actions as they performed the task. We framed the study as examining research and search engine usage behaviors to avoid priming participants to think about disinformation or warnings.

Participants began the task on the Google homepage. We informed participants that they could use either of two specific websites to find a particular piece of information, and that they should start with the first website and could switch to the second for any reason.[3] Control rounds occurred first and third and treatment rounds occurred second and fourth, with the question order randomized within those sets.

[2] Studies of security warnings have shown that this "task focus effect" can bias participants' behavior [11, 100].

[3] This design directly parallels an evaluation of security warnings conducted by Sunshine et al. [11].

We did not prescribe specific search queries to use, but most participants used a similar format: a set of terms relevant to the fact they needed to find combined with the name of the website on which they wanted to find the information. The participant would enter the query, navigate through results to find the requested information, and verbally provide an answer to the researcher. We would then instruct them to return to the Google homepage to begin the next round.

3.4 Data Collection

We took notes during each session to record how the participant selected search results, browsed the websites to seek information, reacted to warnings, and described their thoughts and actions. We also computed two metrics for each warning: *clickthrough rate* (CTR) and *alternative visit rate* (AVR).

Clickthrough Rate CTR is a commonly used metric in studies of security warnings. It measures the proportion of warning encounters where a participant dismisses the warning instead of going back. For contextual warnings, we recorded a clickthrough if the participant clicked a result that had an attached warning and a non-clickthrough if they chose a different result or changed the search query to use the second suggested website. For interstitial warnings, we recorded a clickthrough if the participant clicked "Learn more" and then bypassed the warning. If they clicked "Back to safety" or used the browser back button to go back to the search results, we recorded a non-clickthrough.

Alternative Visit Rate We also recorded whether participants who clicked through a warning returned to the results page and visited an additional source. We used this data to compute each warning's AVR: the proportion of tasks where after attempting to visit an initial source, a participant visits an alternative source before completing the task. AVR is a new metric we devised for empirically measuring the behavioral effects of a disinformation warning. A high AVR indicates that a warning can influence users to decide to visit secondary sources of information.[4] In some cases this will cause a user not to see the disinformation at all, and in all cases it exposes the user to alternative information.

Interview After the final round, we informed participants of the true purpose of the study, then conducted the interview. We first asked about the participant's general understanding of disinformation: how they defined disinformation, what made them believe a website might contain disinformation, and if they had ever encountered disinformation that they knew of. Next we asked the participant to describe their reactions to the warnings they encountered during the study. We prompted participants to elaborate on these responses until we could determine whether the participant had noticed and comprehended the warnings (RQ1 and RQ2).

Before the next round of questions, we showed the participant printouts of the contextual and interstitial warnings used in the study. We asked whether the participant believed the warnings would be effective in use, if they felt that one format would be more effective than the other, and if they had recommendations for how disinformation warnings in general could be made more effective.

Finally, we asked about the participant's demographics, academic background, and level of news consumption. The full set of interview questions is included in Appendix A.

Coding We combined interview transcripts with our notes and observations to form a single report for each session. We then performed open coding on the reports using Dedoose. In the initial phase of coding, we produced 253 unique codes applied to 1343 text excerpts. We condensed similar codes into themes, then grouped themes into categories if they related to the same topic but provided different perspectives. Finally, we analyzed trends in themes across the full population and within separate treatment groups. Quantifying these trends supported our theory development.

3.5 Participant Recruiting

We recruited participants through the Princeton University Survey Research Center as well as distribution lists of various university organizations. We received 76 responses and selected 40 participants. Our participant group consisted of 16 women and 24 men aged 18-28 years old, studying across 17 disciplines.

Clearly this sample is biased in several respects, including age, education level, and social group. We test later in this work whether there is a difference between the results from this sample and a significantly more diverse sample recruited online (see RQ1 in Section 3). In the context of security warnings studies, student populations have been shown to provide similar results to more representative samples [100].

The recruiting and consent materials provided to participants indicated that the study would take 30-45 minutes and focus on the user experience of search engines, with no mention of disinformation or browser warnings. Participants signed consent forms before beginning the study and were paid $15. The study was approved by the Princeton IRB.

3.6 Results

We present quantitative results for the warnings' behavioral effects (Table 1). We also discuss how notice and comprehension related to participant behavior and present qualitative results on users' opinions and perceptions of the warnings.

3.6.1 Behavioral Effects

Contextual The CTR for the contextual warning was very high: $33/40$. There were a total of $11/40$ alternative visits: 7 non-clickthroughs and 4 occasions where a participant who

[4] AVR does not capture users' perceptions of warnings or the accuracy of their beliefs, which is why we pair this approach with qualitative methods. It is an important open question whether encounters with high-AVR warnings are associated with more accurate beliefs, easier correction of misperceptions, or better ability to distinguish disinformation from real news.

Table 1: We measured clickthrough rates (CTR) and alternative visit rates (AVR) in treatment rounds of the laboratory study.

Round	Participant Instructions	Contextual Warning		Interstitial Warning	
		CTR	AVR	CTR	AVR
Control 1	Find the total area of Italy in square kilometers on Wikipedia or WorldAtlas.	–	–	–	–
Control 2	Report the price of a pair of men's New Balance 574 on JoesNewBalanceOutlet or 6pm.com.	–	–	–	–
Treatment 1	Find the political party of Taiwan's Premier on TheNewsLens or FocusTaiwan.	15/20	7/20	7/20	13/20
Treatment 2	Find the name of the girl reported missing in Barbados on Feb 11, 2019 on BarbadosToday or LoopNewsBarbados.	18/20	4/20	11/20	10/20

clicked through a warning went back to search again using the secondary source.

Interstitial The CTR for the interstitial warning was much lower: 18/40. We observed 1 alternative visit after a clickthrough and 22 alternative visits after non-clickthroughs, for a total AVR of 23/40.

3.6.2 Notice and Comprehension

Contextual In the contextual treatment group, 13/20 participants stated during interviews that they were not aware they had been shown disinformation warnings. All of these participants clicked through both warnings. 4 reported that they did not notice the warnings at all; the other 9 noticed the icons but did not notice the text. In other words, over half (9/16) of participants who noticed the contextual warnings did not comprehend their meaning.

Interstitial All 20 participants noticed the interstitial warnings, and 12 understood that the warnings were about about disinformation. 7 others instead believed the warnings communicated a risk of "harm," a "virus," or another "security threat" and quickly clicked to go back without reading any of the text. The other participant who failed to understand the warnings' purpose clicked through in both treatment rounds; when asked why he clicked through he explained that he was focused on getting the information for the study and "probably would have reacted differently" outside of the study.

3.6.3 Opinions on Disinformation Warnings

As part of the interview, we displayed printouts of both warning designs and asked for the participant's opinions about the warnings' relative merits and general effectiveness.

Contextual When asked which warning design they believed would be more effective in general, a small minority (6/40) chose the contextual warning. 5 of these participants were in the interstitial treatment group.

5 participants noted that the contextual warning did not require users to "commit" by clicking a link before seeing the warning. 1 participant explained, "*you're immediately presented with alternatives, whereas for the interstitial I'm already there and committed a click, so I want to go forward.*" Another preferred the contextual warning because it was easier to bypass: "*[I] just wanted to find a link to click on very quickly, it doesn't take as much effort to avoid compared to the interstitial.*"

5 other participants emphasized the "always-on" nature of the contextual warning. 1 participant liked how they could "*always see the warning when browsing Google search results without having to click around.*" Another felt that the contextual warning was paternalistic because it did not allow the opportunity to make an unadulterated choice, instead "*direct[ing] you to which [results] you should visit*."

15 participants said that the contextual warning was not noticeable. 1 specified that "*the exclamation point is very subtle... you're not going to notice it*."

Interstitial 34 participants—an overwhelming majority—felt that the interstitial warning would be more effective in general. When asked why they thought the interstitial was more effective than the contextual warning, 32 mentioned that it was more noticeable. 17 mentioned the color red, with 1 participant noting that "*everybody knows red means stop*."

19 participants remarked on how the the warning cannot be bypassed without input from the user. 1 participant summarized this by saying that "*it stops the flow of the user and forces them to take an action*." Other findings suggest that design cues contributed to the warning's effectiveness; participants mentioned that the red color and text "*implied that the user is in danger*" and that the text was "*more instructive than the text on the contextual warning*."

When asked about drawbacks to the interstitial warning, 5 participants focused on the inconvenience and the potential for warning fatigue. The sentiment was best described by one participant who said that they would "*probably turn it off in the settings*" if the warning showed up frequently.

Improving Warning Designs Many participants (17) suggested that more informative disinformation warnings would be more effective. Suggestions included adding "*more about why this particular site was flagged,*" definitions of terms used, and more explicit descriptions of potential harms. Conversely, 7 others urged "*short and concise*" messages that "*[get] the point across quickly.*"

5 participants suggested using different warnings depending on severity and whether the user had visited the website before. Another 5 recommended continuing to display warnings even after a user had clicked through on that website before, arguing that warnings would be more effective if users were "*consistently reminded that [the page] may not be completely safe or factual.*"

Participants also suggested alternate warning forms: pop-ups, banner messages, or line-specific highlighting showing instances of false information.

Trust The source of the warning was important to many participants. 8 indicated that they were more likely to be deterred by a browser warning if they knew that it was triggered by Google, since they trusted the company's judgment. 1 participant explained that they clicked through the interstitial warning because they understood that Princeton University had flagged the website, and they felt that the university was not a credible source of judgment about online disinformation. The second theme underlying trust judgments was previous experiences with browser warnings. 7 participants expressed that they distrusted the warnings due to previous encounters with false positive warnings or overly restrictive site blockers on institutional networks (e.g., in high school).

Risk 7 participants expressed the opinion that disinformation is not a serious threat or that it is not as harmful as malware. One participant explained that they typically comply with browser warnings but reacted differently to the disinformation warning: "*I don't like the idea of someone telling me where or what I am allowed to access. You can give me suggestions. It was because I realized it was a disinformation warning and not a malware warning that I went back to try and get to the website.*" Another participant characterized this sentiment sharply, saying "*Disinformation warnings should not make it harder to access the site.*"

3.7 Discussion

Contextual vs. Interstitial Warnings The interstitial warning was distinctly more noticeable and comprehensible than the contextual warning, and also far more effective at inducing behavioral changes. Similar findings in security warnings research prompted the field to shift away from contextual formats and towards interstitial formats. Those working to counter disinformation are only just beginning to make this shift; contextual warnings are currently the dominant format in both research and deployment. While contextual warnings may still have a role to play in countering disinformation, interstitial warnings and other forms of active, interruptive warnings clearly merit further consideration.

Impact of Visual Design Choices The iconography, colors, and text styles used in our warnings impacted how well the warnings attracted participants' attention and affected participants' behavior. The icon we added to the contextual warning made the warning more noticeable but did not necessarily aid with comprehension, as many participants who noticed the icon still failed to notice the text. The red background of the interstitial warning contributed to its effectiveness, but may also have reduced comprehension as participants seemed to react quickly upon seeing the red color without taking the time to read or understand the message of the warning. Again drawing from security warnings research, future work should use comparative effectiveness studies to isolate the effects of individual visual design choices.

3.7.1 Mechanisms of Effect

So few participants complied with the contextual warning that it is difficult to draw conclusions about what caused the behavioral effect. For the interstitial warning, however, we found evidence for three different mechanisms by which the warning induced behavioral changes.

Informativeness Warnings science literature focuses on how a warning educates the user and allows them to make an informed decision about how to proceed [86, 104]. In our study, just under half of the alternative visits (11/25) came from participants who comprehended the interstitial warnings. Although other factors may also have influenced these users' behaviors, we can say with certainty that they made informed decisions about how to proceed with the task.

Fear of Harm Another reason that the interstitial warnings exhibited a strong effect on participants' behavior is that the warnings' threatening appearance caused users to believe they were at risk of harm if they continued. In some cases, participants who chose to visit alternative sites without comprehending the warning perceived a risk of harm but did not identify a specific harm that they were concerned about. Others misinterpreted the warning and believed that it described a risk of receiving a computer virus or other security threat. If a warning is going to convey a risk of harm, it should be specific and narrowly scoped; otherwise users may perceive the warnings as irrelevant or false positives, which would reduce the behavioral effect [10, 15, 18, 20, 105]. As long as the specific harm is made clear, however, using design cues to additionally convey a general risk of harm may improve the effectiveness of the warning.

Friction Finally, part of the interstitial warning's strong effect was due to the friction it introduced into participants' workflows. Some participants preferred to go back to the search results and choose another source rather than read the warning, decide whether to believe and comply with it, and click through it. As with the "fear of harm" mechanism, friction must be used carefully to avoid inducing warning fatigue or habituating users to ignore warnings. Further, friction can be thought of as a degradation of the quality of the user's experience. Even if a warning is highly effective at changing behaviors, platforms may be hesitant to deploy that warning if its effect relies on making the platform harder to use.

3.7.2 Limitations

Security warnings studies have observed challenges in studying behavioral responses to risks in laboratory settings, particularly with respect to ecological validity [100]. We encountered similar challenges in this study.

Our sample was small in size and biased in several ways; thus our findings should be understood as illustrative but not representative. It is important to identify if our findings can be replicated by larger, more diverse populations.

Because participants used our computer and we were watching over their shoulder, some participants reported that they behaved differently in the study than they might have in real life. Others may not have reported such because they were reluctant to admit that they behaved with bias or because the effect was unconscious. An experimental design that allows participants to use their own computers in their own environments (i.e., not in a laboratory) could offer more realistic observations of how participants assess risk.

Finally, although participants appeared to be driven to complete the research tasks, they were not personally invested in completing the tasks or finding correct answers to the queries. The role-playing aspect of the study may not have been strongly immersive, and there were no extrinsic rewards or penalties that incentivized correct answers. Further, because our search tasks did not pertain to participants' social or political contexts, participants had little reason to engage in motivated reasoning. Motivated reasoning can strongly influence users' perceptions of information relevant to their social or political beliefs [106], so in those contexts, the warning effects that we demonstrate may be weaker.

4 Crowdworker Study

In our second study, we aimed to verify whether interstitial disinformation warnings are truly effective at changing users' behaviors and investigate the mechanisms behind that effect. Understanding the mechanisms of effect allows us to reason about the utility, limitations, and risks of deploying interstitial disinformation warnings.

Our research questions were:

RQ1: Do interstitial disinformation warnings cause users to choose alternative sources of information? We sought to establish whether the bias of our sample population or the limitations of the laboratory task significantly affected the observed outcomes. We recruited a larger, more diverse participant pool from Amazon Mechanical Turk (Section 4.6) and adjusted the experimental design to account for limitations we observed in the laboratory (Section 4.2).

RQ2: Do interstitial warnings that effectively inform users about the risks of disinformation cause users to choose alternative sources of information? We tested whether participants understood the warnings, then compared the behavioral effects of informative and uninformative warnings to isolate the impact of informativeness on behavior.

RQ3: Do interstitial warnings that communicate a risk of personal harm cause users to choose alternative sources of information? We tested whether warnings that caused users to feel that they were at risk of harm produced substantially different behavioral effects than warnings that did not cause this perception.

RQ4: Does user partisan alignment (with respect to U.S. politics) moderate users' behavioral responses or their perceptions of interstitial warnings? Research in political science shows that political orientation affects judgments of information credibility [107] as well as the effectiveness of misinformation warnings [93]. We included this research question in order to detect if partisan alignment created a bimodal distribution in responses to warnings.

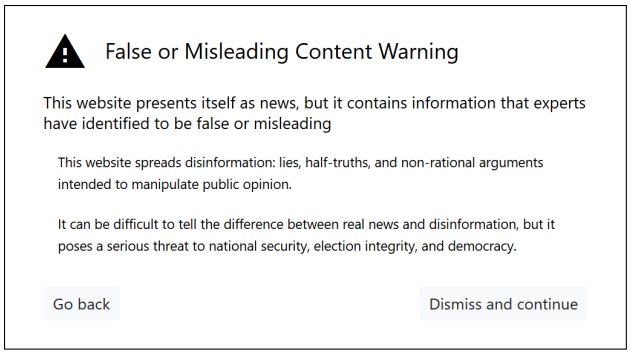
The basic structure of the task and the key behavioral outcomes remained the same as in the laboratory study. We informed participants that they were participating in a study of search engine usage and research behaviors, then guided them through a series of four research tasks using a search engine, alternating between control and treatment rounds. In each treatment round, the participant encountered one of the eight candidate interstitial disinformation warnings upon clicking certain search results. We measured whether the participant clicked through the warning and whether they visited an alternative site. We report the overall CTR and AVR across all observations in order to answer RQ1.

We used surveys after each warning encounter to measure how informative the warning was and how strongly the participant perceived the warning to convey a risk of harm. RQ2 and RQ3 concern the relationship between AVR and these survey responses. A standard analysis approach would have been to randomly assign participants to warnings, then compute statistical tests across the conditions. Unless the differences in effect between warnings were dramatic, however, this approach would have required a massive number of observations on each warning to establish statistical significance.

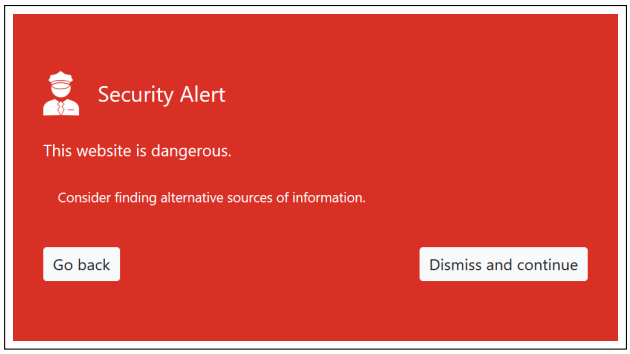
We instead employed a multi-armed bandit algorithm, which allows efficient exploration of a larger design space than is traditionally possible. As we received successive observations, the bandit increased the odds that participants encounter the warnings proving to be most and least informative and most and least effective at conveying fear of harm. After all observations were completed, significantly more participants had encountered these top- and bottom-performing warnings, providing us with the statistical power needed to test our hypotheses. Section 4.5 discusses the design and implementation of the multi-armed bandit.

4.1 Warning Designs

We created eight candidate interstitial disinformation warnings: four designed for informativeness and four designed to evoke fear of harm (see Table 4 in Appendix B.1). We first developed a common layout for the warnings, consisting of an icon, a title, a primary message, an optional detailed message, and two buttons. This layout differed from the laboratory



(a) Informative warning design *i2*



(b) Harm warning design *h3*

Figure 3: Examples of warning designs in the crowdworker study. Table 4 in Appendix B.1 describes all 8 warning designs.

interstitial warning in two ways.

First, in the laboratory warning design, the detailed message (and the "Continue" button) were hidden at first and would only be revealed after the participant clicked "Learn more." In the crowdworker study, we wanted to ensure that the full warning message was always displayed, because part of what we were measuring was the effect of different messages. Therefore we eliminated the "Learn more" button for the new warning designs and instead plainly displayed the detailed message and the "Continue" button on all warnings.

The second change addressed the "Back to safety" button. This button text implied that the user was in danger, which was inappropriate for the informative warnings. We changed the button to read "Go back" and applied this change to both informative and harm-based warnings in order to maintain a common interaction design across all warnings.

For both groups of warnings, we generated a few options for icons, titles, and messages, then created candidate designs by choosing permutations of these elements and inserting them into the layout template.

Informative Warnings We designed the informative warnings to be visually nonthreatening and clearly explanatory in their messages (see Figure 3a). The warnings included one of two icons—either a generic exclamation point icon or a policeman silhouette—and displayed black text against a white background. The warning messages explained the nature and consequences of disinformation in varying detail: some explicitly defined the term "disinformation," some asserted that "experts" had labeled the site as containing "false or misleading" information, and others provided clear guidance on how to behave ("finding alternative sources of information").

Harm Warnings The harm warnings (see Figure 3b) contained little explanatory text and used forceful language, colors, and icons to suggest a serious threat. The warnings used either a skull-and-crossbones icon or the policeman icon, and the content was colored white against a red background. The most extreme warning design simply said: "WARNING: This website is dangerous." The other three were titled "Security Alert" and indicated in the messages that the threat had to do with information quality.

4.2 Task Design

We aimed to keep procedures for the crowdworker study as similar to the laboratory study as possible, but the different setting and research questions necessitated three changes.

First, because crowdworkers participate remotely and use their own computers, we could not instrument their browsers with warnings or control search queries or results. Therefore, instead of using live Google searches, we developed a web application to guide participants through the experiment and simulate a realistic-looking search engine that we controlled (Figure 4). We populated results for each query from real Google results for that query, directly copying search results snippets. To simulate results pages and ensure that all participants saw the exact same content, we used full-page screenshots of results pages, which participants could browse similarly to browsing the real website (Figure 5). Unlike in the laboratory study, we specified the queries to use and did not direct participants to specific sources.

Second, crowdworkers participated in our study to earn a wage, which is a very different motivation than our in-person laboratory subjects for whom study participation was elective. Crowdworkers may be more focused on completing the task quickly (in order to earn the task fee) than on engaging meaningfully with the study and behaving as "good" research participants. One way we addressed the risk was using qualifications to ensure that only workers with track records of submitting quality work participated in a study (see Section 4.6). We also offered a bonus of $1 (43% of the base fee) to participants who submitted correct answers for all four research tasks. The bonus created an incentive for crowdworkers to engage with the tasks, read instructions carefully, seek accurate information, and take disinformation warnings seriously.

Finally, we used a series of surveys in lieu of directly observing participants and conducting interviews. Each participant completed a pre-task survey about their partisan alignment, surveys in each round about their behavior and percep-

Table 2: The crowdworker study task involved two control rounds and two treatment rounds. We report clickthrough rates (CTR) and alternative visit rates (AVR) for each warning in Table 3.

Round	Participant Instructions
Control 1	A train from the Massachusetts Bay Transportation Authority derailed in Boston on June 8. Please find the color of the line that the train was on using the query "MBTA derailment."
Control 2	In Alabama, the Muscogee tribe filed a lawsuit over a casino that was built on sacred ground. Please find the name of the casino using the query "muscogee nation casino lawsuit."
Treatment 1	Please find the names of the two Clatskanie School Board members who were recalled in May 2018 using the query "clatskanie school board recall."
Treatment 2	Mike Sanders, the Executive of Jackson County, Kansas, was sentenced to prison in 2018 for embezzlement. Please find the name of the political party to which he belonged using the query "mike sanders jackson county."

tion of the warning, and a post-task demographic survey.[5]

Search Tasks As in the laboratory study, we selected facts for participants to retrieve that were reported by multiple sources and obscure enough that participants would not likely be familiar with the topic or sources. We also ensured that all search results came from news outlets and that no two results came from the same outlet, giving participants a greater variety of choices when choosing sources of information. All four tasks pertained to events in the U.S. to make the topics more relevant to our U.S.-based participant population. We also designed the treatment tasks to specifically cover political scandals so that participants would find it plausible that news outlets might publish disinformation about about these topics. Table 2 presents the tasks and queries.

Procedures After accepting our job on the crowdworking site, participants visited the website hosting our experimental tool. The first page displayed instructions and a visual aid demonstrating how to use the interface, then directed participants to begin the first search round.

Each round consisted of a research task where the participant used our simulated search engine to find a particular fact from a set of search results. The participant began on a generic search query page (Figure 4) which displayed instructions specifying the fact to search for and the query to use. When the participant submitted the query, they were navigated to a search results page populated with eight results. Clicking on a result led to a story page containing the news article the search result snippet described (Figure 5). These story pages were full-size screenshots of real news article web pages, allowing participants to scroll though and read the articles as if they were browsing the real web pages.

Three of the results in every round were *target results*: the search result snippets clearly pertained to the query topic and the story pages contained the answer to the question. Target story pages provided the answer clearly in the body of the article. The other 5 were *nontarget results*, which linked to story pages that did not contain the answer. Some nontarget results could be identified as erroneous by the results snippet (i.e., they contained some of the search terms but in a different context), while others were germane to the query topic but did not contain the specific piece of information being sought. The search results were ordered so that the top result was always a target result, and the other two target results appeared randomly within the top five. Other than these constraints, which were designed to ensure that target results were easy to find, the order of the results was randomized.

On the story page, an instruction box repeated the question and allowed the participant to submit an answer or return back to the search results. If the participant chose to go back to the results page, all results they had selected so far would be grayed out and the links would be disabled.

In treatment rounds, the participant saw an interstitial warning the first time they clicked a target result. Our multi-armed bandit algorithm (Section 4.5) selected the warning design the participant encountered. All warnings included two buttons: "Go back" or "Dismiss and continue." If the user clicked "Go back" or used the browser back button, they returned to the search results. The user would not be shown a warning when they clicked on another result, even if it was a target. If the user clicked "Dismiss and continue," they were taken to the story page, where they had the option to return to the search results or submit an answer.

When the participant submitted an answer, they were taken to a survey asking why they chose particular search results. This was a misdirection to maintain the false premise that the experiment was studying search engine usage. In control rounds, submitting this survey led to the next round. In treatment rounds, the next page was a second survey designed to capture participants' perceptions of the level of informativeness and risk of harm conveyed by the warning (see Section 4.3). This survey also included an attention check question so that we could discard responses from participants who were not reading instructions carefully.

After completing all four rounds, participants were navigated to a final, demographic survey, then compensated.

[5] See Appendix B.3 for survey details.

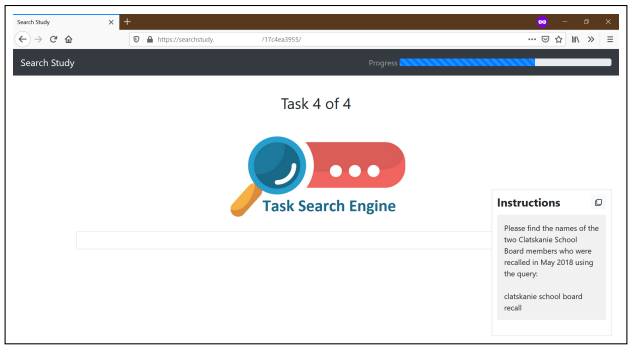


Figure 4: Each round of the crowdworker study began on this search page.

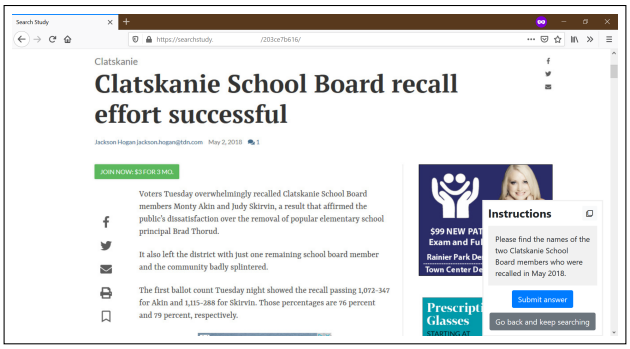


Figure 5: Clicking on a search result led the participant to a story page containing a screenshot of a real news webpage, instructions, and buttons to submit an answer or navigate back.

4.3 Measuring Participant Perceptions of Warnings

After each warning encounter, we used a survey to measure how informative and how harmful the participant found the warning. We developed these survey questions based on our laboratory findings and several small-scale pilot studies.

Informativeness We designed three survey questions to measure whether participants could identify the topic of the warning. In our laboratory study, participants who misunderstood a warning typically believed that the warning had to do with malware or another security threat. Therefore we asked participants in this study to report their belief that the warning described each of three possible topics: malware, information theft, and disinformation. The survey asked how strongly the participant agreed with each of these statements on a Likert scale from 1 ("Strongly disagree") to 5 ("Strongly agree"):

- $info_1$: The warning said that this site is trying to install malware.
- $info_2$: The warning said that this site is trying to steal my information.
- $info_3$: The warning said that this site contains false or misleading information.

We used the survey responses to compute an informativeness score i in the range $[-2,2]$, which captured the participant's certainty that the warning had to do with disinformation. $i_{p,w} = 2$ if participant p "strongly agreed" that warning w was about false or misleading information and "strongly disagreed" about the other topics. For each point deviation from these "correct" responses on the Likert scales, $i_{p,w}$ would be reduced by 1, resulting in a lower score when the participant was uncertain in their answer or had an incorrect understanding of the warning. The scoring formula was:

$$i_{p,w} = max(-2,\ info_3 - info_2 - info_1 - 1)$$

Harm We used two survey questions to measure whether a warning caused a participant to believe that they were at risk of harm. The survey asked how strongly the participant agreed with each of these statements on a Likert scale from 1 ("Strongly disagree") to 5 ("Strongly agree"):

- $harm_1$: After seeing the warning, I believe the website may harm me if I visit it.
- $harm_2$: After seeing the warning, I believe the website may harm other people who visit it.

In our pilots of these survey questions, participants emphasized the distinction between these two forms of harm (personal and societal). Our research questions only concerned perceptions of the risk of personal harm, however, so we computed the harm score h for participant p and warning w by projecting their response to $harm_1$ into the range $[-2,2]$:

$$h_{p,w} = harm_1 - 3$$

4.4 Measuring Participants' Behaviors

We measured the same behavioral outcomes in this study as in the laboratory study (see Section 3.4): clickthrough rate (CTR) and alternative visit rate (AVR).

CTR represents the proportion of warning encounters where the participant clicked "Dismiss and continue."

AVR measures how often participants visited an alternative source before submitting an answer. We recorded an alternative visit when a participant clicked at least one additional result after viewing a warning and before submitting an answer. In this study, we measured AVR in both treatment and control rounds, which allows for estimation of the treatment effect with respect to a base rate (the AVR in control rounds when no warning is shown).

4.5 Assigning Warnings

To answer our research questions about mechanisms of effect (RQ2 and RQ3), we measured for differences in the behavioral effects between the warnings that achieved the highest and lowest scores on the perception measures.

The standard method for comparing effect sizes between treatments is a randomized control test, in which participants are randomly assigned to treatments (usually in equal numbers). The key variable determining how many observations are needed is the estimated difference in effect size between treatments. If this difference is small, the study will require

a large number of observations for each treatment to achieve statistically significant confidence.

When designing our study, we observed that the difference in effect sizes between treatments could be small, meaning that a large sample size could be necessary to evaluate our hypotheses. Our observations were expensive, however, and although we were testing 16 different conditions (8 treatments with 2 outcomes each), we were only interested in comparing the effects of 4 of those conditions (the top and bottom performers for each outcome). Therefore we sought a method to efficiently assign participants to warnings so that the top- and bottom-performing warnings achieved high confidence levels, but the other warnings (which we would not end up using in our hypothesis tests) would receive fewer observations and therefore consume fewer experimental resources.

In this study, we used an adaptive bandit algorithm to assign participants to warnings based on observations of previous participants. With each new observation, bandit algorithms update the probability of each condition in a study according to some *reward function* that aligns with the researchers' scientific goals [108]. Bandits have been widely used in clinical trials, software design optimization, and political opinion research [109–111]. We discuss the full details of our multi-armed bandit implementation in Appendix B.2.

The reward function in our adaptive experiment preferred disinformation warnings that achieved high and low mean scores for harm and informativeness. For the first $n = 80$ participants, the algorithm assigned all warnings equally. For the remaining participants, the reward function prioritized warnings with the highest and lowest scores for harm and informativeness in order to maximize our ability to differentiate between warnings. As these maximum and minimum warnings emerged, the algorithm improved the precision of our estimates by directing further participants to receive those warnings rather than less influential ones.

4.6 Participant Recruiting

We collected data from 250 Amazon Mechanical Turk (MTurk) workers who were based in the U.S. and had completed more than 5,000 jobs with an approval rate above 97%. We discarded data from 12 workers who failed an attention check question, leaving a sample population of 238. The population was roughly two-thirds male and over half of participants were between the ages of 30 and 49. The majority consumed news media at least five days a week and paid fairly high attention to politics and current events. We show the full population demographics in Table 6 in Appendix B.3.

The recruitment materials, consent form, and instructions described the task as pertaining to the use of search engines and made no mention of warnings or disinformation. We estimated the task duration at 15-20 minutes and compensated participants $2.33 with the opportunity to earn a $1 bonus (43%) for retrieving the correct answer for all four queries. If a participant abandoned the task partway through or exceeded a 2-hour time limit, we discarded their data and recruited a replacement participant.

Our study was approved by the Princeton University IRB.

Table 3: We report alternative visit rates (AVR), clickthrough rates (CTR), and mean informativeness ($\bar{i}$) and harm ($\bar{h}$) scores with 95% confidence intervals.

	Liberal					Conservative				
	#	AVR	CTR	$\bar{i}$	$\bar{h}$	#	AVR	CTR	$\bar{i}$	$\bar{h}$
Control	318	20%	–	–	–	158	16%	–	–	–
Treatment	318	87%	16%	–	–	158	85%	17%	–	–
Selected treatments										
h1	120	85%	18%	−1.94 ±0.06	1.18 ±0.18	46	83%	17%	−1.91 ±0.11	–
h3	73	84%	18%	–	–	27	81%	22%	–	1.15 ±0.46
i3	39	87%	13%	1.41 ±0.43	–	10	90%	10%	–	–
i4	17	82%	12%	–	−0.76 ±0.69	25	76%	24%	0.88 ±0.69	−0.2 ±0.62

4.7 Results

We pre-registered our analysis methods [112]. We computed mean ratings and 95% confidence intervals for harm and informativeness scores (see Figures 6 and 7 in Appendix B.2). As top and bottom performers, we selected the warnings with the highest and lowest mean ratings where the 95% confidence intervals did not overlap (see Table 3).

Informativeness Liberal participants identified *i3* as the most informative warning, producing a very high mean score of 1.41 (on the scale $[-2, 2]$). Liberals also gave two other warnings high, consistent ratings for informativeness (*i2* and *i4*). On average, conservatives gave much lower informativeness ratings to all of the warnings that we designed to be informative. Conservatives selected *i4* as the most informative design, with a high mean score of .88. The next most informative warning conservatives selected was a warning we had intended to convey harm: *h4*.

The least informative warning was *h1*, which achieved consistent, extremely low informativeness scores from both liberals and conservatives. *h1* was the most extreme warning in the harm category; the only text it contained was "WARNING: This website is dangerous."

Harm Liberals identified *h1* as the most threatening warning with high confidence, providing an average score of 1.18. Conservatives also gave high harm ratings to *h1*, but just slightly higher ratings to *h3* (1.15).

Both groups selected *i4* as the least harmful warning, although liberals were more likely to give the warning negative ratings for harm (−0.76) while conservatives gave it more neutral ratings (−0.2).

CTR The cumulative CTR across all treatments was 16%, which was noticeably lower than what we observed for interstitial warnings in the laboratory (40%). No individual

warning in this study demonstrated a higher CTR than the interstitial warning we tested in the laboratory.[6]

AVR The cumulative AVR across all treatments was 86%, compared to 19% in control rounds. We evaluated whether this difference was significant using a one-sided z-test. We rejected the null hypothesis that the AVR was equivalent in control and treatment rounds with a z-score of 22.44 and a p-value of $1.48e^{-111}$.

We tested whether there was a significant difference in AVR between top- and bottom-rated warnings for harm within the liberal or conservative groups. For liberals, we used a one-sided z-test and failed to reject the null hypothesis that the AVR was equal for the two warnings ($p = 0.41$). For conservatives, we used Fisher's exact probability test (due to the small sample size of conservative participants) and also failed to reject the null hypothesis ($p = 0.54$).

Next, we tested whether there was a significant difference in AVR between top- and bottom-rated warnings for informativeness within the liberal or conservative groups. Again, we failed to reject both null hypotheses that AVR was equivalent, using a z-test for liberals ($p = 0.27$) and Fisher's test for conservatives ($p = 0.54$).

4.8 Discussion

This study replicated the strong behavioral effect of interstitial disinformation warnings that we observed in the laboratory study. We also demonstrated that it is possible to design interstitial disinformation warnings that effectively inform users that a website may contain disinformation. Conveying that a website may contain disinformation is a necessary step in prompting users to think critically about the website's trustworthiness. Past studies have shown that critical thinking is among the most important predictors of a user's ability to correctly judge the accuracy of information [107, 113].

Our instrument, however, did not measure a significant difference in effect between informative and uninformative warnings. This suggests that the behavioral effects we observed may not be strongly associated with comprehension of the warning's message or use of critical thinking. If informativeness is not the key causal factor in changing behaviors, then warnings are not achieving the intended goal of helping users make informed decisions. Future work will need to consider how to develop warnings that do achieve this goal.

As we found that neither informativeness or fear of harm explained the warnings' behavioral effects, we hypothesize that the friction and inconvenience introduced by the warning may be important causal factors. We established evidence for this theory in our laboratory study (Section 3.7.1), although we did not test it in the crowdworker study.

Limitations There may have been moderator variables we did not measure that affected the relationship between warning perception and behavioral changes. User interviews and detailed qualitative surveys, like those we conducted in our laboratory study, can help surface these variables, but these methods are difficult in a crowdworker setting.

We also note that while our crowdworker sample was more diverse than our laboratory sample, neither sample was representative of the American population. The behavioral effects we observed were fairly consistent across demographic groups, however. Importantly, our study population only included Americans. Cross-cultural research is needed to understand if the effects we observed apply globally.

5 Conclusion

In this section, we provide recommendations for future evaluation and deployment of disinformation warnings.

5.1 Directions for Future Research

Future work can adapt our methods to measure how the effects we demonstrate interact with other factors known to impact adherence to warnings and receptivity to disinformation. These factors include repetition of warnings [10, 15, 18, 20, 105], users' age and level of digital literacy [114, 115], their tendency towards cognitive reflection [113, 116, 117] whether users are repeatedly exposed to the information that they are being warned about [19, 89], and whether that information has a partisan skew that aligns or conflicts with the user's ideology [115, 118, 119].

Another important direction is to measure for unintended, potentially harmful effects of interstitial disinformation warnings. These may include an effect on perceptions of information that does not have a warning attached (either implied truth [93] or general distrust [89]) or a decrease in trust of the institution identified as the source of the warning. There may also be effects caused by repeated exposure to interstitial disinformation warnings over time, such as warning fatigue [20].

Finally, future work should evaluate the effects of other types of interstitial warnings that command users' attention and require interaction. Two active formats that platforms have begun to experiment with are popup windows, which users must dismiss, and overlays that obscure content until the user clicks to show the content [120, 121].

5.2 Informing Platforms' Disinformation Warnings

Interstitial warnings have the potential to be highly effective tools for countering disinformation. Compared to contextual warnings, interstitials are easier for users to notice and comprehend. Platforms that continue to use contextual warnings should be aware that the warnings' effects are minimal, and that making the warnings more prominent and explicit may increase their effectiveness.

More generally, platforms must take evidence-based approaches to developing and deploying disinformation warnings. By releasing data, conducting internal evaluations, and collaborating with independent researchers, platforms can sig-

[6] We did not perform significance tests for CTR, as our research questions pertained only to AVR.

nificantly advance their capacity to counter disinformation with warnings, just as browser vendors did when developing warnings for other online threats [10, 13–15, 17, 19].

References

[1] Samantha Bradshaw and Philip N. Howard. *The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation*. Tech. rep. University of Oxford, Sept. 26, 2018. URL: https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2019/09/CyberTroop-Report19.pdf.

[2] Diego A. Martin, Jacob N. Shapiro, and Michelle Nedashkovskaya. "Recent Trends in Online Foreign Influence Efforts". In: *Journal of Information Warfare (JIW)* 18 (3 2019). URL: https://scholar.princeton.edu/sites/default/files/jns/files/martin_d._shapiro_j._nedashkovskaya_m._recent_trends_in_online_foreign_influence_efforts.pdf.

[3] Adam Mosseri. *Addressing Hoaxes and Fake News*. Facebook Newsroom. Dec. 15, 2016. URL: https://about.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/ (visited on 06/03/2020).

[4] Justin Kosslyn and Cong Yu. *Fact Check now available in Google Search and News around the world*. Google Official Blog. July 2017. URL: https://www.blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/ (visited on 03/05/2020).

[5] Microsoft Bing. *Bing adds Fact Check label in SERP to support the ClaimReview markup*. Bing Blogs. Sept. 14, 2017. URL: https://www.blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/ (visited on 03/05/2020).

[6] Davey Alba and Kate Conger. *Twitter Moves to Target Fake Videos and Photos*. New York Times. Feb. 2020. URL: https://www.nytimes.com/2020/02/04/technology/twitter-fake-videos-photos-disinformation.html (visited on 04/18/2020).

[7] Taylor Hatmaker. *Twitter adds a warning label fact-checking Trump's false voting claims*. TechCrunch. May 26, 2020. URL: https://techcrunch.com/2020/05/26/twitter-trump-labels-fact-checking-tweet/ (visited on 06/03/2020).

[8] Min Wu, Robert C Miller, and Simson L Garfinkel. "Do security toolbars actually prevent phishing attacks?" In: *Proceedings of the 2006 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Apr. 2006. DOI: 10.1145/1124772.1124863.

[9] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. "You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings". In: *Proceedings of the 2008 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Apr. 2008. DOI: 10.1145/1357054.1357219.

[10] Devdatta Akhawe and Adrienne Porter Felt. "Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness". In: *Proceedings of the 22nd USENIX Security Symposium (USENIX Security)*. Aug. 2013. URL: https://www.usenix.org/system/files/conference/usenixsecurity13/sec13-paper_akhawe.pdf.

[11] Joshua Sunshine et al. "Crying Wolf: An Empirical Study of SSL Warning Effectiveness". In: *Proceedings of the 18th USENIX Security Symposium (USENIX Security)*. Aug. 2009. URL: https://www.usenix.org/legacy/event/sec09/tech/full_papers/sunshine.pdf.

[12] Marian Harbach et al. "Sorry, I Don't Get It: An Analysis of Warning Message Texts". In: *Proceedings of the 17th International Conference on Financial Cryptography and Data Security (FC)*. Vol. 7859. Lecture Notes in Computer Science (LNCS). Apr. 2013. DOI: 10.1007/978-3-642-41320-9_7.

[13] Adrienne Porter Felt et al. "Experimenting At Scale With Google Chrome's SSL Warning". In: *Proceedings of the 2014 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Apr. 2014. DOI: 10.1145/2556288.2557292.

[14] Adrienne Porter Felt et al. "Improving SSL Warnings: Comprehension and Adherence". In: *Proceedings of the 2015 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Apr. 2015. DOI: 10.1145/2702123.2702442.

[15] Joel Weinberger and Adrienne Porter Felt. "A Week to Remember: The Impact of Browser Warning Storage Policies". In: *Proceedings of the 12th Symposium On Usable Privacy and Security (SOUPS)*. June 2016. URL: https://www.usenix.org/system/files/conference/soups2016/soups2016-paper-weinberger.pdf.

[16] Malkin, Nathan and Mathur, Arunesh and Harbach, Marian and Egelman, Serge. "Personalized Security Messaging: Nudges for Compliance with Browser Warnings". In: *Proceedings of the EuroUSEC 2017 – The 2nd European Workshop on Usable Security*. Apr. 2017. DOI: 10.14722/eurousec.2017.23008.

[17] Robert W. Reeder et al. "An Experience Sampling Study of User Reactions to Browser Warnings in the Field". In: *Proceedings of the 2018 ACM SIGCHI Conference on Human Factors in Computing Systems (CHI)*. Apr. 2018. DOI: 10.1145/3173574.3174086.

[18] Serge Egelman and Stuart E. Schechter. "The Importance of Being Earnest (In Security Warnings)". In: *Proceedings of the 17th International Conference on Financial Cryptography and Data Security (FC)*. Vol. 7859. Lecture Notes in Computer Science (LNCS). Apr. 2013. DOI: 10.1007/978-3-642-39884-1_5.

[19] Hazim Almuhimedi et al. "Your Reputation Precedes You: History, Reputation, and the Chrome Malware Warning". In: *Proceedings of the 10th Symposium On Usable Privacy and Security (SOUPS)*. June 2014. URL: https://www.usenix.org/system/files/conference/soups2014/soups14-paper-almuhimedi.pdf.

[20] Cristian Bravo-Lillo et al. "Harder to Ignore? Revisiting Pop-Up Fatigue and Approaches to Prevent It ". In: *Proceedings of the 10th Symposium On Usable Privacy and Security (SOUPS)*. June 2014. URL: https://dl.acm.org/doi/abs/10.5555/3235838.3235847.

[21] Mustafa Emre Acer et al. "Where the Wild Warnings Are: Root Causes of Chrome HTTPS Certificate Errors". In: *Proceedings of the 24th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. Oct. 2017. DOI: 10.1145/3133956.3134007.

[22] Andreas Vlachos and Sebastian Riedel. "Identification and Verification of Simple Claims about Statistical Properties". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Sept. 2015. URL: https://www.aclweb.org/anthology/D15-1312.pdf.

[23] James Thorne and Andreas Vlachos. "An Extensible Framework for Verification of Numerical Claims". In: *Proceedings of the Software Demonstrations of the 15th Annual Meeting of the Association for Computational Linguistics (ACL)*. Apr. 2017. URL: https://www.aclweb.org/anthology/E17-3010.pdf.

[24] Naeemul Hassan et al. "Toward Automated Fact-Checking: Detecting Check-worthy Factual Claims by ClaimBuster". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery in Data Mining (KDD)*. July 2017. DOI: 10.1145/3097983.3098131.

[25] James Thorne and Andreas Vlachos. "Automated Fact Checking: Task Formulations, Methods and Future Directions". In: *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*. Aug. 2018. URL: https://www.aclweb.org/anthology/C18-1283.

[26] James Thorne et al. "FEVER: a large-scale dataset for Fact Extraction and VERification". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*. (Long paper). June 2018. DOI: 10.18653/v1/N18-1074.

[27] Hannah Rashkin et al. "Truth of Varying Shades: Analyzing Language in Fake News and Political Fact-Checking". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Sept. 2017. DOI: 10.18653/v1/D17-1317.

[28] Martin Potthast et al. "A Stylometric Inquiry into Hyperpartisan and Fake News". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. (Long paper). July 2016.

[29] Benjamin D. Horne and Siebl Adali. "This Just In: Fake News Packs A Lot In Title, Uses Simpler, Repetitive Content in Text Body, More Similar To Satire Than Real News". In: *Proceedings of the 11th International Conference on Weblogs and Social Media (ICWSM)*. May 2017. URL: https://aaai.org/ocs/index.php/ICWSM/ICWSM17/paper/view/15772.

[30] Austin Hounsel et al. *Supporting Early and Scalable Discovery of Disinformation Websites*. Feb. 28, 2020. arXiv: 2003.07684.

[31] Fan Yang et al. "Automatic detection of rumor on Sina Weibo". In: *Proceedings of the 2012 ACM Workshop on Automated Decision Making for Active Cyber Defense (SafeConfig)*. Aug. 2012. DOI: 10.1145/2350190.2350203.

[32] Aditi Gupta et al. "Faking Sandy: Characterizing and Identifying Fake Images on Twitter during Hurricane Sandy". In: *Proceedings of the 22nd World Wide Web Conference (WWW)*. May 2010. DOI: 10.1145/2487788.2488033.

[33] Leon Derczynski and Kalina Bontcheva. "PHEME: Veracity in Digital Social Networks". In: *Posters, Demos, Late-breaking Results and Workshop Proceedings of the 22nd Conference on User Modeling, Adaptation, and Personalization (UMAP)*. July 7, 2014. URL: http://ceur-ws.org/Vol-1181/pros2014_paper_05.pdf.

[34] Ke Wu, Song Yang, and Kenny Q Zhu. "False rumors detection on Sina Weibo by propagation structures". In: *Proceedings of the 31st International Conference on Data Engineering (ICDE)*. Apr. 2015. DOI: 10.1109/ICDE.2015.7113322.

[35] Zhe Zhao, Paul Resnick, and Qiaozhu Mei. "Enquiring Minds: Early Detection of Rumors in Social Media from Enquiry Posts". In: *Proceedings of the 24th World Wide Web Conference (WWW)*. May 2015. DOI: 10.1145/2736277.2741637.

[36] Jing Ma, Wei Gao, and Kam-Fai Wong. "Detect Rumors in Microblog Posts Using Propagation Structure via Kernel Learning". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. (Long paper). Aug. 2017. DOI: 10.18653/v1/P17-1066.

[37] Ramy Baly et al. "Predicting Factuality of Reporting and Bias of News Media Sources". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Oct. 2018.

[38] Kashyap Popat et al. "Where the Truth Lies: Explaining the Credibility of Emerging Claims on the Web and Social Media". In: *Companion Proceedings of the The Web Conference (WWW) 2017*. Apr. 2017. DOI: 10.1145/3041021.3055133.

[39] Michela Del Vicaro et al. "The spreading of misinformation online". In: *Proceedings of the National Academy of Sciences (PNAS)* 113 (3 Jan. 2016). DOI: 10.1073/pnas.1517441113.

[40] Hunt Allcott and Matthew Gentzkow. "Social Media and Fake News in the 2016 Election". In: *Journal of Economic Perspectives (JEP)* 31.2 (2017). DOI: 10.1257/jep.31.2.211.

[41] Hunt Allcott, Matthew Gentzkow, and Chuan Yu. "Trends in the diffusion of misinformation on social media". In: *Research and Politics* 6 (2 2019). DOI: 10.1177/2053168019848554.

[42] Peter Burger et al. "The reach of commercially motivated junk news on Facebook". In: *PLoS One* 14 (8 Aug. 2019). DOI: 10.1371/journal.pone.0220446.

[43] Andrew M. Guess, Brendan Nyhan, and Jason Reifler. "Exposure to untrustworthy websites in the 2016 US election". In: *Nature Human Behavior* (Mar. 2020). DOI: 10.1038/s41562-020-0833-x.

[44] Arkaitz Zubiaga et al. "Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads". In: *PLoS One* 11 (3 Mar. 2016). DOI: 10.1371/journal.pone.0150989.

[45] Philip N. Howard et al. *Social Media, News and Political Information during the US Election: Was Polarizing Content Concentrated in Swing States?* Tech. rep. University of Oxford, Sept. 28, 2017. URL: Bradshaw,%20Samantha%20and.

[46] Chengcheng Shao et al. "The spread of low-credibility content by social bots". In: *Nature Communications* (Nov. 2018). DOI: 10.1038/s41467-018-06930-7.

[47] Soroush Vosoughi, Deb Roy, and Sinan Aral. "The spread of true and false news online". In: *Science* 359.6380 (Mar. 9, 2018). DOI: 10.1126/science.aap9559.

[48] Adam Badawy, Kristina Lerman, and Emilio Ferrara. "Who Falls for Online Political Manipulation?" In: *Companion Proceedings of the The Web Conference (WWW) 2019*. May 2019. DOI: 10.1145/3308560.3316494.

[49] Alexandre Bovet and Hernán A. Makse. "Influence of fake news in Twitter during the 2016 US presidential election". In: *Nature Communications* (Jan. 2019). DOI: 10.1038/s41467-018-07761-2.

[50] Nir Grinberg et al. "Fake news on Twitter during the 2016 U.S. presidential election". In: *Science* 363.6425 (Jan. 25, 2019). DOI: 10.1126/science.aau2706.

[51] Srijan Kumar, Robert West, and Jure Leskovec. "Disinformation on the Web: Impact, Characteristics, and Detection of Wikipedia Hoaxes". In: *Proceedings of the 25th World Wide Web Conference (WWW)*. Apr. 2016. DOI: 10.1145/2872427.2883085.

[52] Danaë Metaxa and Nicolàs Torres-Echeverry. *Google's Role in Spreading Fake News and Misinformation*. Tech. rep. Stanford Law School Law and Policy Lab, Nov. 5, 2017. DOI: 10.2139/ssrn.3062984.

[53] Daniel Bush and Alex Zaheer. *Bing's Top Search Results Contain an Alarming Amount of Disinformation*. Stanford Internet Observatory Blog. Dec. 27, 2019. URL: https://cyber.fsi.stanford.edu/io/news/bing-search-disinformation (visited on 05/15/2020).

[54] Adam Fourney et al. "Geographic and Temporal Trends in Fake News Consumption During the 2016 US Presidential Election". In: *Proceedings of the 26th ACM International Conference on Information and Knowledge Management (CIKM)*. Nov. 2017. DOI: 10.1145/3132847.3133147.

[55] Chengcheng Shao et al. "Anatomy of an online misinformation network". In: *PLoS One* 13 (4 Apr. 2018). DOI: 10.1371/journal.pone.0196087.

[56] Emilio Ferrara. "Disinformation and Social Bot Operations in the Run Up to the 2017 French Presidential Election". In: *First Monday* 22 (8 Aug. 2017). DOI: 10.2139/ssrn.2995809.

[57] Michele Cantarella, Nicolò Fraccaroli, and Roberto Volpe. *Does Fake News Affect Voting Behaviour?* Tech. rep. DEMB Working Paper Series n.146, June 12, 2019. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3402913.

[58] Kate Starbird et al. "Ecosystem or Echo-System? Exploring Content Sharing across Alternative Media Domains". In: *AAAIICWSM2018*. URL: https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17836.

[59] Claire Wardle and Hossein Derakhshan. *Information Disorder: Toward an interdisciplinary framework for research and policy making*. Tech. rep. Council of Europe, Sept. 27, 2017. URL: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c.

[60] Samantha Bradshaw and Phillip N. Howard. *Why Does Junk News Spread So Quickly Across Social Media?* Tech. rep. Oxford Internet Institute, Jan. 28, 2018. URL: https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/142/original/Topos_KF_White-Paper_Howard_V1_ado.pdf.

[61] Priyanjana Bengani. *Hundreds of 'pink slime' local news outlets are distributing algorithmic stories and conservative talking points*. Columbia Journalism Review. Dec. 18, 2019. URL: https://www.cjr.org/tow_center_reports/hundreds-of-pink-slime-local-news-outlets-are-distributing-algorithmic-stories-conservative-talking-points.php (visited on 02/15/2020).

[62] Tara Calishain. *Look out for Junk Sources in Google News*. ResearchBuzz. Nov. 24, 2019. URL: https://researchbuzz.me/2019/11/24/look-out-for-junk-sources-in-google-news/ (visited on 02/15/2020).

[63] Alex Kasprak et al. *Hiding in Plain Sight: PAC-Connected Activists Set Up 'Local News' Outlets*. Snopes. Mar. 4, 2019. URL: https://www.snopes.com/news/2019/03/04/activists-setup-local-news-sites/ (visited on 02/15/2020).

[64] David M. J. Lazer et al. "The science of fake news". In: *Science* 359.6380 (Mar. 9, 2018). DOI: 10.1126/science.aao2998.

[65] Leon Yin et al. *Your Friendly Neighborhood Troll:The Internet Research Agency's Use of Local andFake News in the 2016 US Presidential Campaign*. Tech. rep. SMaPP Data Report, Jan. 2018. URL: https://s18798.pcdn.co/smapp/wp-content/uploads/sites/1693/2018/11/SMaPP_Data_Report_2018_01_IRA_Links.pdf.

[66] Stephan Lewandowsky et al. "Understanding persuasive elements in phishing e-mails: A categorical content and semantic network analysis". In: *Psychological Science in the Public Interest (PSPI)* 13 (3 Dec. 2012). DOI: 10.1177/1529100612451018.

[67] Edson C. Tandoc Jr., Zheng Wei Lim, and Richard Ling. "Defining "Fake News"". In: *Digital Journalism* 6 (2 Feb. 2018). DOI: 10.1080/21670811.2017.1360143.

[68] James H. Kuklinski et al. "Misinformation and the Currency of Democratic Citizenship". In: *The Journal of Politics (J Polit)* 62 (3 Aug. 2000). DOI: 10.1111/0022-3816.00033.

[69] Rakoen Maertens, Frederik Anseel, and Sander van der Linden. "Combatting climate change misinformation: Evidence for longevity of inoculation and consensus messaging effects". In: *Environmental Psychology* (June 12, 2020). (Pre-print). DOI: 10.1016/j.jenvp.2020.101455.

[70] Christina Nemr and William Gangware. *Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age*. Tech. rep. Park Advisors, Mar. 2019. URL: https://www.park-advisors.com/disinforeport.

[71] Edith Greene, Marlene S Flynn, and Elizabeth F Loftus. "Inducing resistance to misleading information". In: *The European Journal of Psychology Applied to Legal Context* 7 (1 Jan. 2015). DOI: 10.1016/S0022-5371(82)90571-0.

[72] Ullrich KH Ecker, Stephan Lewandowsky, and David TW Tang. "Explicit warnings reduce but do not eliminate the continued influence of misinformation". In: *Memory & Cognition (MC)* 38 (8 Dec. 2010). DOI: 10.3758/MC.38.8.1087.

[73] Goldman, Eric. *Content Moderation Strategies*. Presentation at Princeton University CITP Luncheon Series. Apr. 9, 2019.

[74] Emily Taylor, Stacie Walsh, and Samantha Bradshaw. *Industry Responses to the Malicious Use of Social Media*. NATO Stratcom CoE. Oct. 9, 2018. URL: https://www.stratcomcoe.org/industry-responses-malicious-use-social-media (visited on 06/09/2020).

[75] Masahiro Kimura, Kazumi Saito, and Hiroshi Motoda. "Minimizing the Spread of Contamination by Blocking Links in a Network". In: *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*. July 2008. DOI: 10.1145/1357054.1357219.

[76] Ceren Budak, Divyakant Agrawal, and Amr El Abbadi. "Limiting the spread of misinformation in social networks". In: *Proceedings of the 20th World Wide Web Conference (WWW)*. Mar. 2011. DOI: 10.1145/1963405.1963499.

[77] Nam P Nguyen et al. "Containment of misinformation spread in online social networks". In: *Proceedings of the 4th ACM Conference on Web Science (WebSci)*. June 2012. DOI: 10.1145/2380718.2380746.

[78] Senzhang Wang et al. "Negative influence minimizing by blocking nodes in social networks". In: *Proceedings of the 27th AAAI Conference on Artificial Intelligence (AAAI)*. July 2013.

[79] Mehrdad Farajtabar et al. "Fake News Mitigation via Point Process Based Intervention". In: *Proceedings of the 34th International Conference on Machine Learning (ICML)*. Aug. 2017. DOI: 10.5555/3305381.3305495.

[80] Yi Li, Ruidong Yan, and Weili Wu. "Nonlinear Combinatorial Optimization". In: 2019. Chap. Optimization on Content Spread in Social Network Studies. DOI: 10.1007/978-3-030-16194-1_13.

[81] Twitter. *Disclosing new data to our archive of information operations*. Twitter Company Blog. Sept. 20, 2019. URL: https://blog.twitter.com/en_us/topics/company/2019/info-ops-disclosure-data-september-2019.html (visited on 02/07/2020).

[82] Nathaniel Gleicher and Oscar Rodriguez. *Removing Additional Inauthentic Activity from Facebook*. Facebook Newsroom. Aug. 11, 2018. URL: https://about.fb.com/news/2018/10/removing-inauthentic-activity/ (visited on 02/07/2020).

[83] Nick Clegg. *Combating COVID-19 Misinformation Across Our Apps*. Facebook Newsroom. Mar. 25, 2020. URL: https://about.fb.com/news/2020/03/combating-covid-19-misinformation/ (visited on 04/18/2020).

[84] Jeff Smith, Grace Jackson, and Seetha Raj. *Designing Against Misinformation*. Facebook Design. Dec. 20, 2017. URL: https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2 (visited on 02/07/2020).

[85] Manish Singh. *WhatsApp tests new feature to fight misinformation: Search the web*. TechCrunch. Mar. 21, 2020. URL: https://techcrunch.com/2020/03/21/whatsapp-search-web-coronavirus/ (visited on 04/18/2020).

[86] Michael S. Wogalter. "Handbook of warnings". In: Jan. 20, 2006. Chap. Purposes and scope of warnings. DOI: 10.1201/9781482289688. URL: http://www.safetyhumanfactors.org/wp-content/uploads/2011/12/269wogalter2006Chap11.pdf.

[87] Jillian C. York, David Greene, and Geniie Gebhart. *Censorship Can't Be The Only Answer to Disinformation Online*. Electronic Frontier Foundation. May 1, 2019. URL: https://www.eff.org/deeplinks/2019/05/censorship-cant-be-only-answer-disinformation-online (visited on 04/18/2020).

[88] Daniela C. Manzi. "Managing the Misinformation Marketplace: The First Amendment and the Fight Against Fake News and the Fight Against Fake News". In: *Fordham Law Review* 87 (6 2019). URL: https://ir.lawnet.fordham.edu/flr/vol87/iss6/12/.

[89] Gordon Pennycook, Tyrone D Cannon, and David G Rand. "Prior Exposure Increases Perceived Accuracy of Fake News". In: *The Journal of Experimental Psychology: General* 147 (12 Dec. 2018). DOI: 10.1037/xge0000465.

[90] Björn Ross et al. "Fake News on Social Media: The (In)Effectiveness of Warning Messages". In: *ICIS 2018 Proceedings*. Dec. 2018. URL: https://aisel.aisnet.org/icis2018/social/Presentations/16/.

[91] Mingkun Gao et al. "To Label or Not to Label: The Effect of Stance and Credibility Labels on Readers' Selection and Perception of News Articles". In: *Proceedings of the ACM on Human-Computer Interaction (PACM HCI)*. Nov. 2018. DOI: 10.1145/3274324.

[92] Katherine Clayton et al. "Real Solutions for Fake News? Measuring the Effectiveness of General Warnings and Fact-Check Tags in Reducing Belief in False Stories on Social Media". In: *Political Behavior* (Feb. 2019). DOI: 10.1007/s11109-019-09533-0.

[93] Gordon Pennycook et al. "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings". In: *Management Science* (Feb. 21, 2020). DOI: 10.1287/mnsc.2019.3478.

[94] Haeseung Seo, Aiping Xiong, and Dongwon Lee. "Trust It or Not: Effects of Machine-Learning Warnings in Helping Individuals Mitigate Misinformation". In: *Proceedings of the 10th ACM Conference on Web Science (WebSci)*. June 2019. DOI: 10.1145/3292522.3326012.

[95] Guy Rosen et al. *Helping to Protect the 2020 US Elections*. Facebook Newsroom. Aug. 21, 2019. URL: https://about.fb.com/news/2019/10/update-on-election-integrity-efforts/ (visited on 04/18/2020).

[96] Ben Collins. *Twitter is testing new ways to fight misinformation — including a community-based points system*. NBC News. Feb. 20, 2020. URL: https://www.nbcnews.com/tech/tech-news/twitter-testing-new-ways-fight-misinformation-including-community-based-points-n1139931 (visited on 04/18/2020).

[97] Leticia Bode and Emily K Vraga. "In Related News, That Was Wrong: The Correction of Misinformation Through Related Stories Functionality in Social Media". In: *Journal of Communication* 65 (4 Aug. 2015). DOI: 10.1111/jcom.12166.

[98] R. Kelly Garrett and Brian E. Weeks. "The promise and peril of real-time corrections to political misperceptions". In: *Proceedings of the 2013 ACM Conference on Computer Supported Cooperative Work (CSCW)*. Feb. 2013. DOI: 10.1145/2441776.2441895.

[99] Stuart E Schechter et al. "The Emperor's New Security Indicators". In: *Proceedings of the 28th IEEE Symposium on Security & Privacy (S&P)*. May 2007. DOI: 10.1109/SP.2007.35.

[100] Andreas Sotirakopoulos, Kirsite Hawkey, and Konstantin Beznosov. "On the challenges in usable security lab studies: lessons learned from replicating a study on SSL warnings". In: *Proceedings of the 7th Symposium On Usable Privacy and Security (SOUPS)*. June 2011. DOI: 10.1145/2078827.2078831.

[101] Lorrie F Cranor. "A framework for reasoning about the human in the loop". In: *Proceedings of the 1st Conference on Usability, Psychology, and Security (UPSEC)*. Apr. 14, 2008. URL: https://www.usenix.org/legacy/event/upsec08/tech/full_papers/cranor/cranor.pdf.

[102] Vincent C. Conzola and Michael S. Wogalter. " A Communication–Human Information Processing (C–HIP) approach to warning effectiveness in the workplace". In: *Journal of Risk Research* 4 (4 Apr. 15, 2011). DOI: 10.1080/13669870110062712.

[103] Adrienne Porter Felt et al. "Rethinking Connection Security Indicators". In: *Proceedings of the 12th Symposium On Usable Privacy and Security (SOUPS)*. June 2016. URL: https://www.usenix.org/system/files/conference/soups2016/soups2016-paper-porter-felt.pdf.

[104] David W. Stewart and Ingrid M. Martin. "Intended and Unintended Consequences of Warning Messages: A Review and Synthesis of Empirical Research". In: *Journal of Public Policy & Marketing (JPP&M)* 13 (1 Mar. 1994). DOI: 10.1177/074391569401300101.

[105] Anthony Vance et al. "Tuning out security warnings: a longitudinal examination of habituation through fMRI, eye tracking, and field experiments". In: *Management Information Systems Quarterly (MISQ)* 42 (2 June 2018). DOI: 10.25300/MISQ/2018/14124.

[106] Dan M. Kahan. *Ideology, Motivated Reasoning, and Cognitive Reflection: An Experimental Study*. Tech. rep. Yale Law School Cultural Cognition Project, Nov. 29, 2012. DOI: 10.2139/ssrn.2182588.

[107] Gordon Pennycook and David G. Rand. "Fighting misinformation on social media using crowdsourced judgments of news source quality". In: *Proceedings of the National Academy of Sciences (PNAS)* 116 (7 Feb. 2019). DOI: 10.1073/pnas.1806781116.

[108] Tor Lattimore and Csaba Szepesvàri. *Bandit Algorithms*. July 31, 2020. URL: https://tor-lattimore.com/downloads/book/book.pdf.

[109] Feifang Hu and William F. Rosenberger. *The Theory of Response-Adaptive Randomization in Clinical Trials*. Apr. 10, 2006. DOI: 10.1002/047005588X.

[110] John White. *Bandit Algorithms for Website Optimization*. Jan. 23, 2013.

[111] Molly Offer-Westort, Alexander Coppock, and Donald P. Green. *Adaptive Experimental Design: Prospects and Applications in Political Science*. Tech. rep. Feb. 12, 2019. DOI: 10.2139/ssrn.3364402.

[112] *Pre-registered analysis plan*. OSF. URL: https://osf.io/q3ymr.

[113] Gordon Pennycook and David G Rand. "Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning". In: *Cognition* 188 (July 2019). DOI: 10.1016/j.cognition.2018.06.011.

[114] Kevin Munger et al. "The (Null) Effects of Clickbait Headlines on Polarization, Trust, and Learning ". In: *Public Opinion Quarterly* (Apr. 30, 2020). DOI: 10.1093/poq/nfaa008.

[115] Andrew Guess, Jonathan Nagler, and Joshua Tucker. "Less than you think: Prevalence and predictors of fake news dissemination on Facebook". In: *Science Advances* 5.1 (2019). DOI: 10.1126/sciadv.aau4586.

[116] Ben Tappin, Gordon Pennycook, and David Rand. *Rethinking the Link Between Cognitive Sophistication and Politically Motivated Reasoning*. Dec. 3, 2018. PsYarXiv: 10.31234/osf.io/yuzfj.

[117] Robert Ross, David Rand, and Gordon Pennycook. *Beyond "fake news": The role of analytic thinking in the detection of inaccuracy and partisan bias in news headlines*. Nov. 13, 2019. PsYarXiv: 10.31234/osf.io/cgsx6.

[118] Brian F. Schaffner and Cameron Roche. "Misinformation and Motivated Reasoning: Responses to Economic News in a Politicized Environment ". In: *Public Opinion Quarterly* 81 (1 Mar. 1, 2017). DOI: 10.1093/poq/nfw043.

[119] Dan M. Kahan. *Misconceptions, Misinformation, and the Logic of Identity-Protective Cognition*. Tech. rep. Yale Law School Cultural Cognition Project, May 24, 2017. DOI: 10.2139/ssrn.2973067.

[120] Instagram. *Combatting Misinformation on Instagram*. Instagram Announcements. Dec. 16, 2019. URL: https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram/ (visited on 05/13/2020).

[121] Yoel Roth and Nick Pickles. *Updating our Approach to Misleading Information*. Twitter Product Blog. May 11, 2020. URL: https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html (visited on 05/13/2020).

[122] *Party Identification 3-Point Scale 1952 -2016*. American National Election Studies (ANES). URL: https://electionstudies.org/resources/anes-guide/top-tables/?id=22 (visited on 04/18/2020).

A Laboratory Study Interview Questions

1. How would you define disinformation?
2. What is the difference between disinformation and misinformation?
3. What features or elements of a website would make you think that a site may contain disinformation?
4. What do you look for in a site before you consider it trustworthy?
5. Have you ever encountered disinformation before this study?
6. What was your first reaction when you saw the disinformation warning?
7. Have you encountered browser warnings before?
8. Do you know what these browser warnings were warning against?
9. How effective are browser warnings in your experience?
10. How effective do you think browser warnings are for the average internet user?
11. Between a full-page and an in-context search result warning, which browser warning do
12. you think would be more effective in deterring users from clicking through or consuming information from a website?
13. Why do you think your chosen browser warning type is more effective?
14. Would you prefer a browser with or without warnings?
15. Do you consider browser warnings inconvenient?
16. How would you improve browser warnings to be more effective?
17. What is your age?
18. What is your gender?
19. What is your nationality?
20. What is your area of study?
21. Do you have any experience studying disinformation or related areas?
22. How many minutes do you spend reading the news on average per day?
23. On a scale of 1 to 5, with 1 being strongly liberal and 5 being strongly conservative, where would you put yourself?

B Crowdworker Study Supporting Materials

B.1 Warning Designs

We show the contents of all eight warning designs in Table 4. Figure 3 in Section 4 shows examples of informative and harm-based warnings.

B.2 Multi-armed Bandit Implementation

Our bandit was based on the ε-greedy algorithm [108], which consists of two phases. In exploration, the algorithm assigned arms randomly to subjects. After a defined number of exploration iterations e, the algorithm moved to exploitation, where it assigned the top-performing arm (i.e., the arm that maximized the reward function) with a probability of 1-ε; with probability ε it assigned a random arm. Since subjects were each assigned two warnings, the bandit selected from the ε-distribution independently for each arm.

Each of our eight warning designs corresponded to two bandit arms: one for harm scores and one for informativeness scores. In the two treatment rounds of the experiment, the bandit drew one arm using the harm scores and the other using the informativeness scores, with the draw order randomized. Further, in order to answer RQ4, we modeled subjects' partisan alignment as a mediator variable and maintained independent response sets for liberals and conservatives. Therefore our bandit was divided into four distributions of eight arms each: liberal-harm, liberal-informative, conservative-harm, and conservative-informative.

Bandit Reward Function Each time the bandit selected an arm (to assign a warning to a subject in a treatment round), it computed the current reward for all arms and used these reward values to weight the distribution from which it drew. Arms that had a larger reward value were more likely to be selected.

The primary criteria that affected our reward function was *extremeness*. Arms that had received significantly higher or lower mean ratings than the average arm accrued larger rewards so that subjects were more likely to be assigned to top and bottom performing arms.

The secondary criteria was *precision*. Our reward function preferred arms for which we had high confidence that successive ratings would not significantly change the mean rating. In other words, when many subjects had provided similar ratings for an arm, that arm received a larger reward, and when there were few ratings or those ratings were divergent, the reward was smaller. Including the precision component in the reward function was intended to reduce the confidence intervals for the mean ratings of the selected arms.

The reward function for arm a, $f(a)$, was the weighted sum of the extremeness component e_a and the precision component p_a. Before summing, each component was multiplied by a weight, w_e and w_p, such that $w_e + w_p = 1$ and $w_e \ll w_p$.

$$f_a(t) = w_e(e_a) + w_p(p_a)$$

We computed e_a as the deviation of the arm's mean rating ($\bar{x}_a$) from the mean rating of all n arms.

$$e_a = |\bar{x}_a - \Sigma_{k=0}^{n}(\bar{x}_k)|$$

We computed p_a as the number of observations (N_a) the arm had received divided by the variance (σ_a^2) in ratings. Thus p_a grew as variance decreased (because subjects were agreeing) and shrank as the number of observations grew (because the next marginal observation became less likely to change the mean score). We also used a function $f(\sigma_a^2)$ to avoid undefined values for p_a in the cases where σ_a^2 was zero (because

Table 4: We developed 8 interstitial warning designs for the crowdworker study.

	Harm (white on red background)				Informativeness (black on white background)			
ID	**h1**	**h2**	**h3**	**h4**	**i1**	**i2**	**i3**	**i4**
Icon	Skull	Skull	Policeman	Policeman	Exclamation	Policeman	Policeman	Exclamation
Title	WARNING	Security Alert	Security Alert	Security Alert	False or Misleading Content Warning	Fake News Warning	False or Misleading Content Warning	Fake News Warning
Primary message	This website is dangerous.	This website contains misleading or false information.	This website is dangerous.	This website contains misleading or false information.	This website presents itself as news, but it contains information that experts have identified to be false or misleading	This website contains misleading or false information.	This website contains misleading or false information.	This website presents itself as news, but it contains information that experts have identified to be false or misleading
Details	None	None	Consider finding alternative sources of information.	None	This website spreads disinformation: lies, half-truths, and non-rational arguments intended to manipulate public opinion. It can be difficult to tell the difference between real news and disinformation, but it poses a serious threat to national security, election integrity, and democracy.	This website spreads disinformation: lies, half-truths, and non-rational arguments intended to manipulate public opinion. It can be difficult to tell the difference between real news and disinformation, but it poses a serious threat to national security, election integrity, and democracy.	Consider finding alternative sources of information.	Consider finding alternative sources of information.

all ratings were equal) or undefined (because there were no ratings yet). z was a parameter value close to zero.

$$p_a = \frac{N_a}{f(\sigma_a^2)}$$

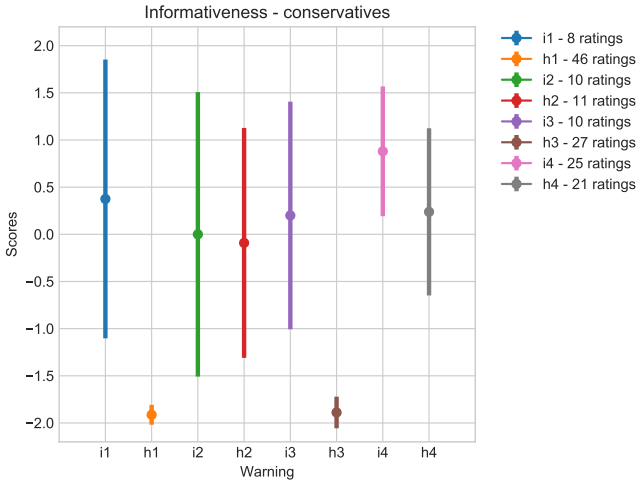
$$f(\sigma_a^2) = \begin{cases} z & \text{if } \sigma_a^2 = 0 \text{ or } N_a = 0 \\ \sigma_a^2 & \text{else} \end{cases}$$

Simulation and Parameter Selection We designed a simulation procedure to guide selection of assignments for the bandit's parameters, which include the number of exploration rounds e, the randomness factor ε, the weights of the two reward components w_p, w_e, and the constant replacement value z used when an arm's variance is 0. The simulation generated user ratings by drawing from discrete normal distributions with static biases assigned to a few arms to encode our belief that some arms would receive categorically higher or lower scores than others. For harm ratings, we assigned a $+1$ bias to one arm and a -1 bias to another arm at random. For informativeness ratings, we assigned two positive and two negative biases. These biases differ for the two types of ratings because pilot tests of the warning designs reported generally neutral harm scores for most warnings but more diverse opinions about informativeness.

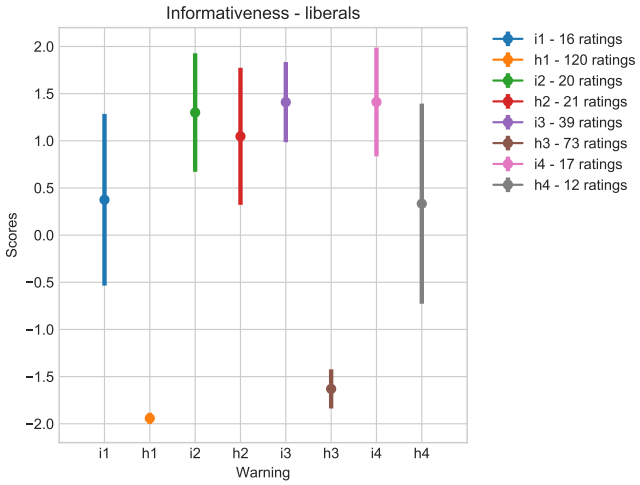
Table 5: For each parameter of the bandit algorithm, we simulated a range of possible assignments, then selected the final assignments that achieved optimal performance.

	e	ε	w_p	w_e	z
Range	$[30, 40, \ldots, 90]$	$[.05, .1, \ldots, 0.2]$	$[.1, .2, \ldots, -0.9]$	$1 - w_p$	0.2
Final	80	0.2	0.1	0.9	0.2

We simulated the experiment with $n = 250$, marking a simulation as a success if top- and bottom- performers were clearly established (i.e., if the arms with the highest and lowest mean scores had nonoverlapping 95% confidence intervals). We first simulated 224 permutations of assignments 20 times each, then selected assignments with a success rate $\geq 85\%$ and simulated those an additional 100 times. Among the configurations that achieved at least 90% success in the second round of tests, we selected the configuration with the lowest value for e as we wanted to maximize exploitation. The set of parameters we test and the final assignment is shown in Table 5.

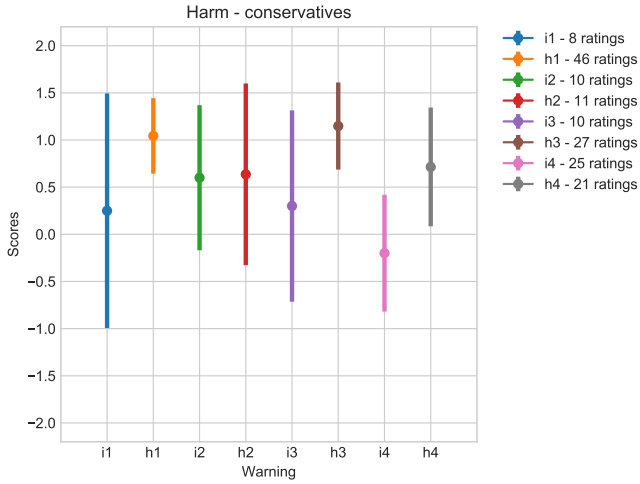


(a)

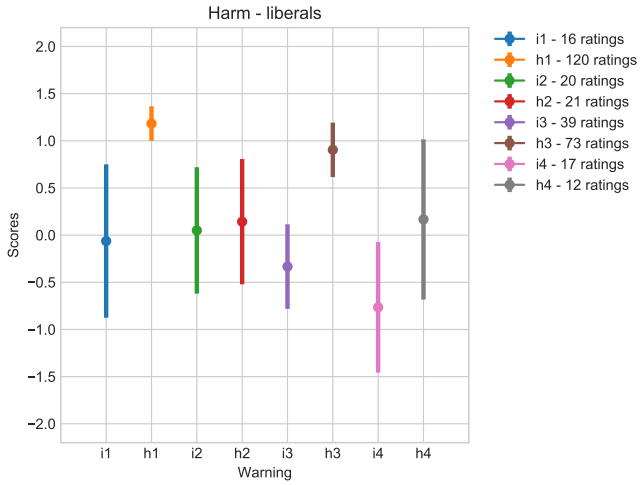


(b)

Figure 6: Mean informativeness ratings and 95% confidence intervals for conservative participants (6a) and liberal participants (6b).



(a)



(b)

Figure 7: Mean harm ratings and 95% confidence intervals for conservative participants (7a) and liberal participants (7b).

B.3 Survey Questions and Responses

Each participant in the crowdworker study answered a series of personal survey questions. We show the questions and response distributions (Table 6) below.

1. How old are you? (18-120)
2. What is your gender? (M, F, Non-binary, Prefer not to answer)
3. Are you currently pursuing education? If so, at what level? (No, High school or equivalent, Undergraduate degree, Master's degree, Professional Degree, Doctorate)
4. What is the highest level of education you have completed? (High school or equivalent, Some college, Undergraduate degree, Master's degree, Professional Degree, Doctorate)
5. When it comes to politics, how would you describe yourself? (Liberal, Conservative, Neither liberal nor conservative) [7]
    - (if Liberal selected) Would you call yourself a strong liberal or a not very strong liberal?
    - (if Conservative selected) Would you call yourself a strong conservative or not a very strong conservative?
    - (if Neither selected) Do you think of yourself as closer to a conservative, closer to a liberal, or neither?
6. During a typical week, how many days do you watch, read, or listen to news on TV, radio, printed newspapers, or the Internet, not including sports? (0-7)
7. How much attention do you pay to news about national politics on TV, radio, printed newspapers, or the Internet? (A great deal, a lot, a moderate amount, a little, or none at all)

Table 6: Response distributions for the demographic survey questions asked during the crowdworker study.

	Liberal	Conservative	Total
Total	159	79	238
Gender			
Male	101	46	147
Female	54	33	87
Non-binary	3	–	3
N/A	1	–	1
Age			
18-29	43	16	59
30-49	92	46	138
50-64	22	12	34
65+	2	5	7
Education			
HS	25	16	41
Up to undergrad	116	59	175
Beyond undergrad	18	4	22
Media Consumption (0-7) and Attention (1-5)			
5+ days, 4-5 attention	63	32	95
5+ days, 2-3 attention	45	23	68
2-4 days, 4-5 attention	8	3	11
2-4 days, 2-3 attention	37	14	51
0-1 days, 1-4 attention	6	7	13