A First Look into Targeted Clickbait and its Countermeasures: The Power of Storytelling

Ankit Shrestha
ankit.shrestha@usu.edu
Utah State University
Logan, Utah, USA

Audrey Flood
audrey.flood@usu.edu
Utah State University
Logan, Utah, USA

Saniat Javid Sohrawardi
saniat.s@mail.rit.edu
Rochester Institute of Technology
Rochester, New York, USA

Matthew Wright
matthew.wright@rit.edu
Rochester Institute of Technology
Rochester, New York, USA

Mahdi Nasrullah Al-Ameen
mahdi.al-ameen@usu.edu
Utah State University
Logan, Utah, USA

ABSTRACT

Clickbait headlines work through superlatives and intensifiers, creating information gaps to increase the relevance of their associated links that direct users to time-wasting and sometimes even malicious websites. This approach can be amplified using *targeted clickbait* that takes publicly available information from social media to align clickbait to users' preferences and beliefs. In this work, we first conducted preliminary studies to understand the influence of targeted clickbait on users' clicking behavior. Based on our findings, we involved 24 users in the participatory design of story-based warnings against targeted clickbait. Our analysis of user-created warnings led to four design variations, which we evaluated through an online survey over Amazon Mechanical Turk. Our findings show the significance of integrating information with persuasive narratives to create effective warnings against targeted clickbait. Overall, our studies provide valuable insights into understanding users' perceptions and behaviors towards targeted clickbait, and the efficacy of story-based interventions.

CCS CONCEPTS

• **Human-centered computing → Human computer interaction (HCI)**; **User studies**; • **Security and privacy → Human and societal aspects of security and privacy**.

KEYWORDS

targeted clickbait, storytelling, qualitative study, quantitative study, interventions

ACM Reference Format:
Ankit Shrestha, Audrey Flood, Saniat Javid Sohrawardi, Matthew Wright, and Mahdi Nasrullah Al-Ameen. 2024. A First Look into Targeted Clickbait and its Countermeasures: The Power of Storytelling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 23 pages. https://doi.org/10.1145/3613904.3642301

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642301

1 INTRODUCTION

Clickbait is a text or a thumbnail link designed to attract attention and entice users to follow that link to visit the linked piece of online content, which is typically deceptive, sensationalized, or otherwise misleading [127].[1] Even worse, clickbait is often used in social engineering attacks, tricking users to click on posts or links that direct them to malicious websites [2, 10, 38, 107, 116, 118, 139, 165]. Since clickbait threatens online security, there have been many attempts to limit it both from industry efforts [20, 67, 90, 143] and academic research [18, 72, 129, 147]. Clickbait, however, is still effective [11, 33, 98]. This can be explained through relevance theory [140, 141], which explains that humans are driven to take in the most relevant information. Clickbait uses definite referring expressions together with superlatives and intensifiers to create an *information gap*, which drives the reader to click on the associated link with the expectation to find relevant information [127].

Relevance theory further states that a stimulus (clickbait) will be optimally relevant if it aligns with the reader's preferences [140, 141]. That explains why clickbait on topics that a user is not inherently interested in may not be effective. Unfortunately, with the wide use of social networking sites, public information about users, including their affiliated institutions, interests, and even their friends, are readily available [66, 79]. Users can be targeted on social media with posts inherently relevant to them using their public information [7, 79, 85]. Relevance theory would suggest that this kind of *targeted clickbait* should be even more effective than non-targeted clickbait.

Therefore, it is important to first understand the public information that could be used to create targeted clickbait *(targeting factors)* and the *countermeasures* that could protect against it **(RQ1)**. To address these initial needs, we conducted a focus group discussion (FGD) with six participants and an online survey with 30 participants in North America (see Figure 1a), which focused on the following: i) brainstorming targeting factors, ii) brainstorming ideas for *(countermeasures)*, and iii) selecting the most effective targeting factors and countermeasures.

Based on the findings from these studies, we developed *story-based interstitial interventions* against targeted clickbait by involving 24 end-users in a participatory design study (see Figure 1b). In this study, we focused on interstitial warnings that interrupt users'

[1] https://en.wikipedia.org/wiki/Clickbait

primary tasks, considering their proven efficacy to protect against phishing and malware [72, 117]. Our focus on storytelling is due to the effectiveness of this approach in persuading users [37, 91, 137].[2] People often engage in risky behavior despite knowing the potential harms [28, 50], but storytelling could be a powerful tool in changing such behavior [54, 71, 137]. Within the framework of story-based interstitial interventions, participants reflected on a key remaining question: what stories should we tell? In other words, our study answers, "How can we leverage user-informed stories in designing interstitial interventions to protect users from targeted clickbait?" **(RQ2)**. The existing literature [17, 101] suggests that involving end users from an early stage results in improved designs. To this end, we leveraged participatory design [45, 59, 155], where participants were facilitated with graphics to create their version of the warning.

The findings from our participatory design study provide us with concepts that users believe should guide the story in a warning against targeted clickbait. These concepts then inform our final warnings that we evaluated through an online survey with 114 participants (see Figure 1c). The study answers, "How do social media users perceive and behave towards user-informed story-based interventions designed to help them make informed decisions about targeted clickbait?" **(RQ3)**. The results from this evaluation survey indicate that user-informed stories can effectively support social media users to understand and counter clickbait.

In summary, our findings first contribute to the identification and taxonomy of targeting factors in social media and the countermeasures against targeted clickbait. Based on these targeting factors, our studies reveal the effectiveness of targeted clickbait and lead to the creation of story-based interventions based on user participation through ideation and design. Finally, our work evaluates the designs that emerged from these efforts, and finds them to be effective against targeted clickbait. To the best of our knowledge, this is the first systematic exploration of targeted clickbait, its efficacy in tricking users, and the design and evaluation of potential countermeasures through involving end users. Taken together, our studies provide valuable insights into users' perceptions, needs, and behavior around targeted clickbait, and the user-informed stories designed to protect them from clicking on a clickbait. We provide a set of recommendations based on our findings, which include using stories interchangeably to avoid habituation, and adapting warnings based the level of threat.

2 BACKGROUND AND RELATED WORK

Users have been found to be vulnerable to the misleading and sensationalized information provided in a clickbait [53, 65, 111–113, 145]. *Relevance theory* suggests that targeted clickbait would be even more effective [21, 126, 127]. We thus provide background on the working and potential of targeted clickbait in tricking users at the beginning of §2.1, followed by a discussion on our motivation to curate targeting factors **(RQ1)**. In §2.2, we discuss prior studies on clickbait that encourage us to explore warning-based interventions **(RQ2 and RQ3)**. Lastly, §2.3 highlights our respective rationales behind using interstitial, story-driven, and user-informed warnings **(RQ2 and RQ3)**.

2.1 Targeted Clickbait and Targeting Factors

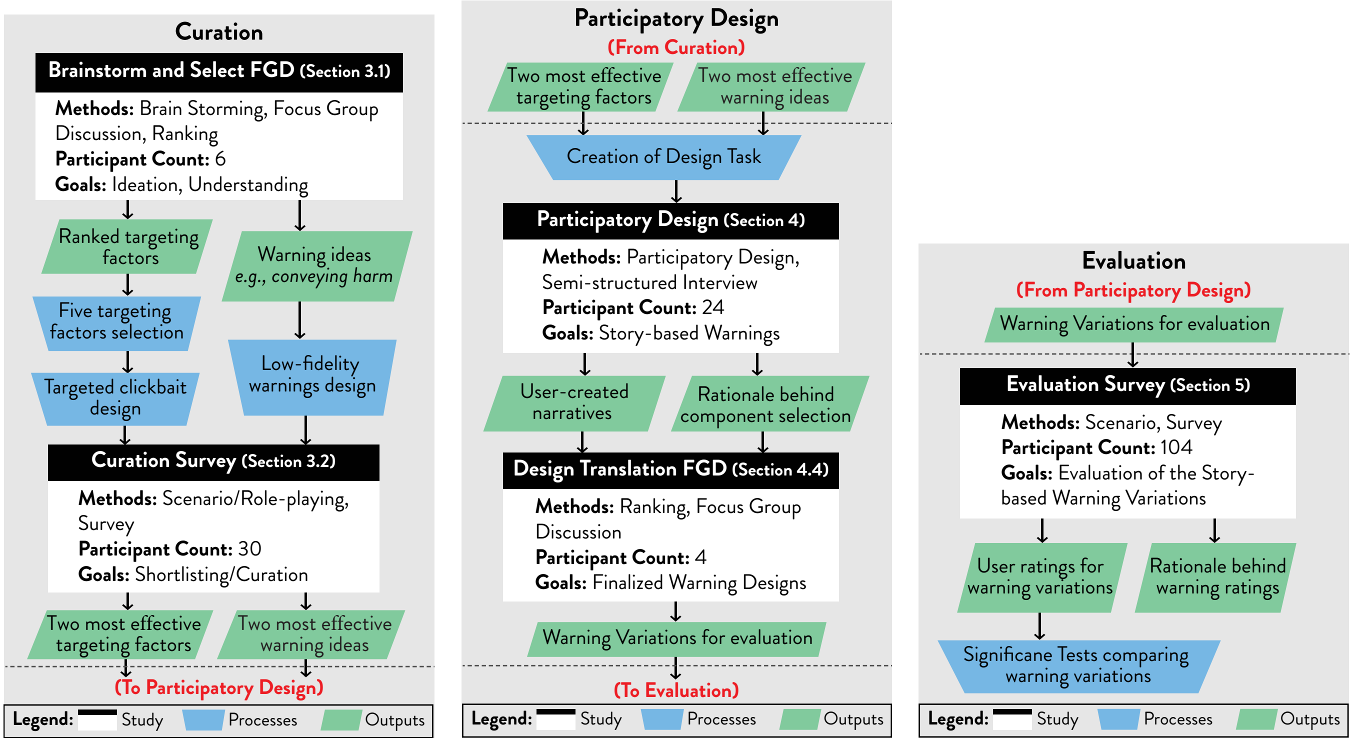
Relevance theory is based on two main principles: the *cognitive principle* and the *communicative principle* [32, 140, 141, 153, 154]. According to cognitive principle, humans are geared to maximizing relevance [32, 140, 141]. Clickbait uses definite referring expressions (e.g. "These were Chávez' last words" use of "These" in place of the words themselves) [21, 126], and superlatives (e.g. terrifying, coolest, genius, unreal) or intensifiers (e.g. ridiculously, crazy, "THIS") to create relevance for its readers [126, 127]. Prior work [89] revealed that these elements contribute to an *information gap* by encouraging readers to construct new conceptual files based on the terms used in a headline while providing little or no content for those files [127]. The instance of an information gap can be seen in a headline such as, "The worst [superlative] day of the week [information gap by hiding simple information] to eat at a restaurant". The information gap then drives a reader to click on the associated link, expecting the article to contain relevant information [127].

In contrast, regular clickbait does not align with the communicative principle, which states that input will be optimally relevant if it is (a) worth the reader's effort to process and (b) the most relevant input allowing for the reader's abilities and preferences [32, 140, 141]. Online communication often occurs in a so-called collapsed context [92, 151]. When offline, we speak to one group in one context, but when online, we communicate across groups of people and contexts [92, 127, 151]. That is true for clickbait creators, as they cannot know who will see and read their work, or when it might be read [127].

This is becoming less true all the time, however. Users often make large amounts of information about themselves available on social media, either publicly or to only loosely defined friend groups, including interests, location, job type, relationship status, and associated institutions. This wealth of information makes it possible to align clickbait to users' preferences and beliefs and thus ensure it is optimally relevant [7, 66, 79, 85, 140, 141]. Advancement in artificial intelligence has made it easy and inexpensive to access high-quality text generation technology to automatically select and produce even more relevant headlines and material based on this information. Furthermore, faceswap and other deepfake image generation technologies are also becoming easier and cheaper to access. A malicious actor can use these tools to weaponize publicly available information about friends to create even more compelling targeted clickbait with pictures and video [13, 30, 49, 52, 76, 77, 129, 144]. As targeted clickbait more strongly aligns with users' beliefs and interests, they may feel that it is worth their effort and time to click on the post and discover the answer [7, 85, 127, 140, 141].

There is a gap in existing literature to understand targeted clickbait and countermeasures against it. We found a few studies [105, 121] in the realm of targeted attacks in general, which mentioned location, friends, relatives, and affiliated institutions as the examples of targeting factors. However, we found a little work that has listed and ranked the targeting factors for targeted clickbait. To this end, we identified and ranked targeting factors and countermeasures against targeted clickbait **(RQ1)**.

[2] https://www.howcommunicationworks.com/blog/2021/5/24/how-to-persuade-people-the-hidden-power-of-stories


Curation
Brainstorm and Select FGD (Section 3.1)
Methods: Brain Storming, Focus Group Discussion, Ranking
Participant Count: 6
Goals: Ideation, Understanding
Ranked targeting factors
Warning ideas e.g., conveying harm
Five targeting factors selection
Targeted clickbait design
Low-fidelity warnings design
Curation Survey (Section 3.2)
Methods: Scenario/Role-playing, Survey
Participant Count: 30
Goals: Shortlisting/Curation
Two most effective targeting factors
Two most effective warning ideas
(To Participatory Design)
Legend: Study Processes Outputs
Participatory Design
(From Curation)
Two most effective targeting factors
Two most effective warning ideas
Creation of Design Task
Participatory Design (Section 4)
Methods: Participatory Design, Semi-structured Interview
Participant Count: 24
Goals: Story-based Warnings
User-created narratives
Rationale behind component selection
Design Translation FGD (Section 4.4)
Methods: Ranking, Focus Group Discussion
Participant Count: 4
Goals: Finalized Warning Designs
Warning Variations for evaluation
(To Evaluation)
Legend: Study Processes Outputs
Evaluation
(From Participatory Design)
Warning Variations for evaluation
Evaluation Survey (Section 5)
Methods: Scenario, Survey
Participant Count: 104
Goals: Evaluation of the Story-based Warning Variations
User ratings for warning variations
Rationale behind warning ratings
Significane Tests comparing warning variations
Legend: Study Processes Outputs

(a) Summary of the curation of targeting factors and countermeasures (see §3)

(b) Summary of the participatory design (see §4)

(c) Summary of the evaluation of warnings (see §5)

Figure 1: Summary of the studies reported in the paper

2.2 Learning from the Studies on Clickbait

Since there are no studies on targeted clickbait, we look into the studies about clickbait in general to begin the process of creating countermeasures against targeted clickbait. Because clickbait is effective [55, 119, 120], prior works aimed at mitigating the problem, where most of them focused on the moderation of clickbait in social media [3, 27, 29, 75, 81, 115, 134, 160, 162, 163]. Moderation often suffers from issues, however, including reliability [35] and false positives/negatives [75] that lead to blocking a substantial amount of non-clickbait posts or vice versa. It is also difficult to estimate how effective the detection and moderation would be for targeted clickbait. While clear cases of clickbait and known malicious sites should be blocked, we argue that less obvious cases can be combated by supporting users to understand the risks in real time.

Only a few works have aimed at supporting end users through interventions to inform and persuade them to avoid clickbait [18, 19, 27, 53, 61, 72, 147]. Bhuiyan et al. [18, 19] developed browser extensions to nudge users to reflect on the credibility of news in social media. A few studies used interventions through identifying clickbait and misinformation, and letting users to block such posts [27, 43, 85, 125]. These studies show some promise in countering clickbait, highlighting the importance of informing and warning users about targeted clickbait. Therefore, we designed and evaluated warning-based interventions **(RQ2 and RQ3)**.

2.3 Reflection, Stories and Users

Warnings can generally be classified into interstitial (blocking) and contextual (non-blocking). Prior work [44, 72, 109, 117, 158] found that users routinely ignore contextual warnings such as banners or pop-ups. They instead notice interstitial interventions that interrupt the user's primary task, allow them to reflect on their actions and respond by seeking information from alternative sources [16, 44, 72, 117, 158]. Here, the reflective design promotes conscious thought and decision-making, and helps users consider their actions [48, 101–103]. The Psychology and Marketing literature [12, 62, 87, 114] support that reflective designs help increase engagement and thoughtful decision-making. The study of Kaiser et al. [72] further informs us that interstitial warnings can effectively inform users about the risk of harm. Therefore, we designed and evaluated interstitial warnings **(RQ2 and RQ3)**.

We combine interstitial warnings with storytelling, leveraging its power to persuade users to avoid targeted clickbait [37, 51, 64, 78, 86, 91, 137]. The prior work [28, 50] referred to cognitive dissonance, a phenomenon representing that even when users know something is bad, they tend to do the action. For instance, people often do not quit smoking despite knowing it can cause lung cancer. In such cases, strategic story, also called *narrative persuasion*, is a powerful tool that combines relevant information with emotion [37, 54, 63, 71, 84, 99, 137]. It changes the persuasive message from a dry listing of information to something that is embedded in a larger narrative, where we get the message by seeing people's stories unfold [37, 54, 63, 78, 91, 99, 137]. Therefore, we leveraged the power of stories to persuade users to avoid targeted clickbait **(RQ2 and RQ3)**.

We focus on a participatory design approach to create stories for the warnings [14, 45, 59, 122, 155]. Designers often overlook the intricate difficulties and challenges end users may face [17, 42, 70, 100, 101, 152]. Prior studies [100, 101, 152] suggest that users are not just the target audience for a design but a collaborative party holding knowledge about its development. These studies also highlight the importance of understanding the perspectives and needs of users from an early stage to facilitate concept generation, as well as increase the adoption of a design [17, 31, 70, 100, 101, 152]. Therefore, we involved end users from the early stages of our design ideation.

While prior studies primarily focused on detecting and moderating regular clickbait, targeted clickbait is yet to be studied. To the best of our knowledge, our work is the first to understand user perspectives towards targeted clickbait, involve users from selecting targeting factors to designing story-based warnings, and evaluate user-informed story-based warnings.

3 TARGETING FACTORS AND COUNTERMEASURES (RQ1)

In §3–§5, we present a series of studies and corresponding findings. A summary of the flow of these studies along with its goal and outputs are presented in Figure 1. For consistency, we use these terms based on the frequency of participants' comments in each study: a few (0-20%), some (21-40%), about half (41-60%), most (61-80%), and almost all (81-100%).

Due to the novelty of targeted clickbait and a lack of prior work in this area, we first need to understand the factors that can be used to create targeted clickbait. Considering possible cognitive load and confusion of participants caused by having too many study variables [34, 74, 93, 96], we broke this process into two studies (see Figure 1a) to curate targeting factors and countermeasures that align with participants' beliefs and interests [60, 104, 136, 138, 149].

3.1 Study I: Brainstorm and Select FGD

We conducted a Focus Group Discussion (FGD) [59, 80, 97] with six participants (FGD1 – FGD6) – including three User Experience (UX) experts – to brainstorm and rank ideas about both targeting factors and countermeasures. The group setting provides a platform to generate ideas and refine them in a single session through collaboration. Table 6 in Appendix B provides the demographic information of our participants.

3.1.1 Methods. We recruited participants through snowball and convenience sampling, where we contacted them via email. The FGD was conducted in person (audio-recorded), and lasted around 75 minutes. The session started with brainstorming targeting factors in social media. Then the participants rated each idea from 1 (not likely to click) to 5 (very likely to click) while providing their rationales. Finally, they were asked to discuss potential countermeasures against targeted clickbait.

We ranked the targeting factors based on participants' ratings, and listed the countermeasures. We transcribed the audio recording and extracted the rationale behind participants' ratings using thematic analysis [15, 22, 40]. Two researchers independently coded the transcript of the focus group, where they read through it and developed codes. Then the codes were compared, and the coders discussed and resolved any discrepancies in the codes.

3.1.2 Findings.

Selection of Targeting Factors. Our participants came up with 12 targeting factors through brainstorming, which we ranked using their ratings. We then selected the following five factors with an average rating greater than the median.

(1) Niche Activities. Participants believed the activities special to them, like skiing or fishing, could create effective targeted clickbait. One participant said, *"I think in this case, one of the factors might be how unique the activities are, so weight loss is pretty common. If they are unique activities that I do, then I will click on it." (FGD1).*
(2) Relevant Location. Participants reported that a place they have recently visited or are planning to visit would make an enticing post. One of them said, *"Like even for the location when we plan for going somewhere, and yeah, the clickbait news pops up in our Facebook like things to do here. That will be so much tempting." (FGD2).*
(3) Field of Study/Work. Participants found their field of study or work to be an effective targeting factor. One participant mentioned, *"If the news is relevant to my research or my field, I would definitely click on it." (FGD4)."*
(4) Friends' images (called *face-swap* hereafter). Participants anticipated that using advancing AI technologies, they could be targeted with fake posts containing face-swapped images of their friends. One participant (FGD1) talked about how the image of the wedding of a friend whom they had not contacted in last ten years could still interest them to know more about their recent life event.
(5) Affiliated Institutions. Participants believed their affiliated institutions could be used to create a compelling clickbait. One of them said, *"I am pretty much inclined to click posts about my university. I definitely would want to know like what happened. ...Even if I knew it might be clickbait, I still might be intrigued to see what was inside." (FGD6).*

List of Countermeasures. The participants brainstormed ideas to counter targeted clickbait, which we present below. Nearly all participants discussed most of these ideas.

(1) *Reporting*. A source (fact checkers, users) reporting the post as misleading. One participant said, *"If it is a really harmful website, people might have reported on that. So like, negative reports [can be helpful]." FGD5.* This could help the site to downgrade the post [106] or warn other users about it.
(2) *Alternate Source.* An addition to the interface that provides an alternate way to find the answer to a question posed by clickbait, e.g. a "Google It" button that leads to a quick search for the information.
(3) *URL credibility.* A warning that raises doubts about the link. One participant proposed, *"Reporting the trustworthiness of the URL itself can be helpful." (FGD1).*

Targeting Factor	Likelihood to click
Face-swap (Friends)	M=4.22, SD=0.92
Affiliated Institutions	M=4.02, SD=0.89
Niche Activities	M=3.85, SD=1.02
Field of Study/Work	M=3.65, SD=1.13
Relevant Location	M=3.53, SD=1.04

Table 1: Likelihood to click on the various targeted clickbait

Countermeasures	Usefulness
Consequences	M=4.03, SD=0.92
Reporting	M=4.00, SD=1.11
URL Credibility	M=3.83, SD=1.11
Alternate Source	M=3.33, SD=1.12
Reveal Mystery	M=3.26, SD=1.36

Table 2: Usefulness of the different countermeasures

(4) *Reveal Mystery*. Providing a summary of answer to the question posed by clickbait. For example, if the headline is "You Won't Believe How Much Money Tom Brady Made During His 22-Year Football Career," then providing the answer "He made roughly $450 million" [148] can diffuse the curiosity generated by the clickbait.

(5) *Consequences.* Showing a warning that conveys the negative consequences (harm) of clicking on a clickbait.

3.1.3 Translation of Findings to Designs. For each targeting factor, we created two posts – one with positive emotional valence and one with negative (see Figures 2a and 2b). Since targeted clickbait increases relevance and feeds on users' emotional reactions [29, 159], we aimed to examine if the change in emotional valence impacts the effectiveness of a targeted clickbait. For the countermeasures, we presented the information in text on top of the post with a red overlay and a generic header (see Figure 2c) aimed at selecting the best idea for countering targeted clickbait.

3.2 Study II: Curation Survey

We selected the two most effective targeted clickbaits that align with users' preferences, and in turn, increases relevance [140, 141]. Similarly, persuasion, which is the basis for behavior change, starts from the values, beliefs, and motives of users [60, 104, 136, 138, 149]. Therefore, we selected the two most effective countermeasures to align with users' values and motives. The Institutional Review Board at our university approved the study.

3.2.1 Methods. Based on the designed targeted clickbait and countermeasures, we conducted a survey with 30 participants. A power analysis indicated that 30 participants would provide a large effect size, which can only detect effects that apply to at least 80% of the population [58, 59]. This effect size is appropriate for this study, given that the goal is to find directions for the next steps: participatory design and evaluation.

Participant Recruitment. We recruited participants over Amazon Mechanical Turk (MTurk). Following the guidelines from prior work [82, 110], we recruited participants with a 99% HIT approval rate [3] to increase the quality of responses in our study. As each design was presented on a separate page with only a single question, i.e., participants did not need to go through multiple questions in a single page, we felt it was reasonable to not include any attention check questions. Participants had to be 18 years or older and live in the United States or Canada to participate in our study. We set the location of target participants as United States or Canada on MTurk, and we confirmed this by geolocating their IP address as recorded by the Qualtrics survey platform. The study took between 10 and 15 minutes to complete. Each participant was compensated with USD $2.00. Table 7 in Appendix B shows the demographic information of our participants.

Procedure. To start, participants were presented with an Informed Consent Document (ICD). After agreeing to the ICD, participants were shown a *scenario* where they encountered a targeted clickbait. An example scenario for face-swap clickbait was: "Imagine you are a friend of the person shown below [a picture of person is shown]. While browsing through social media, you encounter the following post [a face-swapped post using the friend is shown]." We used scenarios in our study, as they are a powerful tool to help participants in imagining their interaction with targeted clickbait [26].

Each participant was shown five scenarios (one for each of the five targeting factors) and ten targeted clickbait (positive and negative for each factor). Participants were asked to rate the likelihood of clicking on each clickbait on a five-point Likert scale (1: Extremely unlikely, 5: Extremely likely). Afterwards, they provided rationale for their choice of the most effective targeted clickbait. A similar process was followed for the countermeasures, where participants rated their usefulness and provided their rationale. At the end, they responded to a demographic questionnaire.

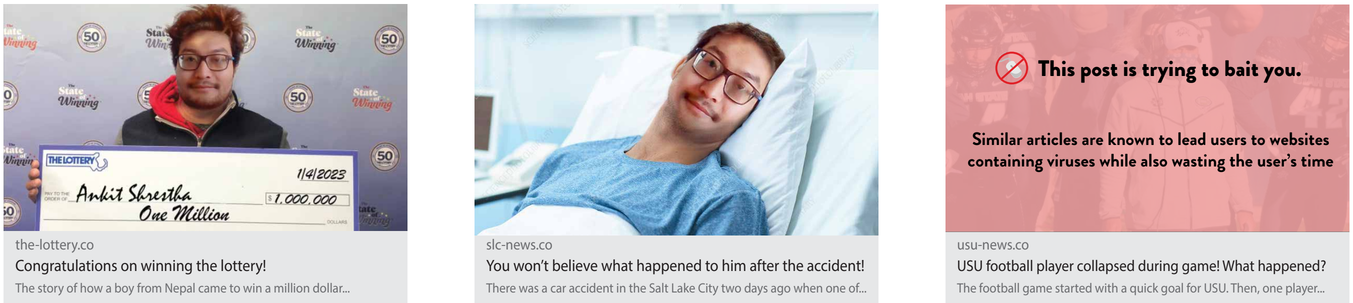
Analysis. We used statistical tests to analyze our quantitative results. We consider the result to be significant when we find p<.05. While comparing two conditions, we used a Wilcoxon signed rank test for the matched pairs of subjects. Wilcoxon tests are similar to t-tests but do not assume the distributions of the compared samples, which is appropriate for our collected data. For the qualitative results from the open-ended questions, we performed a thematic analysis, where two independent researchers coded the responses and later discussed and resolved the discrepancies in the codes. The inter-coder reliability was 91.6%.

3.2.2 Findings.

Emotional Valence. Our findings show that the likelihood to click on targeted clickbait is not significantly different (W=2052.0, p=.38) between the positively (M=3.9, SD=1.16) and negatively valenced posts (M=3.8, SD=1.31). We thus decided to use both variations in our next study (see §4).

Effectiveness of Targeted Clickbait. Our findings reveal that users were most likely to click the targeted clickbait using the face-swap of a friend, followed by the one using affiliated institution (see Table 1). Thus, we selected these for the next phase of our research. We found that users were significantly more likely to click on face-swapped posts than those using location or field of study/work (see Table 3). However, there were no significant differences between

[3] https://www.mturk.com/worker/help



the-lottery.co
Congratulations on winning the lottery!
The story of how a boy from Nepal came to win a million dollar...

slc-news.co
You won't believe what happened to him after the accident!
There was a car accident in the Salt Lake City two days ago when one of...

This post is trying to bait you.

Similar articles are known to lead users to websites containing viruses while also wasting the user's time

usu-news.co
USU football player collapsed during game! What happened?
The football game started with a quick goal for USU. Then, one player...

(a) Face-Swap with Positive Valence **(b) Face-Swap with Negative Valence** **(c) Text-Only Harm Warning**

Figure 2: Designs used in the Curation Survey. (a), (b) show targeted clickbait using face-swaps with friends, and (c) shows the text-only warning conveying harm.

Targeting Factors	Field of Study/Work	Relevant Location	Niche Activities
Face Swap (Friends)	**W=104.0, *p*<.05**	**W=61.5, *p*<.01**	W=99.0, *p*=.13
Affiliated Institutions	W=40.5, *p*=.14	W=45.5, *p*=.07	W=92.0, *p*=.61

Table 3: Significance test (Two most effective targeting factors vs. remaining)

Countermeasures	URL Credibility	Alternate Source	Reveal Mystery
Consequences	W=30.0, *p*=.45	**W=35.0, *p*<.01**	**W=30.0, *p*<.01**
Reporting	W=76.0, *p*=.41	**W=45.5, *p*<.05**	**W=10.0, *p*<.01**

Table 4: Significance test (Two most useful countermeasures vs. remaining)

face-swapped posts and the ones using affiliated institutions (W=58, *p*=.21) or niche activities (see Table 3).

From the open-ended responses, we see that about half of the participants believed that the personal nature of the face-swap post would pique their interest. One of them noted, *"Most are familiar clickbait formats that I would ignore regardless of how well tailored they are to my interests. Seeing a friend's face in clickbait is something I have never experienced before. If it were a real story, I would be fascinated. If it turned out to be fake, I would be very upset and add that to the list of dirty tricks I try not to fall for."* Similarly, some participants believed that the affiliated institution post would grab their attention, as one explained, *"As an avid Utah State University football fan, I would be very concerned if one of the athletes' health were threatened, and I would like further information about it."*

A summary of the taxonomy of targeting factors is provided in Table 8 in the Appendix.

Usefulness of Countermeasures. For the countermeasures, we found that conveying consequence was perceived as the most useful way to prevent users from clicking on a targeted clickbait, followed by reporting the post as misleading (see Table 2). Both of these methods were rated significantly better than the approaches of revealing the mystery and providing alternate sources (see Table 4). Therefore, we selected conveying consequences and reporting posts as misleading as the countermeasures for participatory design study.

Based on the open-ended responses, we observed that about half of the participants found conveying harm effective due to the fear appeal. One participant mentioned, *"Because I do not want to get a virus. I also do not want to support some scammy website that does such a thing. Finally, I do not want to waste my time on clickbait."* Some participants highlighted the effectiveness of reporting posts as misleading due to the social factor associated with it. One of them mentioned, *"This one is giving real-time crowd-sourced information that others have reported this [the post] as misinformation. That makes me think that I want to find another source to find out about this local news event."*

4 PARTICIPATORY DESIGN (RQ2)

We conducted the participatory design study with 24 participants (P1-P24) with the goal of generating story-based warnings for evaluation (see Figure 1b). Here, we used the curated countermeasures (see §3) to create design tasks in our participatory design study. Through these design tasks, participants created story-based warnings that we leveraged to generate concepts for designing the warning. We later address how we evaluate the warnings through an online survey in §5.

We created two variations of the design task, each representing a goal for story-based warning: conveying consequences *(Harm Design)* and reporting posts as misleading *(Report Design)* – the countermeasures selected from the curation survey. These design goals align with behavior change persuasion theories (see Appendix A for further details on theoretical framework used to develop our design tasks). In each design task, participants were given five categories of components: overlays, headers, navigation (buttons), messages, and components for expression (see Figure 3). Overlays, headers, messages, and navigation are based on prior studies that point to essential components in interstitial warnings [44, 72, 117, 158].

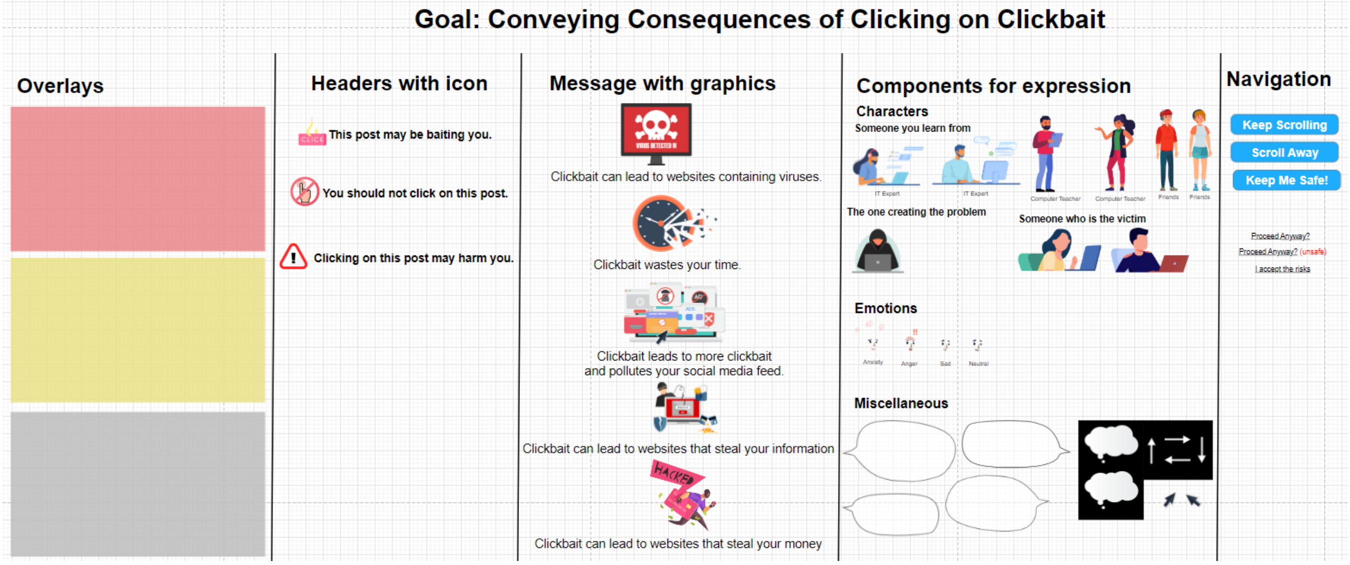

Goal: Conveying Consequences of Clicking on Clickbait
Overlays
Headers with icon
This post may be baiting you.
You should not click on this post.
Clicking on this post may harm you.
Message with graphics
Clickbait can lead to websites containing viruses.
Clickbait wastes your time.
Clickbait leads to more clickbait and pollutes your social media feed.
Clickbait can lead to websites that steal your information
HACKED
Clickbait can lead to websites that steal your money
Components for expression
Characters
Someone you learn from
IT Expert
IT Expert
Computer Teacher
Computer Teacher
Friends
Friends
The one creating the problem
Someone who is the victim
Emotions
Anxiety
Anger
Sad
Neutral
Miscellaneous
Navigation
Keep Scrolling
Scroll Away
Keep Me Safe!
Proceed Anyway?
Proceed Anyway? (unsafe)
I accept the risks

Figure 3: *Harm Design* (conveying harm) with all the design components

Components for expression are to facilitate storytelling in the warning design.

4.1 Methods

We recruited participants via email and by sharing the study information with various university departments. Participants had to be at least 18 years old to participate in this study. We had 13 women and 11 men as participants, aged between 18 and 44. Table 9 in the Appendix shows the demographic information of our participants.

We conducted the study over Zoom (audio-recorded). When a participant showed interest, we emailed them the Informed Consent Document (ICD), which they agreed to before we scheduled a time for a Zoom session. The Institutional Review Board at our university approved the study.

Procedure. First, participants were given an overview of the study. Then they were randomly assigned one of four targeted clickbait designs (2 emotional valences × 2 targeting factors), for which they reported their perceptions and whether they would click on it and why or why not. Next, participants were randomly assigned one of the two design tasks, where they first selected an overlay and a header, followed by providing the rationale behind their selection. Then, participants selected the message they would want to convey. They were asked to express their message through a story using components for expression. They were allowed to ask for additional components for expression (e.g., one participant asked for an elderly character). Then the participants provided their rationale behind the selection of message. Next, they were asked to select the navigation buttons and explain their selection.

Once the design was completed, participants explained their perceptions of why the story depicted in their design would be effective to prevent users from clicking on a targeted clickbait. Thereafter, participants were shown the remaining three targeted clickbaits one by one, and asked to elicit what changes they would make in their warning design considering the variations in targeted clickbaits; they also explained why the changes, if any, were necessary. They were also asked to select the most effective targeted clickbait and explain their choice. Finally, the participants were asked to complete a demographic survey hosted in Qualtrics[4]. They were compensated with a $15 Amazon.com gift card for their participation.

Analysis. The audio recordings from the study were transcribed and combined with the warnings designed by our participants. We performed thematic analysis on our transcriptions and the stories created by our participants [15, 22, 23, 131]. Two independent researchers coded each transcript and the story in the warning. The researchers read through the transcripts and stories of the first few interviews, developed codes, compared them, and then iterated until we had developed a consistent codebook. After the codebook was finalized, two researchers independently coded the remaining interviews. 88.9% of the codes matched between the two reviewers, resulting in Cohen's Kappa score of 0.83. The two coders discussed and agreed on the discrepancies in the remaining codes. Finally, we organized and taxonomized our codes into higher-level categories.

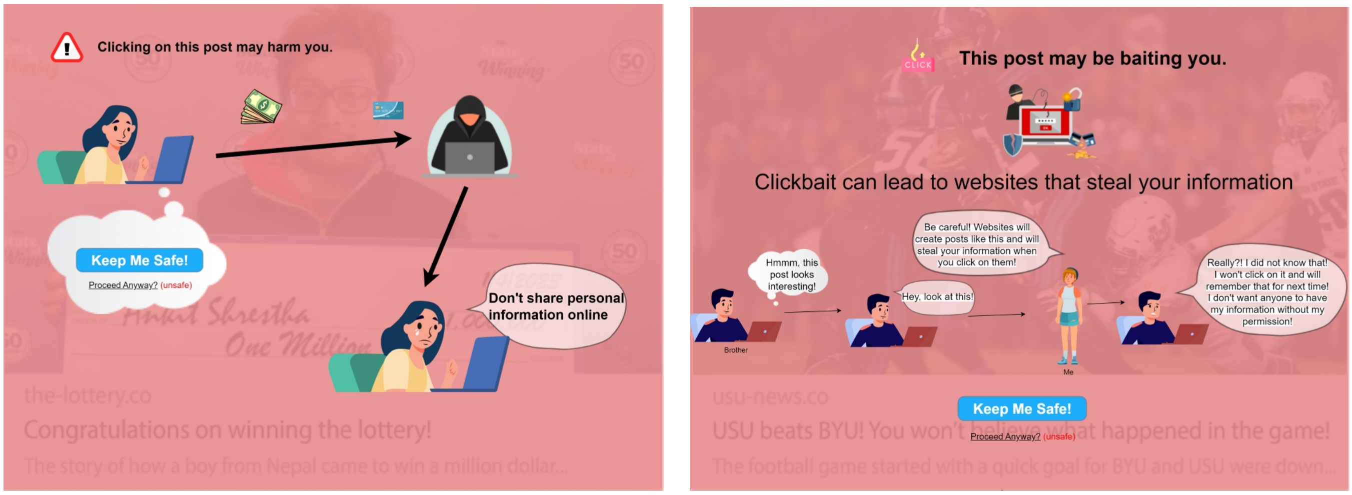
4.2 Findings

4.2.1 *Perceptions of Targeted Clickbait.*

The participants reported their perceptions and rationale behind clicking or avoiding the post (targeted clickbait), where our analysis revealed four prominent themes as presented below.

Relevance. About half of the participants agreed they would click on the post just because it was relevant. Some of them mentioned they would be surprised to see their friends in a social media post, and click any such post about someone they recognize. One of them said, *"Yeah, I probably would [click on the targeted clickbait], especially if I knew the person I would want to see what happened."*

[4] Qualtrics is an online survey platform used to create, distribute, collect, and analyze survey data (www.qualtrics.com).


Clicking on this post may harm you.
Keep Me Safe!
Proceed Anyway? (unsafe)
Don't share personal information online
the-lottery.co
Congratulations on winning the lottery!
The story of how a boy from Nepal came to win a million dollar...
This post may be baiting you.
CLICK
Clickbait can lead to websites that steal your information
Hmmm, this post looks interesting!
Be careful! Websites will create posts like this and will steal your information when you click on them!
Hey, look at this!
Really?! I did not know that! I won't click on it and will remember that for next time! I don't want anyone to have my information without my permission!
Brother
Me
usu-news.co
Keep Me Safe!
USU beats BYU! You won't believe what happened in the game!
Proceed Anyway? (unsafe)
The football game started with a quick goal for BYU and USU were down...

(a) Emotional Story (Harm) created by P1

(b) Competent Peer Story (Harm) created by P5

Figure 4: Story-based warnings conveying harm created by users

(P3). Some participants reported they would click on posts about their university as they would be enticed to know. These responses indicate that the relevancy of targeted clickbait would influence a user's decision to click on it. Participants also mentioned they would click on the post despite a warning identifying it as a clickbait, where one of them said, *"It [posts about university football] is something I usually talk about with friends and family. So I usually click on it even if it is clickbait so I can talk to them about it and know what's happening." (P5).*

Curiosity. Most participants reported that curiosity about the post was enhanced by its relevance. They further agreed that curiosity depended on the headline and the photo presented in the post. Some participants particularly pointed to words like "collapsed" and "you won't believe" in highlighting the role of headlines to grab their attention. One participant said, *"Yeah [I am interested in the post]. It is probably the use of words like collapsed for the player's situation, and then, when it says what happened, I expect to learn what actually happened if I click on it." (P15).*

Suspicion. Some participants raised suspicion when presented with the post due to their prior negative experiences with similar posts on social media. A few of them also identified the post as clickbait. One participant said, *"No [I will not click on it], because the first time I got this kind of post, and I clicked on it, it was fake, so I am afraid to try it." (P12).* A few participants felt the post was sketchy due to attention-grabbing headlines or fake-looking thumbnails, one of them commented, *"Yeah, it definitely catches my eye. Just because any time I see you won't believe what happened, I know it is clickbait. ... I probably wouldn't click on it" (P13).* The responses show that clickbait features like the attention-grabbing headline can be an identifying factor that help users to detect and avoid a clickbait. A few participants also reflected on their experience with posts using image manipulation in social media, similar to a targeted clickbait. One participant said, *"Lately there's tons of post on social media that's like someone you know died. And then all my friends are tagged in it. And I'm like, obviously that's fake." (P7)*

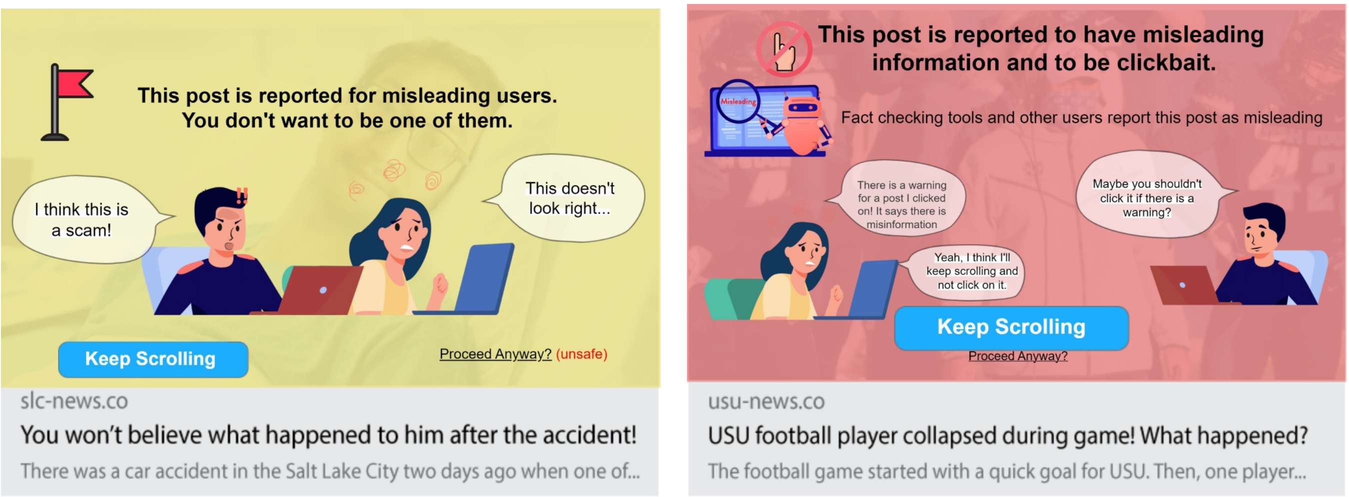
Habituation. Some participants reported their inclination to click on the post despite facing similar posts and identifying it as clickbait due to their non-consequential past experiences. One of them commented, *"Sometimes I do realize, like, okay, this is probably just like some sort of clickbait like trying to get me to click. However, if that is a topic that I am interested in, then I usually will click on it as I don't think it is that bad." (P3).* Our results indicate that participants are unaware of the hidden consequences of clickbait, like the common use of clickbait by malicious sites, and collecting information through cookies. While some participants are aware of consequences such as lots of ads and time wasting, they rarely think these consequences are harmful. These issues habituate users to interact with clickbait.

4.2.2 *Perceptions of Warning Components.*

In this section, we report our participants' perceptions of the warning components.

Overlays. For *Harm Design*, about half of our participants selected the red overlay, citing three advantages: attention, immediate conveyance of danger, and efficacy in warning users. One participant said, *"Red is just a big warning color. ... I mean, since we learn that red means stop, red is the color that makes you pause and, like, have caution. So if you're going to have an eye-catching warning, then it should be red." (P13).* Some participants selected yellow. Only a few selected gray as they did not consider targeted clickbait to be malicious. For *Report Design* as well, about half of our participants selected the red overlay and about half selected the gray overlay. Those selecting the gray overlay argued that red is too strong of a color to convey that a post is misleading.

Headers. In *Harm Design*, about half of our participants selected the goal-oriented header that reads "Clicking on this post may harm you," as it conveyed the risk of clicking a post. One participant commented that the goal-oriented one was better than the other two: *"I probably wouldn't even think about [selecting the first one]. [As for the second one that says] You should not click on this post. Yeah, I probably shouldn't, but I probably would still do it. But having this may harm you would get my attention more." (P4).* About half


This post is reported for misleading users.
You don't want to be one of them.
I think this is a scam!
This doesn't look right...
Keep Scrolling
Proceed Anyway? (unsafe)
slc-news.co
You won't believe what happened to him after the accident!
There was a car accident in the Salt Lake City two days ago when one of...
This post is reported to have misleading information and to be clickbait.
Misleading
Fact checking tools and other users report this post as misleading
There is a warning for a post I clicked on! It says there is misinformation
Maybe you shouldn't click it if there is a warning?
Yeah, I think I'll keep scrolling and not click on it.
Keep Scrolling
Proceed Anyway?
usu-news.co
USU football player collapsed during game! What happened?
The football game started with a quick goal for USU. Then, one player...

(a) Realistic Conversation (Report) created by P3

(b) Competent Peer with Credible Source (Report) created by P8

Figure 5: Story-based warnings conveying post is reported as misleading created by users

of the participants selected the header about baiting them, since it explicitly denoted the post as a clickbait. Only a few participants selected the header about not clicking the post, as they did not like being told what to do. For *Report Design*, about half of our participants selected the goal-oriented header, as they liked being informed about the specific reasons why they should avoid the post. One participant said, *"This one gives a little bit more context than the other headings. Instead of just saying this is clickbait or don't click on it, it says, well, this post is reported as misleading. So, I like that one a little bit more." (P3).*

Messages. In *Harm Design*, about half of our participants preferred to convey the stealing of information due to relevancy, where cookies from sketchy sites often collect user information. Some participants shared their experiences, such as when they clicked on a link and started receiving promotional emails. Some participants related the message to losing their sensitive information as a result of hacking. Only a few participants selected any of the other messages conveying the harm of targeted clickbait. For *Report Design*, about half of the participants preferred fact-checking tools as the source of report, where they believe there would be no bias from artificial intelligence. One participant said, *"[I like] This bottom one: fact-checking tools report this post as misleading. That seems more reliable than the other ones." (P3).* Only a few participants preferred any of the remaining messages.

Navigation. In *Harm Design*, about half of the participants selected *"Keep Scrolling"* button, as they found the suggestive language appropriate. Most participants selected *"Proceed Anyway? (unsafe)"* due to the button acting as a second reminder to the users. One participant commented, *"[I like proceed anyway (unsafe) because] They have the unsafe in red just to give this person an extra opportunity not to click on it." (P1).* In *Report Design*, most participants selected *"Keep Me Safe!"* button as they felt that clicking on it could protect them from misleading posts. One participant mentioned, *"I expect, keep you safe to work in a way that if someone clicks on keep me safe, then this content would be hidden in future." (P20).* About half of the participants preferred including the *"Proceed Anyway (unsafe)"* button, again due to it acting as a second reminder. Here, about half of the participants preferred the *"I accept the risks"* button as it would remind users that the link is risky and put liability on their action.

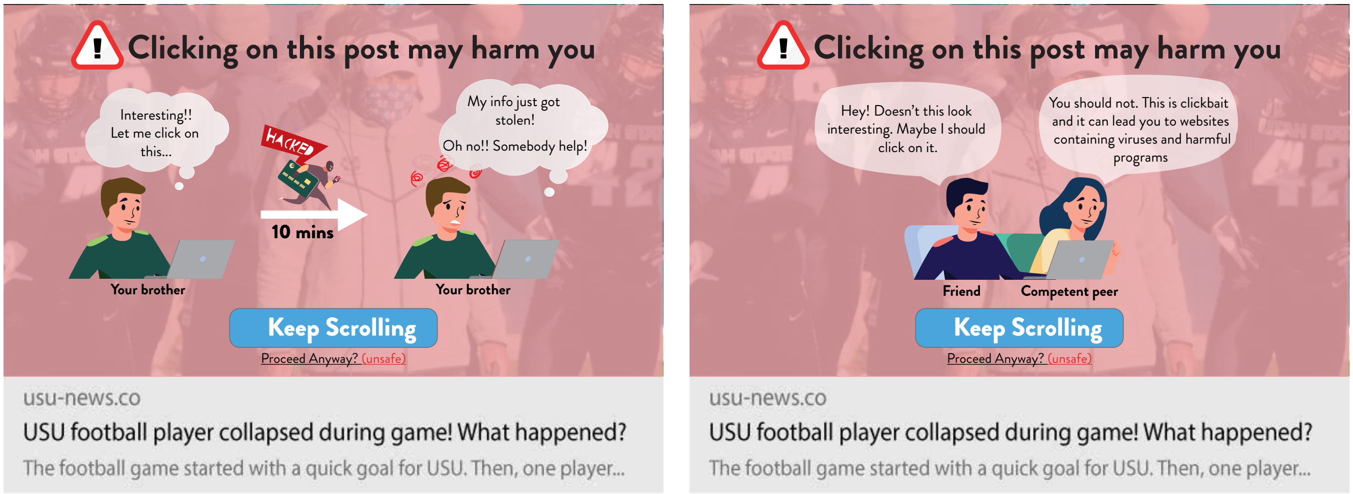
4.2.3 Themes in User Stories.

In this section, we report on the stories from our participants, which are categorized into four themes.

Harm: Emotional Story. About half of the participants who did *Harm Design* wanted to convey an emotional story, where a character faced consequences from targeted clickbait. Almost all of these stories were motivated by negative past experiences of participants or someone they know. One participant (see Figure 4a) mentioned, *"Well, I actually did have a friend before provide information, and she lost several thousand dollars for making that mistake. Okay, I think what I would like to do is have one of her where she's just like neutral. And then another picture of her like down here, of her very sad after what had happened." (P1).* Upon further decomposition of such stories, we identified three key elements in the design: a character who clicked on clickbait, a consequence that the character faced, and a negative emotion depicted on the character's face.

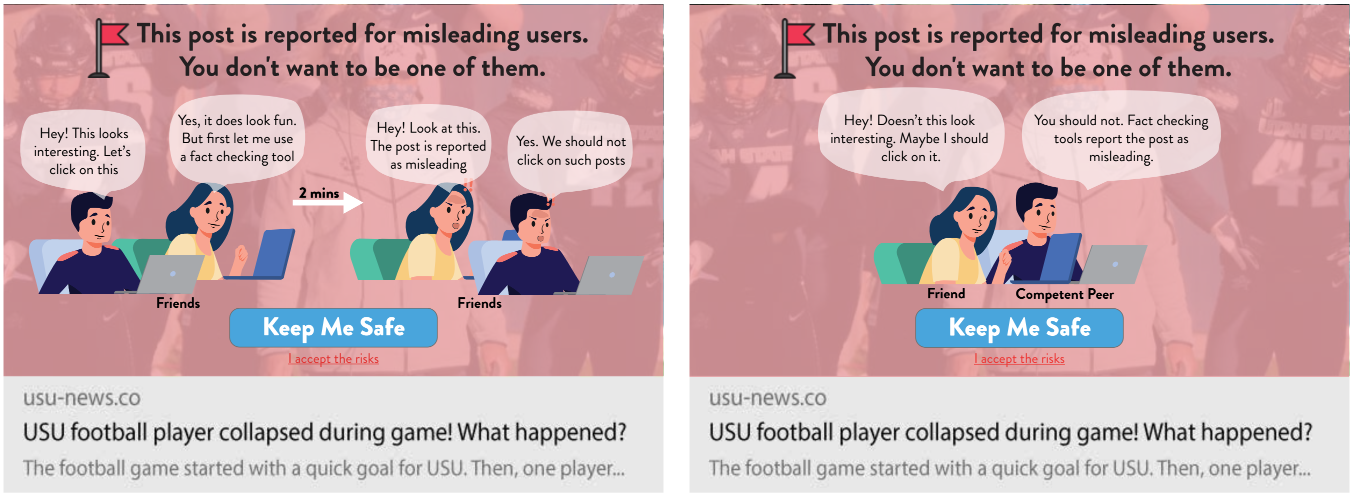
Harm: Competent Peer. About half of the participants depicted a story where a character interested in the targeted clickbait was stopped by another character, conveying its consequences. One participant (see Figure 4b) said, *"So my brother came across a post on social media with a picture and a headline that said, you won't believe what celebrity was arrested for this crazy crime. ... I warned him to let him know that there are lots of websites that create posts like this, and then they will steal information from your account if you click on it." (P5).* Upon further analysis, we found three key elements common across almost all such stories: a character interested in the post, a knowledgeable character, and a consequence explained by the knowledgeable character.

Report: Realistic Conversation. About half of the participants depicted a conversation between two characters in their story. Almost


Clicking on this post may harm you
Interesting!! Let me click on this...
HACKED
10 mins
My info just got stolen! Oh no!! Somebody help!
Your brother
Your brother
Keep Scrolling
Proceed Anyway? (unsafe)
usu-news.co
USU football player collapsed during game! What happened?
The football game started with a quick goal for USU. Then, one player...
Clicking on this post may harm you
Hey! Doesn't this look interesting. Maybe I should click on it.
You should not. This is clickbait and it can lead you to websites containing viruses and harmful programs
Friend
Competent peer
Keep Scrolling
Proceed Anyway? (unsafe)
usu-news.co
USU football player collapsed during game! What happened?
The football game started with a quick goal for USU. Then, one player...

(a) An emotional story of harm **(b) Story with a competent peer conveying harm**

Figure 6: Story-based warnings conveying harm, as used for evaluation


This post is reported for misleading users. You don't want to be one of them.
Hey! This looks interesting. Let's click on this
Yes, it does look fun. But first let me use a fact checking tool
2 mins
Hey! Look at this. The post is reported as misleading
Yes. We should not click on such posts
Friends
Friends
Keep Me Safe
I accept the risks
usu-news.co
USU football player collapsed during game! What happened?
The football game started with a quick goal for USU. Then, one player...
This post is reported for misleading users. You don't want to be one of them.
Hey! Doesn't this look interesting. Maybe I should click on it.
You should not. Fact checking tools report the post as misleading.
Friend
Competent Peer
Keep Me Safe
I accept the risks
usu-news.co
USU football player collapsed during game! What happened?
The football game started with a quick goal for USU. Then, one player...

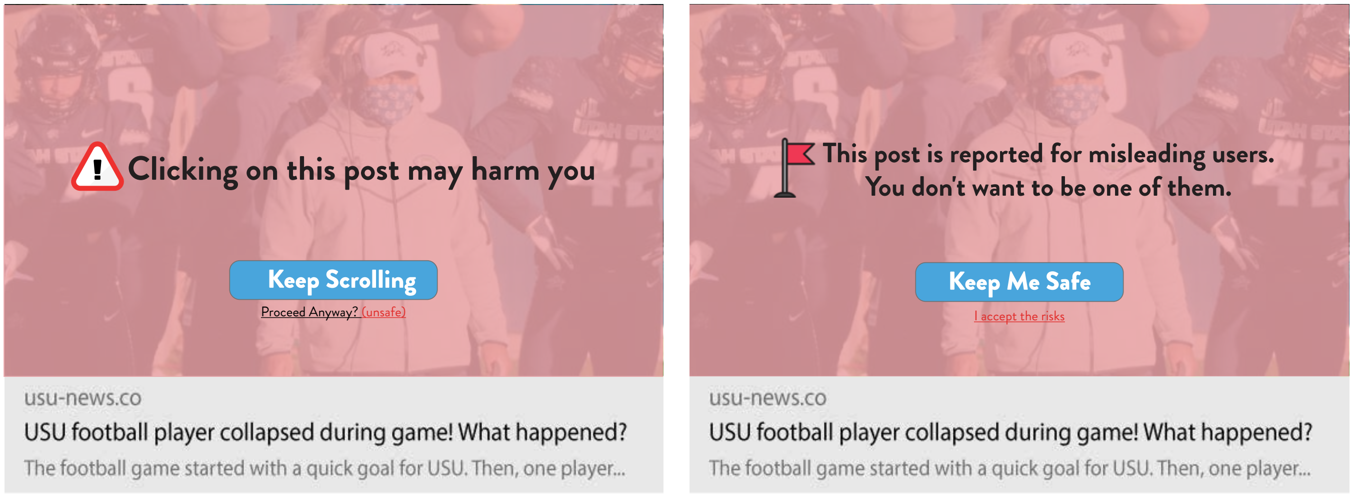
(a) A realistic conversation as a story **(b) Story with a competent peer reporting from a credible source**

Figure 7: Story-based warnings conveying reported posts, as used for evaluation

all of them reported that they usually work or study with a friend, and when they encounter something like this, they would figure out together what to do. One participant (see Figure 5a) said, *"I think it's pretty common for people to be either in a work setting or working on homework or just friends getting together and scrolling on the Internet where one person is looking like, oh, this seems weird. And the other person reassures like, yes, that is weird. It looks like a scam." (P3).* Our decomposition of these stories revealed two key elements: two characters talking with each other, and the source from which they discover that the post is misleading.

Report: Competent Peer with Credible Source. About half of the participants created a story where a character interested in the post was informed about a credible source reporting the post as misleading by another character. One participant (see Figure 5b) said, *"[The story would go like] Oh, shoot! I just clicked on this thing, and this warning came up. I don't know what to do. Another person would say, Whoa! What did you click? And they'd be like, this is reported as misleading from fact-checking tools." (P8).* These stories contained three key components: a character interested in the post, a knowledgeable character, and a credible source of report.

4.2.4 Warning Changes for Targeted Clickbait Variations. Most participants agreed that only a change in emotion did not change the targeted clickbait enough to warrant a change in the warning. Therefore, we would not use emotional valence as a variable in the evaluation survey. About half of the participants agreed that a warning might need to change with the change in targeting factor. We note that these changes are related to increasing the threat level of warnings. One participant mentioned, *"I would probably put more of a stronger warning on this one [lottery with face swap]. Just because I feel like anything with money is just very scammy."*



Clicking on this post may harm you

usu-news.co

USU football player collapsed during game! What happened?

The football game started with a quick goal for USU. Then, one player...

This post is reported for misleading users. You don't want to be one of them.

usu-news.co

USU football player collapsed during game! What happened?

The football game started with a quick goal for USU. Then, one player...

(a) Control condition for warnings conveying harm **(b) Control condition for warnings conveying post is reported**

Figure 8: Control condition warnings with the same components except the story

(P2). While these changes were motivated by the perceived risks regarding a post, any targeted clickbait could be equally harmful. We suggest the conveyance of appropriate threat levels through color changes in overlays, based on classifications suggested by artificial intelligence (see §6.5). Here, our designs are not customized based on the targeted clickbait.

4.3 Translation to Final Designs

Since the participatory design study is qualitative [15, 22, 23], it is not reasonable to select the most used components, especially where the difference is small (e.g., five participants select red overlays, while six participants selected yellow overlays). Therefore, we conducted a focus group discussion (FGD) with four participants (PD1-PD4), including UX experts, graphic designers, and psychology majors. The participants were recruited via email to the respective departments of our university. The study was conducted over Zoom and lasted around an hour. The Institutional Review Board at our university approved the study.

4.3.1 Methods. At the beginning of the FGD, participants were provided with the qualitative and quantitative measures from our findings in the participatory design study. Then they discussed the findings that guided their final selection of components. Our participants first selected the overlays, headers, and navigation buttons. The expression of the message through storytelling varied among our participants in the participatory design study, and we used thematic analysis to identify themes and key components within the stories (see §4.2.3). During the FGD, participants discussed and selected combinations of components, resulting in the stories we finally used. Here, four themes were translated into four story-based warnings, which we evaluated in the next study (see §5). At the end of the FGD, participants completed a demographic survey and were compensated with a $15 Amazon.com gift card.

4.3.2 Story-based Warnings for Evaluation.

Harm: Emotional Story (shortened to Harm: Emotional). For the story elements, participants had the two characters be siblings. One participant expressed, *“Siblings are the characters that users would endear and care about in the warnings”. (PD3).* Similarly, stealing information was selected as the consequence, and anxiety as the negative emotion (see Figure 6a). The participants selected the red overlay, the goal-oriented header, and the ”Keep Scrolling” and ”Proceed Anyway? (unsafe)” buttons for the remaining components. Our participants agreed with the responses from participatory design in selecting these components.

Harm: Competent Peer (shortened to Harm: Peer). Our participants selected the same warning components as for the emotional story for similar reasons. For the story elements, participants had the knowledgeable character and the character interested in the post be friends. One of them mentioned, *“I can imagine two friends talking with each other where one is competent and informs the other”. (PD1).* The participants selected infection of devices through viruses as the consequence since it seemed realistic and convincing to them in a conversation between two friends (see Figure 6b).

Report: Realistic Conversation (shortened as Report: Conversation). The participants discussed and agreed upon two friends as the characters having a conversation and fact-checking tools as the source from which they discover that the post is misleading (see Figure 7a). One participant mentioned, *“Realistically speaking, I think the conversation between two friends is the only option I can think happening in reality.” (PD1).* For warning components, participants selected the red overlay and goal-oriented headers, agreeing with the responses from our participatory design. Participants selected the ”Keep Me Safe!” button, as it felt like a safeguard against misleading posts, and the ”I accept the risks” button, as it conveyed that users would be liable for consequences.

Report: Competent Peer with Credible Source (shortened as Report: Source). Our participants selected the same warning components as with the realistic conversation. For the story, participants selected

Demographic	Demographic Group	*N*
Gender	Male	44
	Female	66
	Prefer not to answer	1
Age range	25-29 years old	2
	30-34 years old	9
	35-39 years old	16
	40-44 years old	29
	45-49 years old	13
	50-54 years old	10
	55-59 years old	13
	60-64 years old	12
	Above 65 years old	7
Race	White	94
	Black/African American	7
	Asian	4
	Hispanic or Latino	2
	Native American	1
	Mixed Race	3
Education	High School Graduate	30
	Two-year College Degree	27
	Four-year College Degree	45
	Graduate degree (MS/PhD)	8
	I prefer not to answer	1

Table 5: Demographic Information of the Participants in the Evaluation Survey (*N*=Number of Participants)

friends as both the knowledgeable character and the character interested in the post. They agreed that fact-checking tools are the most credible source (see Figure 7b).

5 EVALUATION (RQ3)

Along with the four story-based warnings (see §4.3.2), we created two control conditions (one each for harm and report) that include all but the story in their design (see Figure 8). The comparison between controls and story-based warnings thus shows the impact of stories in warning design. We evaluated six warning variations through a Qualtrics survey with 114 participants (medium effect size based on power analysis) over MTurk (see Figure 1c). The Institutional Review Board at our University approved the study.

5.1 Methods

Participant Recruitment. Participants had to be 18 years or older and live in the United States or Canada to participate in our study. We followed the guidelines from prior work [82, 88, 110] to increase the quality of responses, where we recruited participants with a 99% HIT approval rate,[5] and used masters qualification considering the nature and length of the survey. We compensated the participants with USD 2.5. The study took between 12 and 25 minutes to complete. Table 5 shows the demographics of our participants.

Procedure. At the beginning, participants were presented with an Informed Consent Document (ICD). After agreeing to the ICD, they were shown a *scenario* [26] in which they encountered a targeted clickbait (without any warning). An example scenario for the affiliated institution case is: ”Imagine you are a student at Utah State University. While browsing through social media, you encounter the following post [a post about the University is shown].” Participants rated their *Interest* and *Likelihood* to click on the targeted clickbait (without warning) on a 7-point Likert scale (-3: strongly disagree, 3: strongly agree).

Thereafter, each participant was shown the six warning variations, in random order to avoid order effects. After each warning was shown, participants rated it in seven different survey questions on a 7-point Likert scale. The questions were presented in random order, with some questions reversed using antonyms to

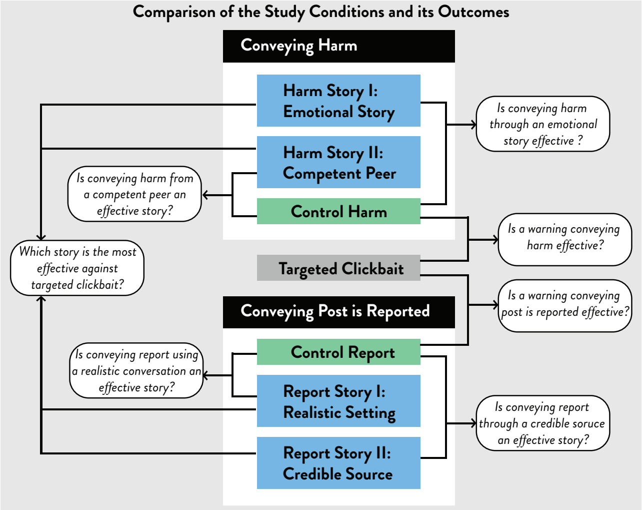
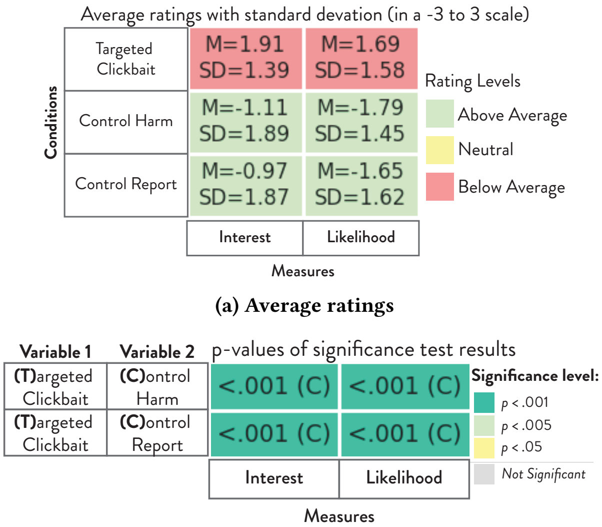

Comparison of the Study Conditions and its Outcomes
Conveying Harm
Harm Story I: Emotional Story
Harm Story II: Competent Peer
Control Harm
Targeted Clickbait
Conveying Post is Reported
Control Report
Report Story I: Realistic Setting
Report Story II: Credible Source
Is conveying harm through an emotional story effective ?
Is conveying harm from a competent peer an effective story?
Is a warning conveying harm effective?
Which story is the most effective against targeted clickbait?
Is a warning conveying post is reported effective?
Is conveying report using a realistic conversation an effective story?
Is conveying report through a credible soruce an effective story?

Figure 9: Comparison of the study conditions and its outcomes



Conditions	Interest	Likelihood

Targeted Clickbait	M=1.91
	SD=1.39
		M=1.69
		SD=1.58

Control Harm	M=-1.11
	SD=1.89
		M=-1.79
		SD=1.45

Control Report	M=-0.97
	SD=1.87
		M=-1.65
		SD=1.62

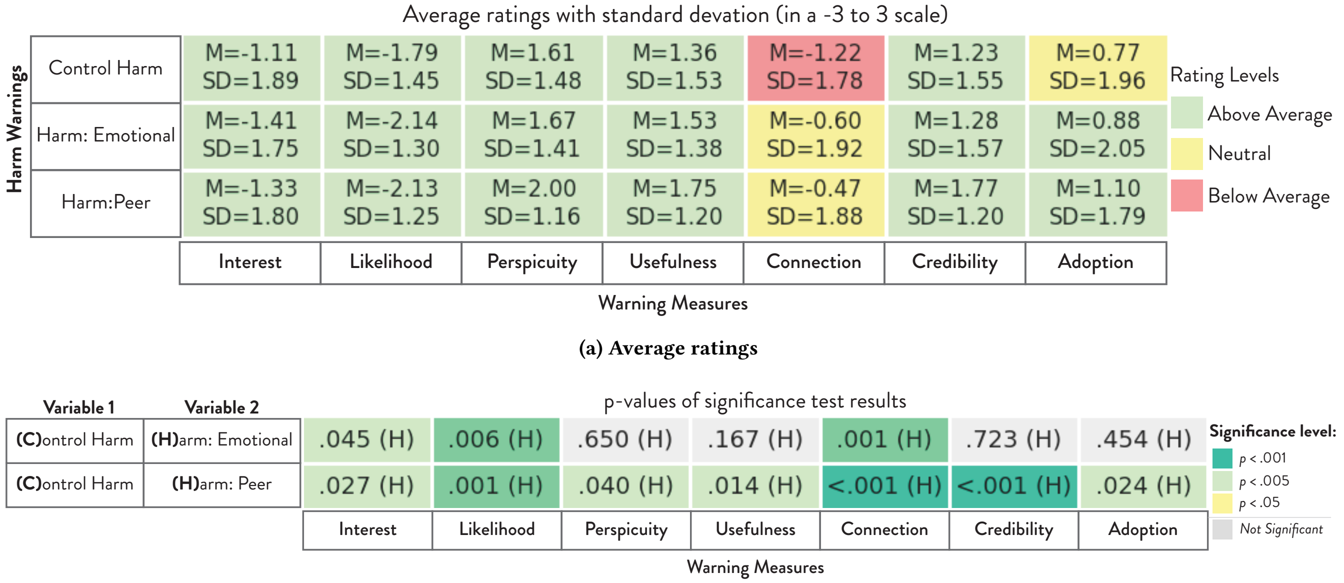
(a) Average ratings

Variable 1	Variable 2	Interest	Likelihood
(T)argeted Clickbait	(C)ontrol Harm	<.001 (C)	<.001 (C)
(T)argeted Clickbait	(C)ontrol Report	<.001 (C)	<.001 (C)

(b) Significance test results

Figure 10: Comparisons between the targeted clickbait post (without warning) and control warnings

[5] https://www.mturk.com/worker/help



Average ratings with standard deviation (in a -3 to 3 scale)

Harm Warnings	Interest	Likelihood	Perspicuity	Usefulness	Connection	Credibility	Adoption
Control Harm	M=-1.11 SD=1.89	M=-1.79 SD=1.45	M=1.61 SD=1.48	M=1.36 SD=1.53	M=-1.22 SD=1.78	M=1.23 SD=1.55	M=0.77 SD=1.96
Harm: Emotional	M=-1.41 SD=1.75	M=-2.14 SD=1.30	M=1.67 SD=1.41	M=1.53 SD=1.38	M=-0.60 SD=1.92	M=1.28 SD=1.57	M=0.88 SD=2.05
Harm:Peer	M=-1.33 SD=1.80	M=-2.13 SD=1.25	M=2.00 SD=1.16	M=1.75 SD=1.20	M=-0.47 SD=1.88	M=1.77 SD=1.20	M=1.10 SD=1.79

Warning Measures

Rating Levels
- Above Average
- Neutral
- Below Average

(a) Average ratings

p-values of significance test results

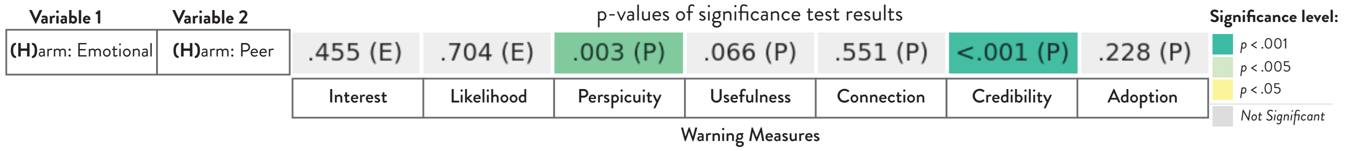
Variable 1	Variable 2	Interest	Likelihood	Perspicuity	Usefulness	Connection	Credibility	Adoption
(C)ontrol Harm	(H)arm: Emotional	.045 (H)	.006 (H)	.650 (H)	.167 (H)	.001 (H)	.723 (H)	.454 (H)
(C)ontrol Harm	(H)arm: Peer	.027 (H)	.001 (H)	.040 (H)	.014 (H)	<.001 (H)	<.001 (H)	.024 (H)

Warning Measures

Significance level:
- $p < .001$
- $p < .005$
- $p < .05$
- *Not Significant*

(b) Significance test results

Figure 11: Comparisons for harm warnings: control vs. story-based



p-values of significance test results

Variable 1	Variable 2	Interest	Likelihood	Perspicuity	Usefulness	Connection	Credibility	Adoption
(H)arm: Emotional	(H)arm: Peer	.455 (E)	.704 (E)	.003 (P)	.066 (P)	.551 (P)	<.001 (P)	.228 (P)

Warning Measures

Significance level:
- $p < .001$
- $p < .005$
- $p < .05$
- *Not Significant*

Figure 12: Significance test results comparing the two story-based harm warnings

avoid bias [36, 150]. The questions asked participants to evaluate the warning based on its *Perspicuity*, and *Usefulness* [123] using UEQ+, a validated scale of user experience [124]. We also added custom questions – similar to prior studies [39, 161, 164] – where we asked participants about their *Interest* and *Likelihood* to click on the targeted clickbait with the warning; they also rated the warning in terms of personal *Connection*, *Credibility*, and *Adoption*.

We included six attention checks in this survey [69, 83]. Three participants failed at least one attention check question, and their responses were removed from the analysis. Participants were also asked two open-ended questions about their feedback on each warning and their rationale behind adopting (or not adopting) it. At the end, participants answered a set of demographic questions.

Analysis. We used statistical tests to analyze our quantitative results. We consider results to be significant when we find p<.05 using a Wilcoxon signed rank test. Figure 9 highlights the comparisons of our study conditions and corresponding outcomes. We performed thematic analysis for the qualitative results from our open-ended questions, where two independent researchers coded the responses and later discussed and resolved the discrepancies in their codes. The coding included a total of 1332 responses (111 participants × 6 warning variations × 2 open-ended questions), where the inter-coder reliability was 87.3%.

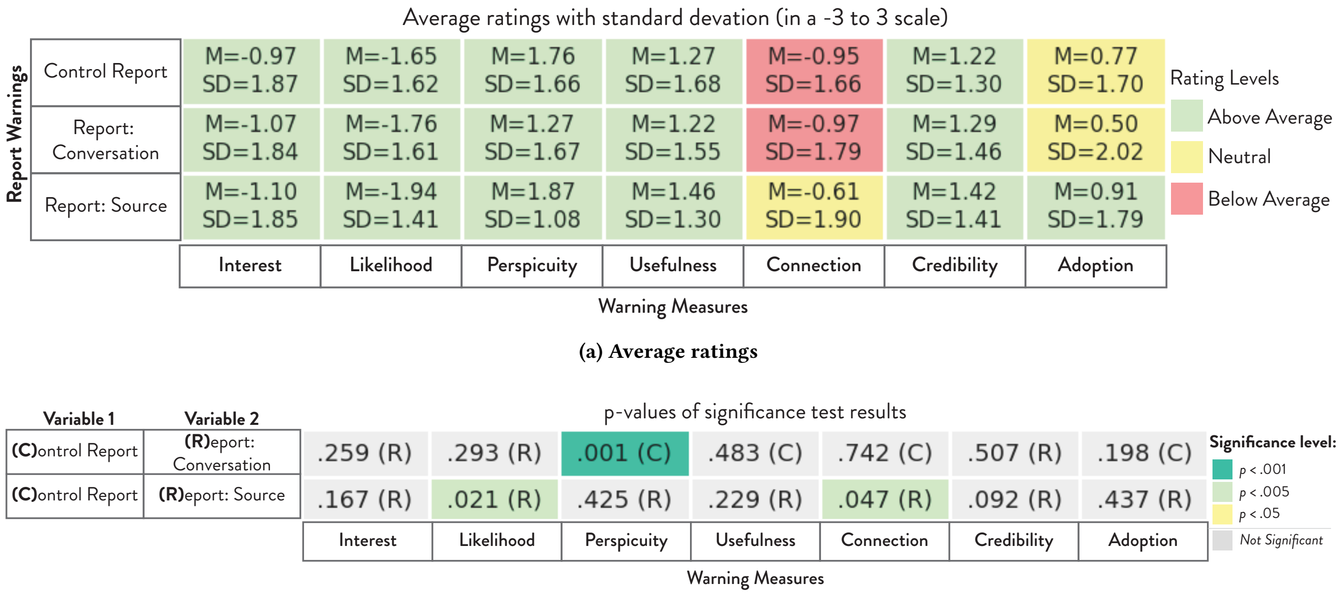
5.2 Findings

In this section, the reported means of the measures are on a −3 to 3 scale. Based on the UEQ handbook,[6] values between −0.8 and 0.8 represent a neutral evaluation, values > 0.8 represent a positive evaluation, and values < −0.8 represent a negative evaluation in a non-benchmarked scale. Since *Interest* and *Likelihood* are undesirable, we reverse the colors for these measures in Figures 11a, 10a, and 13a. Further, UEQ points out that due to the calculation of means over a range of diverse participants and answer tendencies (for example the avoidance of extreme answer categories), values close to +2 or −2 are considered extremities. We will use these classifications for our measures except for *Perspicuity*, which is benchmarked and has defined levels for positive, negative, and neutral evaluations.

5.2.1 *Is Targeted Clickbait Effective?*

For the targeted clickbait, participants rated their *Interest* and *Likelihood* to click. We observed high scores for *Interest* **(M=1.91, SD=1.39)**, which implies that based on the scenario, participants found the post relevant. Our findings from the curation survey and participatory design highlight the importance of relevance in participants' decision to click

[6] https://www.ueq-online.org/Material/Handbook.pdf



Average ratings with standard deviation (in a -3 to 3 scale)

Report Warnings	Interest	Likelihood	Perspicuity	Usefulness	Connection	Credibility	Adoption
Control Report	🟩 M=-0.97 SD=1.87	🟩 M=-1.65 SD=1.62	🟩 M=1.76 SD=1.66	🟩 M=1.27 SD=1.68	🟥 M=-0.95 SD=1.66	🟩 M=1.22 SD=1.30	🟨 M=0.77 SD=1.70
Report: Conversation	🟩 M=-1.07 SD=1.84	🟩 M=-1.76 SD=1.61	🟩 M=1.27 SD=1.67	🟩 M=1.22 SD=1.55	🟥 M=-0.97 SD=1.79	🟩 M=1.29 SD=1.46	🟨 M=0.50 SD=2.02
Report: Source	🟩 M=-1.10 SD=1.85	🟩 M=-1.94 SD=1.41	🟩 M=1.87 SD=1.08	🟩 M=1.46 SD=1.30	🟨 M=-0.61 SD=1.90	🟩 M=1.42 SD=1.41	🟩 M=0.91 SD=1.79

Warning Measures

Rating Levels
🟩 Above Average
🟨 Neutral
🟥 Below Average

(a) Average ratings

p-values of significance test results

Variable 1	Variable 2	Interest	Likelihood	Perspicuity	Usefulness	Connection	Credibility	Adoption
(C)ontrol Report	**(R)**eport: Conversation	⬜ .259 (R)	⬜ .293 (R)	🟦 .001 (C)	⬜ .483 (C)	⬜ .742 (C)	⬜ .507 (R)	⬜ .198 (C)
(C)ontrol Report	**(R)**eport: Source	⬜ .167 (R)	🟩 .021 (R)	⬜ .425 (R)	⬜ .229 (R)	🟩 .047 (R)	⬜ .092 (R)	⬜ .437 (R)

Warning Measures

Significance level:
🟦 $p < .001$
🟩 $p < .005$
🟨 $p < .05$
⬜ *Not Significant*

(b) Significance test results

Figure 13: Comparisons for report warnings: control vs. story-based

on the post. Similar results are echoed from our evaluation survey, where the *Likelihood* to click on a post is quite high **(M=1.69, SD=1.58)**. Further, these two measures (Interest and Likelihood) are strongly correlated **(r=0.94, *p*<0.001)**. These findings are in line with the communicative principle of relevance [140, 141], showing that the relevance of targeted clickbait is a key factor in its efficacy.

5.2.2 *Do Control Warnings without Stories Work?*

In this section, we report on the efficacy of control warnings (i.e., warnings without stories). We observe that even without stories, warnings substantially decrease *Interest* in the targeted clickbait (see Figure 10a). As expected, given that *Interest* is an important aspect of users' decision-making process, we also see a considerable decline in the *Likelihood* measure.

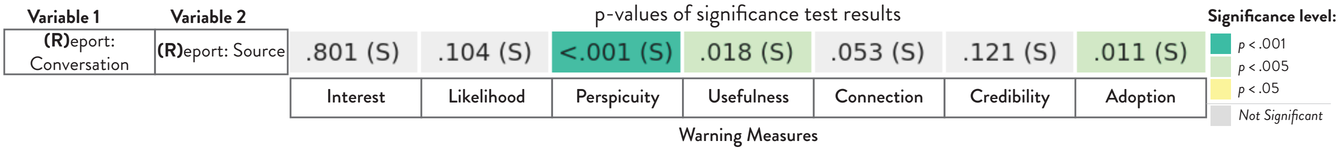
Significance tests between the targeted clickbait post (without warnings) and the control conditions revealed that both of our control warnings significantly reduced *Interest* and *Likelihood* (see Figure 10b). Open-ended responses support these results, where participants highlight the power of reflection due to the interruptions caused interstitial warnings. One participant noted, *"I probably would adopt the warning in real life because it does not stop me from clicking the article if I really wanted to. It makes me stop and think why this warning would be there and discourages me from reading an unsafe article"*. Even without stories or details in the warnings, some participants reflected on their actions and decided to find an alternate source to their answer. One of them mentioned, *"The warning offers to let me continue to the story, but it reiterates that it is unsafe. I'm going to keep scrolling, as the message suggests. I can always find out if the story is true by other means."* Further, some participants pointed to the other benefits of the warnings, including saving time and avoiding fake information.

5.2.3 *Do Stories about Harm Improve Warning Efficacy?*

We observe that the control warning for harm is effective. The control was rated above average in all of the measures except *Connection* and *Adoption* (see Figure 11a). Now we aim to understand if adding user-informed stories further improves the control warning. From the survey results, we observe that story-based warnings conveying harm were rated higher than the control in all measures (see Figure 11a). Notably, *Likelihood* for the story-based warnings is rated below -2.1, indicating their effectiveness.

In open-ended responses, about half of our participants highlighted the lack of specificity of harm in the control harm warning. Such specificity is provided in story-based warnings, which helps explain their higher ratings. One participant commented, *"The [control] warning doesn't do a good job of explaining exactly how the article may harm me. Are they going to attempt to phish me? Is it harmful to me in some kind of emotional sense (like it may cause me a negative reaction based on false information) or something? I don't really like this warning, and it isn't very clear, and I now have more questions than answers about what the site might be about."*. Thus, most participants reported that they might not adopt the control warning, supporting the below-average rating for its *Adoption*. One participant mentioned, *"I am not sure whether or not I would adopt this warning in real life, if I am being honest. I find myself slightly torn. The general idea is that the user should not click this link, as it may cause harm, but it's unclear exactly what type of harm. The more I think of it, the more I lean towards not adopting it."*

Significance tests pointed to the efficacy of the *"Harm: Emotional"* warning, which significantly reduced the *Interest* and *Likelihood* as compared to the control (see Figure 11b). About half of our participants appreciated the communication of harm through a story. One of them noted, *"I can tell that it is trying to protect me and prevent me*



p-values of significance test results

Variable 1	Variable 2	Interest	Likelihood	Perspicuity	Usefulness	Connection	Credibility	Adoption
(R)eport: Conversation	(R)eport: Source	.801 (S)	.104 (S)	<.001 (S)	.018 (S)	.053 (S)	.121 (S)	.011 (S)

Warning Measures

Significance level:

🟩	$p < .001$
🟩	$p < .005$
🟨	$p < .05$
⬜	*Not Significant*

Figure 14: Significance test results comparing the two story-based report warnings

from making a bad decision in clicking on the link. I appreciate that it is communicating what would happen and making me aware of the risk of this link." The *"Harm: Peer"* warning performed significantly better than the control in all the measures (see Figure 11b). Open-ended responses support the findings from our significant tests in *Perspicuity*, *Credibility*, *Connection*, and *Adoption*, where most participants reported positively about the warning's clarity, use of a competent peer, trust in their friends, and conveyance of specific harm, respectively. One of them noted, *"I would adopt the warning in real life because it is coming from a competent peer that I know. I would feel a personal connection to them, and I would heed their warning. I would want my friend to have a good opinion of me, and I would trust their judgment."*. Some participants also highlighted the relevance of a story presented in the warning. One of them mentioned, *"It's pretty easy to understand. It's like talking to your friend. It makes sense to me to speak to my friend in this matter so that I can protect her from harm."*

When comparing the two-story based warnings, we observe that *"Harm: Peer"* rates higher than *"Harm: Emotional"* in all measures except *Interest* and *Likelihood* (see Figure 11a). Significance tests between these two warnings revealed that *"Harm: Peer"* performed significantly better in terms of *Credibility* and *Perspicuity* (see Figure 12). Open-ended responses suggest that some participants found the message of harm coming from a friend more credible than the story depicting a single character facing harm. They also mentioned that conversations are easier to understand, which could explain the ratings for *Perspicuity*. One participant commented, *"I feel that the conversation gives a few more details that makes it easier for me to understand what could happen if I click on the link."* In light of the scores for *Interest* and *Likelihood*, however, we are unable to declare a champion story-based warning conveying harm. Most critically, our findings show that adding user-informed stories can enhance the efficacy of warnings conveying harm.

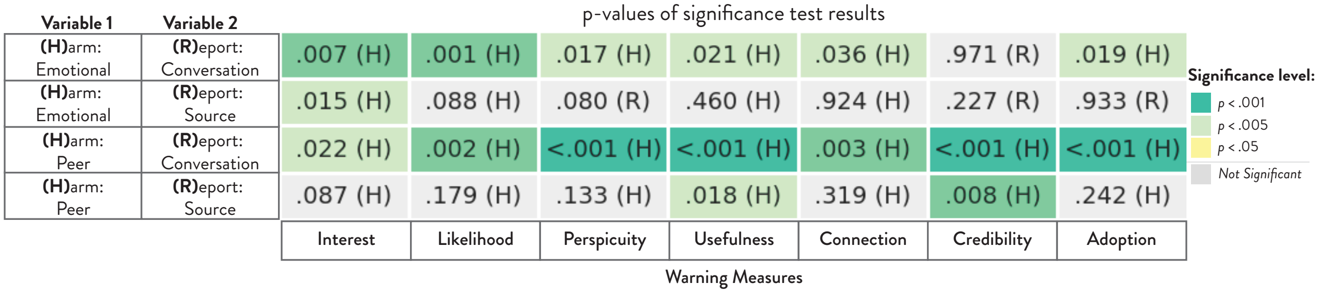
5.2.4 Do Stories about Reporting Improve Warning Efficacy? As with control warnings conveying harm, we observe that control warnings conveying report – i.e., the post is reported as misleading – are effective. The control was again rated above average in all measures except *Connection* and *Adoption*. Now we compare these with story-based warnings about reporting. First, we see that the *"Report: Source"* warning is rated higher than the control in all measures (see Figure 13a). As with the warnings conveying harm, about half of our participants conveyed the need for additional information in the control. One participant said about control, *"The warning is very basic, and doesn't give a lot of context. While the warning is clear, it doesn't do a good job of explaining why the post is misleading."*. The story element of the *"Report: Source"* warning addressed this need.

Significance tests revealed that the *"Report: Source"* warning significantly reduced the *Likelihood* to click compared to the control (see Figure 13b). In open-ended responses, about half of our participants appreciated providing the source of a report in the warning. One participant noted, *"It gives some information about peer reviews, so I like that it explains why there is a warning."* The *"Report: Source"* warning also performed significantly better than the control in terms of *Connection* (see Figure 13b). Similar to the harm warnings, some participants found the information coming from a friend personal and relatable. One of them mentioned, *"The situation seems more realistic than the others, and the cartoon is done in a more positive light in that the character is trying to help the other one and giving a good reason of why they shouldn't click on it."*.

The *"Report: Conversation"* warning, however, rates lower than the control in *Perspicuity*, *Usefulness*, *Connection*, and *Adoption* (see Figure 13a). Further, we find that the control was significantly better than the *"Report: Conversation"* warning in terms of *Perspicuity* (see Figure 13b). This warning attempts to depict via a comic-like presentation how two characters are conversing to discover the report. Even though open-ended responses for the control warning highlight the need for specific information and more details, participants found this presentation to be complicated and difficult to comprehend. One participant mentioned, *"It's a good way to warn people about viruses but it is also very complicated and not straightforward. I would rather use a different approach and one that doesn't have many steps to it."* According to some of our participants' comments, the difficulty in comprehension negatively impacted their perceptions of warning's *Usefulness* and their response for its *Adoption*.

When comparing the two story-based report warnings, we observe that *"Report: Source"* is rated higher than *"Report: Conversation"* in all measures (see Figure 13a). Significance tests between these two warnings revealed that *"Report: Source"* performed significantly better in terms of *Perspicuity*, *Usefulness*, and *Adoption* (see Figure 14). As described above (for *"Report: Conversation"*), these results can be explained through the difficulty of participants in understanding the concept of two characters discovering that a post is reported.

On the *Connection* measure, we perhaps surprisingly found that all six warnings performed poorly (see Figure 11a and 13a); the reason is unclear from the open-ended responses. We speculate that participants did not feel personally connected to warnings, as they would stop them from performing their primary task. Further studies are needed to have more in-depth understanding of participants' perceptions and ratings on personal connection. We note that story-based warnings performed significantly better than the control warnings for *Connection* (see Figures 11b and 13b).



p-values of significance test results

Variable 1	Variable 2	Interest	Likelihood	Perspicuity	Usefulness	Connection	Credibility	Adoption
(H)arm: Emotional	(R)eport: Conversation	.007 (H)	.001 (H)	.017 (H)	.021 (H)	.036 (H)	.971 (R)	.019 (H)
(H)arm: Emotional	(R)eport: Source	.015 (H)	.088 (H)	.080 (R)	.460 (H)	.924 (H)	.227 (R)	.933 (R)
(H)arm: Peer	(R)eport: Conversation	.022 (H)	.002 (H)	<.001 (H)	<.001 (H)	.003 (H)	<.001 (H)	<.001 (H)
(H)arm: Peer	(R)eport: Source	.087 (H)	.179 (H)	.133 (H)	.018 (H)	.319 (H)	.008 (H)	.242 (H)

Warning Measures

Significance level:
- $p < .001$
- $p < .005$
- $p < .05$
- Not Significant

Figure 15: Significance test results comparing the story-based warnings conveying harm and report

5.2.5 Is There a Champion Story? Finally, we compare all four story-based warnings (two each for harm and report). We observe that the *"Harm: Peer"* warning is rated the highest in all measures except *Interest* and *Likelihood* (see Figures 11a and 13a). In these two measures, *"Harm: Emotional"* is rated the highest. *"Report: Conversation"* warning is rated the lowest in all of the measures except *Credibility*, in which measure *"Harm: Emotional"* is rated the lowest (see Figure 11a and 13a).

In terms of significance tests, we find that both harm warnings significantly outperform *"Report: Conversation"* in nearly every measure (see Figure 15). *"Harm: Emotional"* and *"Report: Source"* are roughly even, with only on significant result for *"Harm: Emotional"* in the *Interest* measure. *"Harm: Peer"* is at least slightly better across the board versus *"Report: Source"*, but with only two significant differences: *Credibility* and *Usefulness*. Given that *"Harm: Peer"*, *"Harm: Emotional"*, and *"Report: Source"* are all effective with few significant differences, we cannot select a clear champion.

6 DISCUSSION AND IMPLICATIONS

Our studies report on the understanding and behavior of users towards targeted clickbait, the design of user-informed story-based warnings, and their effectiveness against targeted clickbait. In this section, we discuss the implications of our findings and provide guideline for future research in these directions.

6.1 Moving Towards a User-Informed Design Process

Prior studies [17, 70, 100, 101, 152] highlight the difficulties and challenges faced by end users that are often overlooked by designers. According to Norman [101], information that the designers want to convey through warnings may differ from the information perceived by end users. Therefore, our studies include end users, starting from the ideation of targeting factors and countermeasures and continuing throughout the design process.

During our user-informed activities, we faced challenges, particularly in the participatory design study. We had to conduct multiple pilot sessions with non-technical users and internal feedback sessions to align our design process, design activities, and interview questions with the understanding of the users. Based on our observations, users find it difficult to design artifacts without significant structure and guidelines and may struggle to express their ideas through designs. These observations led us to create a guided step-by-step design process, where the design is divided into multiple steps (e.g., select an overlay for the warning) and are guided by the researchers (e.g., think of the size for the overlay).

Our findings indicate the efficacy of this process, where the evaluation of our six user-informed warnings (including control warnings, where the components outside of stories were also user-informed) point towards their effectiveness against targeted clickbait (see §5.2.2, §5.2.3, and §5.2.4). Based on the positive reception of our designs, we encourage the HCI community to adopt more user-informed design processes and involve users from the early stages. We believe that this approach will benefit the community in aligning designs to users' needs and expectations.

6.2 A Challenge: The High Relevance of Targeted Clickbait

Scott [127] explains the working of clickbait through the *cognitive* principle of relevance: it makes the information behind the link suddenly seem relevant to answer a question created in the user's mind by the clickbait title [140, 141]. Untargeted clickbait does not, however, align with the *communicative* principle of relevance: it should be the most relevant point for the user. The user has other reasons to be on social media that may hold more sway than the clickbait's curiosity-inducing title. Targeted clickbait aims to satisfy the conditions for both cognitive and communicative principles of relevance at the same time.

Our findings support the working of targeted clickbait based on relevance theory. Our results from the curation survey represent high scores for users' *Likelihood* to click on targeted clickbait (see §3.2.2). Similar results are echoed in the participatory design study, where most participants' decision to click on the post was influenced by its relevance (see §4.2.1). In the evaluation study, we see high scores for *Interest* in the post. Moreover, *Interest* has a strong correlation with *Likelihood* of users to click on the post (see §5.2.1). These findings support the working of targeted clickbait through relevance theory, and suggest that targeted clickbait can be a big threat in the near future with the increasing ease of access to artificial intelligence (text generation and image manipulation) and public information in social media. We encourage industry and academia to adopt preventive measures against the threat of targeted clickbait by developing and deploying appropriate tools and technologies. In doing that, we believe the user-informed warnings from our studies will function as an initial reference for future research and development.

6.3 Efficacy of Reflection through Interruption

We observe that users are primarily motivated to click on targeted clickbait due to its relevance and the manufactured information gap. Interstitial warnings shift users' focus from their desire to click on the post to a reflection of their action through interruption [72]. The effect of interruption is apparent from the efficacy of control warnings and our participants' open-ended responses (see §5.2.2). Although control warnings do not contain any stories, thanks to their interstitial nature, they induced participants to reflect on their actions. These findings support the prior literature in phishing and malware, demonstrating the efficacy of interstitial warnings [44, 72, 117, 132, 158].

6.4 The Power of Storytelling

Similar to prior work that leveraged persuasive narratives (stories) to change user behavior [37, 54, 63, 71, 84, 99, 137], our findings show that users can be persuaded to avoid targeted clickbait using storytelling, where persuasive narratives in warnings helps them to reflect and make an informed decision about their online safety. The efficacy of story-based warnings is clear from the significant reduction in the participants' *Interest* and *Likelihood* to click on targeted clickbait (see §5.2.3, and §5.2.4). Further, our findings highlight the efficacy of story-based warnings in measures such as *Perspicuity*, *Usefulness*, and *Credibility*, highlighting the importance of combining information with stories (see §5.2.3, and §5.2.4).

We found that a few participants had already experienced targeted clickbait through the manipulated pictures of their friends in social media (see §4.2.1). As access to image-manipulation and text-generation technologies increase the viability of targeted clickbait, story-based warning can be a reliable way to support users' in making informed decisions. We suggest leveraging the effective variations of stories interchangeably to avoid *habituation*. Vance et al. [146] point to habituation as a primary inhibitor to the efficacy of security warnings, and suggests using variations of the warning to address this issue.

6.5 Threat Personalization in Warnings

Our findings lead us to recommend the personalization of warnings against targeted clickbait. We observe that users have varying perceptions of the threat level of a post (see §4.2.4). Our participants also perceived different levels of threats from two targeting factors, even though both of them could be equally harmful. Therefore, we emphasize that a warning against clickbait should convey the correct threat level. Our findings suggest using colors in the overlays of warnings based on color theory to portray threat levels. The correct conveyance of a threat level offers multiple benefits. For instance, users immediately understand the threat they will face if they click on the post, persuading them to avoid malicious websites – otherwise, some users might choose to ignore the warning, thinking the post is not harmful. If all warnings use a red overlay and convey a high level of threat, then a user may lose trust in the warning if they keep encountering them for posts that they find non-malicious but only waste their time. Such experiences can lead them to ignore similar warnings in the future [146].

We acknowledge the need to develop artificial intelligence to scalably identify the threat levels of clickbait with accuracy, and thus recommend future research in this direction.

6.6 Limitations and Future Work

The participant pool in our studies is limited to users from the U.S. and Canada. We note that the societal and cultural background, literacy rate, public policy, economic condition, and infrastructural support could impact users' perceptions and behavior towards targeted clickbait and the story-based warnings designed to protect against it. Recent studies [5, 6, 41, 108, 129, 133, 147] point towards the importance of looking beyond Western contexts. To this end, we encourage future research to validate and extend our work, and include participants from diverse backgrounds and geographic regions, including developing countries.

Twenty four participants took part in our participatory design study, where we followed widely-used methods for qualitative research [9, 15, 22, 23, 130], focusing in depth on a small number of participants. We acknowledge the limitations of such study that a different set of samples might yield varying results. Thus, we do not draw any quantitative, generalizable conclusions from this study. Rather, we leveraged the findings from participatory design to conduct an evaluation survey, where we targeted a medium effect size based on our power analysis.

Since users' security and privacy perceptions are positively influenced by their knowledge and technical efficacy [68, 94, 128], and the majority of our participants are educated, we speculate that the perceptions and behavior of users reported in this paper represent an upper bound in the context of protecting against targeted clickbait. We recommend future work to focus on less-educated population in understanding their behavior towards targeted clickbait and identify the scope of enhancing our warning designs to address their needs and expectations.

7 CONCLUSION

Our findings contribute to the taxonomy of targeting factors in targeted clickbait and countermeasures against it **(RQ1)**, which provides directions and framework for future exploration on this problem. Using the taxonomy, we create story-based warnings against targeted clickbait through user participation **(RQ2)**. Our evaluation then shows the efficacy of these story-based warnings **(RQ3)**. Our study reports on the efficacy of targeted clickbait in tricking users (see §5.2.1) and the motivations behind users' decisions to interact with targeted clickbait, including relevance of the post, the curiosity gap created by the post, and habituation due to non-consequential past experiences (see §4.2.1). It also shows that at least three variations of story-based warnings designed through user participation can be effective against targeted clickbait (see §5.2.3, and §5.2.4). Finally, we recommend threat-level personalization and interchangeable use of user-stories to resist habituation.

ACKNOWLEDGMENTS

We thank our participants in this study. We are thankful to the reviewers for their thoughtful and detailed feedback. This work was supported in part by the National Science Foundation under Grant Nos. 1949699 and 1949694.

REFERENCES

[1] Charles Abraham and Susan Michie. 2008. A taxonomy of behavior change techniques used in interventions. *Health psychology* 27, 3 (2008), 379.

[2] Sally Adee. 2016. Scammer AI can tailor clickbait to you for phishing attacks. https://www.newscientist.com/article/2101483-scammer-ai-can-tailor-clickbait-to-you-for-phishing-attacks/.

[3] Amol Agrawal. 2016. Clickbait detection using deep learning. In *2016 2nd International Conference on Next Generation Computing Technologies (NGCT)*. 268–272. https://doi.org/10.1109/NGCT.2016.7877426

[4] Icek Ajzen. 1991. The theory of planned behavior. *Organizational behavior and human decision processes* 50, 2 (1991), 179–211.

[5] Mahdi Nasrullah Al-Ameen and Huzeyfe Kocabas. 2020. "I cannot do anything": User's Behavior and Protection Strategy upon Losing, or Identifying Unauthorized Access to Online Account. In *Symposium on Usable Privacy and Security (Poster Session)*.

[6] Mahdi Nasrullah Al-Ameen, Huzeyfe Kocabas, Swapnil Nandy, and Tanjina Tamanna. 2021. "We, three brothers have always known everything of each other": A Cross-cultural Study of Sharing Digital Devices and Online Accounts. *Proceedings on Privacy Enhancing Technologies* 2021, 4 (2021), 203–224.

[7] Jennifer Allen, Cameron Martel, and David G Rand. 2022. Birds of a feather don't fact-check each other: Partisanship and the evaluation of news in Twitter's Birdwatch crowdsourced fact-checking program. In *CHI Conference on Human Factors in Computing Systems*. 1–19.

[8] Christopher J Armitage and Mark Conner. 2001. Efficacy of the theory of planned behaviour: A meta-analytic review. *British journal of social psychology* 40, 4 (2001), 471–499.

[9] Hanieh Atashpanjeh, Arezou Behfar, Cassity Haverkamp, Maryellen McClain Verdoes, and Mahdi Nasrullah Al-Ameen. 2022. Intermediate Help with Using Digital Devices and Online Accounts: Understanding the Needs, Expectations, and Vulnerabilities of Young Adults. In *International Conference on Human-Computer Interaction*. Springer, 3–15.

[10] Jeffrey Avery, Mohammed Almeshekah, and Eugene Spafford. 2017. Offensive deception in computing. In *International Conference on Cyber Warfare and Security*. Academic Conferences International Limited, 23.

[11] Arun Babu, Annie Liu, and Jordan Zhang. 2017. New updates to reduce clickbait headlines. *Facebook Newsroom* (2017).

[12] Eunsoo Baek, Ho Jung Choo, Xiaoyong Wei, and So-Yeon Yoon. 2020. Understanding the virtual tours of retail stores: how can store brand experience promote visit intentions? *International Journal of Retail & Distribution Management* (2020).

[13] Natasha Bajema, Craig S Smith, and Dan Garisto. 2022. AI's Real Worst-Case Scenarios: Who needs Terminators when you have precision clickbait and ultra-deepfakes? *IEEE Spectrum* 59, 1 (2022), 8–14.

[14] M Baskinger and B Hanington. 2008. Sustaining Autonomous Living for Older People Through Inclusive Strategies for Home Appliance Design. In *Designing Inclusive Futures*. Springer, 117–126.

[15] Kathy Baxter, Catherine Courage, and Kelly Caine. 2015. *Understanding Your Users: A Practical Guide to User Research Methods* (2 ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[16] Arezou Behfar, Hanieh Atashpanjeh, and Mahdi Nasrullah Al-Ameen. 2023. Can Password Meter be More Effective Towards User Attention, Engagement, and Attachment?: A Study of Metaphor-based Designs. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 164–171.

[17] Nigel Bevan and Ian Curson. 1998. Planning and implementing user-centred design. In *CHI 98 conference summary on Human factors in computing systems*. 111–112.

[18] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–30.

[19] Md Momen Bhuiyan, Kexin Zhang, Kelsey Vick, Michael A Horning, and Tanushree Mitra. 2018. FeedReflect: A tool for nudging users to assess news credibility on Twitter. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 205–208.

[20] Monika Bickert. 2019. Combatting vaccine misinformation. *Facebook Newsroom* 1 (2019).

[21] Jonas Nygaard Blom and Kenneth Reinecke Hansen. 2015. Click bait: Forward-reference as lure in online news headlines. *Journal of Pragmatics* 76 (2015), 87–100.

[22] Richard E Boyatzis. 1998. *Transforming qualitative information: Thematic analysis and code development.* sage, Thousand Oaks, CA, USA.

[23] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[24] Paolo Buono, Giuseppe Desolda, Francesco Greco, and Antonio Piccinno. 2023. Let warnings interrupt the interaction and explain: designing and evaluating phishing email warnings. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–6.

[25] José Luis Caivano. 1998. Color and semiotics: A two-way street. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur* 23, 6 (1998), 390–401.

[26] Stuart Candy and Jake Dunagan. 2017. Designing an experiential scenario: The people who vanished. *Futures* 86 (2017), 136–153.

[27] Abhijnan Chakraborty, Bhargavi Paranjape, Sourya Kakarla, and Niloy Ganguly. 2016. Stop clickbait: Detecting and preventing clickbaits in online news media. In *2016 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*. IEEE, 9–16.

[28] Natalia P Chapanis and Alphonse Chapanis. 2017. Cognitive dissonance: Five years later. *Attitude change* (2017), 116–153.

[29] Yimin Chen, Niall J Conroy, and Victoria L Rubin. 2015. Misleading online content: recognizing clickbait as" false news". In *Proceedings of the 2015 ACM on workshop on multimodal deception detection*. 15–19.

[30] Bobby Chesney and Danielle Citron. 2019. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.* 107 (2019), 1753.

[31] Susan Ciccantelli and Jason Magidson. 1993. Consumer idealized design: involving consumers in the product development process. *Journal of Product Innovation Management: An International Publication of the Product Development & Management Association* 10, 4 (1993), 341–347.

[32] Billy Clark. 2013. *Relevance theory.* Cambridge University Press.

[33] Colin Crowell. 2017. Our approach to bots and misinformation. *Twitter public policy* (2017).

[34] Brita Curum and Kavi Kumar Khedo. 2021. Cognitive load management in mobile learning systems: principles and theories. *Journal of Computers in Education* 8 (2021), 109–136.

[35] Maria D. Molina, S Shyam Sundar, Md Main Uddin Rony, Naeemul Hassan, Thai Le, and Dongwon Lee. 2021. Does clickbait actually attract more clicks? Three clickbait studies you must read. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–19.

[36] Justin A DeSimone, Peter D Harms, and Alice J DeSimone. 2015. Best practice recommendations for data screening. *Journal of Organizational Behavior* 36, 2 (2015), 171–181.

[37] Laurence Dessart and Willem Standaert. 2023. Strategic storytelling in the age of sustainability. *Business Horizons* 66, 3 (2023), 371–385.

[38] Lindsey Donato. 2016. The Dark Side of Clickbait – This Will Shock You. https://www.blumshapiro.com/insights/clickbait-cybersecurity-ransomware-ct-ma-ri/.

[39] Dana Lynn Driscoll and Allen Brizee. 2010. Creating good interview and survey questions. *Purdue Online Writing Lab* (2010).

[40] Prakriti Dumaru and Mahdi Nasrullah Al-Ameen. 2023. "After she fell asleep, it went to my next podcast, which was about a serial killer": Unveiling Needs and Expectations Regarding Parental Control within Digital Assistant. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 17–21.

[41] Prakriti Dumaru, Ankit Shrestha, Rizu Paudel, Arezou Behfar, Hanieh Atashpanjeh, and Mahdi Nasrullah Al-Ameen. 2023. "I Have Learned that Things are Different here": Understanding the Transitional Challenges with Technology Use After Relocating to the USA. In *International Conference on Human-Computer Interaction*. Springer, 201–220.

[42] Prakriti Dumaru, Ankit Shrestha, Rizu Paudel, Cassity Haverkamp, Maryellen Brunson McClain, and Mahdi Nasrullah Al-Ameen. 2023. "... I have my dad, sister, brother, and mom's password": unveiling users' mental models of security and privacy-preserving tools. *Information & Computer Security* (2023).

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

[44] Serge Egelman, Lorrie Faith Cranor, and Jason Hong. 2008. You've been warned: An empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1065–1074.

[45] Pelle Ehn. 2017. Scandinavian design: On participation and skill. In *Participatory design*. CRC Press, 41–77.

[46] Paul Ekman et al. 1999. Basic emotions. *Handbook of cognition and emotion* 98, 45-60 (1999), 16.

[47] Andrew J Elliot and Markus A Maier. 2012. Color-in-context theory. In *Advances in experimental social psychology*. Vol. 45. Elsevier, 61–125.

[48] Yu-Min Fang, Kuen-Meau Chen, and Yi-Jhen Huang. 2016. Emotional reactions of different interface formats: Comparing digital and traditional board games. *Advances in Mechanical Engineering* 8, 3 (2016), 1687814016641902.

[49] Hany Farid. 2022. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety* 1, 4 (2022).

[50] Leon Festinger. 1962. Cognitive dissonance. *Scientific American* 207, 4 (1962), 93–106.

[51] Klaus Fog, Christian Budtz, and Baris Yakaboylu. 2005. *Storytelling*. Springer.

[52] Dilrukshi Gamage, James Stomber, Farnaz Jahanbakhsh, Bill Skeet, and Gautam Kishore Shahi. 2022. Designing Credibility Tools To Combat Mis/Disinformation: A Human-Centered Approach. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–4.

[53] Christine Geeng, Savanna Yee, and Franziska Roesner. 2020. Fake news on Facebook and Twitter: Investigating how people (don't) investigate. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–14.

[54] David A Gilliam and Karen E Flaherty. 2015. Storytelling by the sales force and its effect on buyer–seller exchange. *Industrial Marketing Management* 46 (2015), 132–142.

[55] Nathaniel Gleicher. 2019. Removing coordinated inauthentic behavior from China. *Facebook Newsroom* 19 (2019).

[56] Peter Leo Gorski, Yasemin Acar, Luigi Lo Iacono, and Sascha Fahl. 2020. Listen to developers! a participatory design study on security warnings for cryptographic apis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[57] Linda J Griffith and S David Leonard. 1997. Association of colors with warning signal words. *International Journal of Industrial Ergonomics* 20, 4 (1997), 317–325.

[58] JoAnn T Hackos and Janice Redish. 1998. *User and task analysis for interface design*. Vol. 1. Wiley New York.

[59] Bruce Hanington and Bella Martin. 2019. *Universal methods of design expanded and revised: 125 Ways to research complex problems, develop innovative ideas, and design effective solutions*. Rockport publishers.

[60] Marja Harjumaa and Harri Oinas-Kukkonen. 2007. Persuasion theories and IT design. In *Persuasive Technology: Second International Conference on Persuasive Technology, PERSUASIVE 2007, Palo Alto, CA, USA, April 26-27, 2007, Revised Selected Papers 2*. Springer, 311–314.

[61] Naeemul Hassan, Mohammad Yousuf, Md Mahfuzul Haque, Javier A. Suarez Rivas, and Md Khadimul Islam. 2019. Examining the roles of automation, crowds and professionals towards sustainable fact-checking. In *Companion Proceedings of The 2019 World Wide Web Conference*. 1001–1006.

[62] Steffi Heidig, Julia Müller, and Maria Reichelt. 2015. Emotional design in multimedia learning: Differentiation on relevant design features and their effects on emotions and learning. *Computers in Human behavior* 44 (2015), 81–95.

[63] Jeremy Heyer, Nirmal Kumar Raveendranath, and Khairi Reda. 2020. Pushing the (visual) narrative: the effects of prior knowledge elicitation in provocative topics. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[64] Hans Hoeken, Matthijs Kolthoff, and José Sanders. 2016. Story perspective and character similarity as drivers of identification and narrative persuasion. *Human communication research* 42, 2 (2016), 292–311.

[65] Y Linlin Huang, Kate Starbird, Mania Orand, Stephanie A Stanek, and Heather T Pedersen. 2015. Connected through crisis: Emotional proximity and the spread of misinformation online. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*. 969–980.

[66] Markus Huber, Stewart Kowalski, Marcus Nohlberg, and Simon Tjoa. 2009. Towards automating social engineering using social networking sites. In *2009 International Conference on Computational Science and Engineering*, Vol. 3. IEEE, 117–124.

[67] Taylor Hughes, Jeff Smith, and Alex Leavitt. 2018. Helping People Better Assess the Stories They See in News Feed with the Context Button| Facebook Newsroom. *Facebook Newsroom* (2018).

[68] Iulia Ion, Rob Reeder, and Sunny Consolvo. 2015. "…No One Can Hack My Mind": Comparing Expert and Non-Expert Security Practices. In *Proceedings of the Eleventh USENIX Conference on Usable Privacy and Security* (Ottawa, Canada) *(SOUPS '15)*. USENIX Association, USA, 327–346.

[69] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. 2010. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*. 64–67.

[70] Kristina Risom Jespersen. 2008. *User Driven Product Development: Creating a User-Involving Culture*. Samfundslitteratur.

[71] Michael D Jones and Holly Peterson. 2017. Narrative persuasion and storytelling as climate communication strategies. In *Oxford Research Encyclopedia of Climate Science*.

[72] Ben Kaiser, Jerry Wei, Eli Lucherini, Kevin Lee, J Nathan Matias, and Jonathan Mayer. 2021. Adapting security warnings to counter online disinformation. In *30th USENIX Security Symposium (USENIX Security 21)*. 1163–1180.

[73] Michael J Kalsher and Linda J Frederick. 1998. Hazard level perceptions of warning components and configurations. *International Journal of Cognitive Ergonomics* 2, 1-2 (1998), 123–143.

[74] Slava Kalyuga. 2011. Cognitive load theory: How many types of load does it really need? *Educational Psychology Review* 23 (2011), 1–19.

[75] Hema Karande, Rahee Walambe, Victor Benjamin, Ketan Kotecha, and TS Raghu. 2021. Stance detection with BERT embeddings for credibility analysis of information on social media. *PeerJ Computer Science* 7 (2021), e467.

[76] Jan Kietzmann, Linda W Lee, Ian P McCarthy, and Tim C Kietzmann. 2020. Deepfakes: Trick or treat? *Business Horizons* 63, 2 (2020), 135–146.

[77] Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. 2017. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*. 3677–3685.

[78] Rebecca J Krause and Derek D Rucker. 2020. Strategic storytelling: When narratives help versus hurt the persuasive power of facts. *Personality and Social Psychology Bulletin* 46, 2 (2020), 216–227.

[79] Katharina Krombholz, Heidelinde Hobel, Markus Huber, and Edgar Weippl. 2015. Advanced social engineering attacks. *Journal of Information Security and applications* 22 (2015), 113–122.

[80] Richard A Krueger. 2014. *Focus groups: A practical guide for applied research*. Sage publications.

[81] Vaibhav Kumar, Dhruv Khattar, Siddhartha Gairola, Yash Kumar Lal, and Vasudeva Varma. 2018. Identifying clickbait: A multi-strategy approach using neural networks. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. 1225–1228.

[82] Franki Y.H. Kung, Navio Kwok, and Douglas J. Brown. 2018. Are attention check questions a threat to scale validity? *Applied Psychology* 67, 2 (2018), 264–283. https://doi.org/10.1111/apps.12108

[83] Franki YH Kung, Navio Kwok, and Douglas J Brown. 2018. Are attention check questions a threat to scale validity? *Applied Psychology* 67, 2 (2018), 264–283.

[84] Linda C Lederman and Lisa M Menegatos. 2011. Sustainable recovery: The self-transformative power of storytelling in Alcoholics Anonymous. *Journal of Groups in Addiction & Recovery* 6, 3 (2011), 206–227.

[85] Stephan Lewandowsky, Ullrich KH Ecker, Colleen M Seifert, Norbert Schwarz, and John Cook. 2012. Misinformation and its correction: Continued influence and successful debiasing. *Psychological science in the public interest* 13, 3 (2012), 106–131.

[86] Patrick J Lewis. 2011. Storytelling as research/research as storytelling. *Qualitative inquiry* 17, 6 (2011), 505–510.

[87] Gitte Lindgaard, Cathy Dudek, Devjani Sen, Livia Sumegi, and Patrick Noonan. 2011. An exploration of relations between visual appeal, trustworthiness and perceived usability of homepages. *ACM Transactions on Computer-Human Interaction (TOCHI)* 18, 1 (2011), 1–30.

[88] Eric Loepp and Jarrod T Kelly. 2020. Distinction without a difference? An assessment of MTurk Worker types. *Research & Politics* 7, 1 (2020), 2053168019901185.

[89] George Loewenstein. 1994. The psychology of curiosity: A review and reinterpretation. *Psychological bulletin* 116, 1 (1994), 75.

[90] Tessa Lyons. 2017. Replacing disputed flags with related articles. *Facebook Newsroom* 20 (2017).

[91] Roberto Martinez-Maldonado, Vanessa Echeverria, Gloria Fernandez Nieto, and Simon Buckingham Shum. 2020. From data to insights: A layered storytelling approach for multimodal learning analytics. In *Proceedings of the 2020 chi conference on human factors in computing systems*. 1–15.

[92] Alice E Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011), 114–133.

[93] Richard E Mayer and Roxana Moreno. 2003. Nine ways to reduce cognitive load in multimedia learning. *Educational psychologist* 38, 1 (2003), 43–52.

[94] Michelle L. Mazurek, Saranga Komanduri, Timothy Vidas, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Patrick Gage Kelley, Richard Shay, and Blase Ur. 2013. Measuring Password Guessability for an Entire University. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security* (Berlin, Germany) *(CCS '13)*. Association for Computing Machinery, New York, NY, USA, 173–186. https://doi.org/10.1145/2508859.2516726

[95] Susan Michie, Michelle Richardson, Marie Johnston, Charles Abraham, Jill Francis, Wendy Hardeman, Martin P Eccles, James Cane, and Caroline E Wood. 2013. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine* 46, 1 (2013), 81–95.

[96] Roxana Moreno and Alfred Valdez. 2005. Cognitive load and learning effects of having students organize pictures and words in multimedia environments: The role of student interactivity and feedback. *Educational Technology Research and Development* 53, 3 (2005), 35–45.

[97] David L Morgan. 1996. *Focus groups as qualitative research*. Vol. 16. Sage publications.

[98] Adam Mosseri. 2018. Helping ensure news on Facebook is from trusted sources. *Facebook Newsroom. Online verfügbar unter https://newsroom. fb. com/news/2018/01/trusted-sources/, zuletzt aktualisiert am* 19 (2018), 2018.

[99] Elizabeth L Murnane, Xin Jiang, Anna Kong, Michelle Park, Weili Shi, Connor Soohoo, Luke Vink, Iris Xia, Xin Yu, John Yang-Sammataro, et al. 2020. Designing ambient narrative-based interfaces to reflect and motivate physical activity. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[100] Don Norman. 2002. Emotion & design: attractive things work better. *interactions* 9, 4 (2002), 36–42.

[101] Don Norman. 2013. *The design of everyday things: Revised and expanded edition*. Basic books.

[102] Donald A Norman. 2004. Introduction to this special section on beauty, goodness, and usability. *Human–Computer Interaction* 19, 4 (2004), 311–318.

[103] Donald A Norman and Andrew Ortony. 2003. Designers and users: Two perspectives on emotion and design. In *Symposium on foundations of interaction design*. 1–13.

[104] Daniel J O'keefe. 2015. *Persuasion: Theory and research*. Sage Publications.

[105] Daniela Oliveira, Harold Rocha, Huizi Yang, Donovan Ellis, Sandeep Dommaraju, Melis Muradoglu, Devon Weir, Adam Soliman, Tian Lin, and Natalie Ebner. 2017. Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing. In *Proceedings of the 2017 chi conference on human factors in computing systems*. 6412–6424.

[106] Will Oremus, Chris Alcantara, Jeremy B. Merrill, and Artur Galocha. 2021. How Facebook shapes your feed. https://www.washingtonpost.com/technology/interactive/2021/how-facebook-algorithm-works/.

[107] Andy O'Donnell. 2018. What is Clickbait?: What's really happening when you click that link to finish an irresistible story. https://www.lifewire.com/the-dark-side-of-clickbait-2487506.

[108] Rizu Paudel, Prakriti Dumaru, Ankit Shrestha, Huzeyfe Kocabas, and Mahdi Nasrullah Al-Ameen. 2023. A Deep Dive into User's Preferences and Behavior around Mobile Phone Sharing. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–22.

[109] Rizu Paudel, Ankit Shrestha, Prakriti Dumaru, and Mahdi Nasrullah Al-Ameen. 2023. " It doesn't just feel like something a lawyer slapped together." Mental-Model-Based Privacy Policy for Third-Party Applications on Facebook. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 298–306.

[110] Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods* 46, 4 (2014), 1023–1031.

[111] Gordon Pennycook, Jonathon McPhetres, Bence Bago, and David G Rand. 2022. Beliefs about COVID-19 in Canada, the United Kingdom, and the United States: A novel test of political polarization and motivated reasoning. *Personality and Social Psychology Bulletin* 48, 5 (2022), 750–765.

[112] Gordon Pennycook and David G Rand. 2019. Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition* 188 (2019), 39–50.

[113] Gordon Pennycook and David G Rand. 2020. Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of personality* 88, 2 (2020), 185–200.

[114] Dorian Peters, Rafael A Calvo, and Richard M Ryan. 2018. Designing for motivation, engagement and wellbeing in digital experience. *Frontiers in psychology* (2018), 797.

[115] Martin Potthast, Sebastian Köpsel, Benno Stein, and Matthias Hagen. 2016. Clickbait detection. In *European conference on information retrieval*. Springer, 810–817.

[116] Elissa M Redmiles, Neha Chachra, and Brian Waismeyer. 2018. Examining the Demand for Spam: Who Clicks?. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. ACM, 212.

[117] Robert W Reeder, Adrienne Porter Felt, Sunny Consolvo, Nathan Malkin, Christopher Thompson, and Serge Egelman. 2018. An experience sampling study of user reactions to browser warnings in the field. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–13.

[118] Kris Rides. 2017. Clickbait Malware Sites. https://www.linkedin.com/pulse/clickbait-malware-sites-kris-rides/.

[119] Yoel Roth and Del Harvey. 2018. How Twitter is fighting spam and malicious automation. *Twitter [blog], June* (2018).

[120] Twitter Safety. 2019. Information operations directed at Hong Kong. *Twitter Blog, August* 19 (2019).

[121] Fatima Salahdine and Naima Kaabouch. 2019. Social engineering attacks: A survey. *Future internet* 11, 4 (2019), 89.

[122] Elizabeth B-N Sanders, Eva Brandt, and Thomas Binder. 2010. A framework for organizing the tools and techniques of participatory design. In *Proceedings of the 11th biennial participatory design conference*. 195–198.

[123] Martin Schrepp, Andreas Hinderks, and Jörg Thomaschewski. 2014. Applying the user experience questionnaire (UEQ) in different evaluation scenarios. In *International Conference of Design, User Experience, and Usability*. Springer, 383–392.

[124] Martin Schrepp and Jörg Thomaschewski. 2019. Handbook for the modular extension of the User Experience Questionnaire. *Retrieved from www. ueq-online. org* (2019).

[125] Yaacov Schul. 1993. When warning succeeds: The effect of warning on success in ignoring invalid information. *Journal of Experimental Social Psychology* 29, 1 (1993), 42–62.

[126] Kate Scott. 2019. *Referring expressions, pragmatics, and style: Reference and beyond*. Cambridge University Press.

[127] Kate Scott. 2021. You won't believe what's in this paper! Clickbait, relevance and the curiosity gap. *Journal of pragmatics* 175 (2021), 53–66.

[128] Sovantharith Seng, Huzeyfe Kocabas, Mahdi Nasrullah Al-Ameen, and Matthew Wright. 2019. Poster: Understanding User's Decision to Interact with Potential Phishing Posts on Facebook Using a Vignette Study. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security* (London, United Kingdom) *(CCS '19)*. Association for Computing Machinery, New York, NY, USA, 2617–2619. https://doi.org/10.1145/3319535.3363270

[129] Farhana Shahid, Srujana Kamath, Annie Sidotam, Vivian Jiang, Alexa Batino, and Aditya Vashistha. 2022. " It Matches My Worldview": Examining Perceptions and Attitudes Around Fake Videos. In *CHI Conference on Human Factors in Computing Systems*. 1–15.

[130] Ankit Shrestha, Prakriti Dumaru, Rizu Paudel, and Mahdi Nasrullah Al-Ameen. 2023. Understanding the Challenges in Academia to Prepare Nursing Students for Digital Technology Use at Workplace. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing*. 96–100.

[131] Ankit Shrestha, Danielle M Graham, Prakriti Dumaru, Rizu Paudel, Kristin A Searle, and Mahdi Nasrullah Al-Ameen. 2022. Understanding the Behavior, Challenges, and Privacy Risks in Digital Technology Use by Nursing Professionals. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–22.

[132] Ankit Shrestha, Rizu Paudel, Prakriti Dumaru, and Mahdi Nasrullah Al-Ameen. 2023. Towards Improving the Efficacy of Windows Security Notifier for Apps from Unknown Publishers: The Role of Rhetoric. In *International Conference on Human-Computer Interaction*. Springer, 101–121.

[133] Ankit Shrestha, Tanusree Sharma, Pratyasha Saha, Syed Ishtiaque Ahmed, and Mahdi Nasrullah Al-Ameen. 2023. A first look into software security practices in bangladesh. *ACM Journal on Computing and Sustainable Societies* 1, 1 (2023), 1–24.

[134] Kai Shu, Suhang Wang, Thai Le, Dongwon Lee, and Huan Liu. 2018. Deep headline generation for clickbait detection. In *2018 IEEE International Conference on Data Mining (ICDM)*. IEEE, 467–476.

[135] Mario Silic, Dianne Cyr, Andrea Back, and Adrian Holzer. 2017. Effects of color appeal, perceived risk and culture on user's decision in presence of warning banner message. In *Silic, M., Cyr, D., Back, A., & Holzer, A.(2017, January). Effects of Color Appeal, Perceived Risk and Culture on User's Decision in Presence of Warning Banner Message. In Proceedings of the 50th Hawaii International Conference on System Sciences*.

[136] Ana Siljak. 2004. Power and Persuasion: Ideology and Rhetoric in Communist Yugoslavia, 1944–1953.

[137] Annette Simmons. 2019. *The story factor: Inspiration, influence, and persuasion through the art of storytelling*. Basic books.

[138] Herbert W Simons and Jean Jones. 2011. *Persuasion in society*. Taylor & Francis.

[139] Fioravante Souza. 2015. Analyzing a Facebook Clickbait Worm. https://blog.sucuri.net/2015/06/analyzing-a-facebook-clickbait-worm.html.

[140] Dan Sperber, Francesco Cara, and Vittorio Girotto. 1995. Relevance theory explains the selection task. *Cognition* 57, 1 (1995), 31–95.

[141] Dan Sperber and Deirdre Wilson. 1986. *Relevance: Communication and cognition*. Vol. 142. Citeseer.

[142] Tom G Stevens. 1998. *You Can Choose to be Happy:" rise Above" Anxiety, Anger and Depression*. Wheeler Sutton Publishing Company.

[143] Sara Su. 2017. New Test with Related Articles. *Facebook Newsroom, April* 25 (2017).

[144] Rashid Tahir, Brishna Batool, Hira Jamshed, Mahnoor Jameel, Mubashir Anwar, Faizan Ahmed, Muhammad Adeel Zaffar, and Muhammad Fareed Zaffar. 2021. Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[145] Jacqueline Urakami, Yeongdae Kim, Hiroki Oura, and Katie Seaborn. 2022. Finding Strategies Against Misinformation in Social Media: A Qualitative Study. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[146] Anthony Vance, Brock Kirwan, Daniel Bjornn, Jeffrey Jenkins, and Bonnie Brinton Anderson. 2017. What do we really know about how habituation to warnings occurs over time? A longitudinal fMRI study of habituation and polymorphic warnings. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2215–2227.

[147] Rama Adithya Varanasi, Joyojeet Pal, and Aditya Vashistha. 2022. Accost, Accede, or Amplify: Attitudes towards COVID-19 Misinformation on WhatsApp in India. In *CHI Conference on Human Factors in Computing Systems*. 1–17.

[148] Mike Vulpo. 2022. You Won't Believe How Much Money Tom Brady Made During His 22-Year Football Career. https://www.eonline.com/news/1318262/you-wont-believe-how-much-money-tom-brady-made-during-his-22-year-football-career.

[149] Andrew Wachtel. 2003. Power and Persuasion: Ideology and Rhetoric in Communist Yugoslavia, 1944-1953. By Carol S. Lilly. Boulder, Colo.: Westview Press, 2001. xii, 272 pp. Notes. Bibliography. Index. Illustrations. $45.00, paper. *Slavic Review* 62, 1 (2003), 166–167.

[150] Bert Weijters and Hans Baumgartner. 2012. Misresponse to reversed and negated items in surveys: A review. *Journal of Marketing Research* 49, 5 (2012), 737–747.

[151] Michael Wesch. 2009. YouTube and you: Experiences of self-awareness in the context collapse of the recording webcam. *Explorations in media ecology* 8, 2 (2009), 19–34.

[152] Christopher R Wilkinson and Antonella De Angeli. 2014. Applying user centred and participatory design approaches to commercial product development. *Design Studies* 35, 6 (2014), 614–631.

[153] Deirdre Wilson and Dan Sperber. 2006. Relevance theory. *The handbook of pragmatics* (2006), 606–632.

[154] Deirdre Wilson and Dan Sperber. 2012. *Meaning and relevance.* Cambridge University Press.

[155] Terry Winograd. 1996. *Bringing design to software.* ACM.

[156] Michael S Wogalter, Dave DeJoy, and Kenneth R Laughery. 1999. *Warnings and risk communication.* CRC Press.

[157] Michael S Wogalter, Amy B Magurno, Ann W Carter, Julie A Swindell, William J Vigilante, and Jason G Daurity. 1995. Hazard associations of warning header components. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 39. SAGE Publications Sage CA: Los Angeles, CA, 979–983.

[158] Min Wu, Robert C Miller, and Simson L Garfinkel. 2006. Do security toolbars actually prevent phishing attacks?. In *Proceedings of the SIGCHI conference on Human Factors in computing systems.* 601–610.

[159] Zhan Xu, Mary Laffidy, and Lauren Ellis. 2023. Clickbait for climate change: comparing emotions in headlines and full-texts and their engagement. *Information, Communication & Society* 26, 10 (2023), 1915–1932.

[160] Savvas Zannettou, Sotirios Chatzis, Kostantinos Papadamou, and Michael Sirivianos. 2018. The good, the bad and the bait: Detecting and characterizing clickbait on YouTube. In *2018 IEEE Security and Privacy Workshops (SPW).* IEEE, 63–69.

[161] Leah Zhang-Kennedy, Sonia Chiasson, and Robert Biddle. 2013. Password advice shouldn't be boring: Visualizing password guessing attacks. In *2013 APWG eCrime Researchers Summit.* 1–11. https://doi.org/10.1109/eCRS.2013.6805770

[162] Hai-Tao Zheng, Jin-Yuan Chen, Xin Yao, Arun Kumar Sangaiah, Yong Jiang, and Cong-Zhi Zhao. 2018. Clickbait convolutional neural network. *Symmetry* 10, 5 (2018), 138.

[163] Yiwei Zhou. 2017. Clickbait detection in tweets using self-attentive network. *arXiv preprint arXiv:1710.05364* (2017).

[164] Verena Zimmermann, Karola Marky, and Karen Renaud. 2022. Hybrid password meters for more secure passwords–a comprehensive study of password meters including nudges and password information. *Behaviour & Information Technology* (2022), 1–44.

[165] ZoneAlarm.com. 2016. Click Bait Scams you should stay away from on Social Media. https://www.zonealarm.com/blog/2016/03/click-bait-scams-in-social-media/.

APPENDIX

A DESIGN TASK AND ITS THEORETICAL FRAMEWORK

The theory of planned behavior [4, 8] is the foremost theory of behavior change, which has brought about multiple studies developing Behavior Change Techniques (BCTs). The BCTs are clustered and taxonomized by the studies of Abraham et al. [1] and Michie et al. [95]. Based on these studies [1, 95], the conveyance of harm, and influence through others (reporting) are two of the most effective BCTs. These theories are in line with our findings from the curation survey (see §3), and motivate our design tasks (see §4).

Overlays. Based on color theory [25, 47], we provided three variations of overlays: red, yellow, and gray to convey the threat level (from high to neutral). Prior studies [57, 135] showed that colors in warnings could immediately convey the threat to users, helping them make informed decisions.

Headers. A header represents the first message users see in a warning [56, 73, 156, 157]. We provided two generic headers and a goal-oriented header. Generic header has two variations conveying that the post is baiting users, and users should not click on the post (see Figure 3). The goal-oriented header conveys that the post can cause harm (for *Harm Design*), and the post is misleading (for *Report Design*).

Navigation. Since users make their decisions upon reflecting on the content of an interstitial warning [24, 117], we provided two types of navigation buttons (see Figure 3): blue button *(Scroll Buttons)* that users can click to avoid the post, and a text button *(Ignore Buttons)* that would remove the warning and let users click through the post. In our study, we provided three variations of scroll buttons: "Keep Scrolling" (conveying suggestion), "Scroll Away" (conveying action), and "Keep Me Safe!" (conveying desirable outcome). We also provided three variations for the ignore buttons: "Proceed Anyway?" (conveying action), "Proceed Anyway? (unsafe)" (conveying a secondary reminder), and "I accept the risks" (conveying liability to users).

Messages. The messages depend on the goal of design tasks. For *Harm Design*, the messages varied based on different possible consequences of clicking on a clickbait (see Figure 3). For *Report Design*, messages varied based on the source, which reported the post as misleading. These sources include social media users, anonymous crowd of users, fact-checking tools, and professional fact-checkers.

Components for expression. The selected message is expressed through a story in the warning. We presented the story through the portrayal of characters, emotions, and other graphical components, including thought bubbles, speech bubbles, and arrows that are commonly used in comics (see Figure 3). We leveraged the taxonomy of basic emotions [46, 142] to choose a negative emotion surrounding anxiety/fear, sadness, or anger that is depicted on a victim's face.

B SUPPORTING TABLES FOR CURATION STUDIES

PID	Gender	Age Range	Current Education	Major
FGD1	Male	30-34 years old	Graduate degree (Ph.D.)	Civil Engineering (Water)
FGD2	Male	25-29 years old	Graduate degree (MS)	Plant Science
FGD3	Female	25-29 years old	Graduate degree (Ph.D.)	Computer Science (HCI)
FGD4	Female	25-29 years old	Graduate degree (Ph.D.)	Computer Science (HCI)
FGD5	Female	25-29 years old	Graduate degree (MS)	Computer Science (HCI)
FGD6	Female	30-34 years old	Four-year college degree	Physical Therapy

Table 6: Summary of the demographics information about the gender, age range, current education, and major of the 6 participants who took part in the brainstorm and select FGD

Demographic	Demographic Group	N
Gender	Male	17
	Female	13
Age range	18-24 years old	3
	25-29 years old	5
	30-34 years old	5
	35-39 years old	7
	40-44 years old	6
	45-49 years old	3
	50-54 years old	1

Race	White	24
	Asian	2
	Black/African American	1
	Hispanic or Latino	1
	Native American	1
	Mixed Race	1
Education	High School Graduate	6
	Two-year College Degree	2
	Four-year College Degree	17
	Graduate degree (MS/PhD)	5

Table 7: Demographic Information of the Participants in the Curation Survey (N=Number of Participants)

Targeting Factor	Description	*Likelihood*
Face-swap (Friends)	Users are targeted through posts using face-swapped images of their friends (can be created through AI technology)	83.33%
Affiliated Institutions	Users are targeted through posts about any educational or work institution that they are affiliated with	76.66%
Niche Activities	Users are targeted through posts about their unique interests and activities that are not quite common (e.g., fishing)	73.33%
Field of Study/Work	Users are targeted through posts related to their professional or educational field or discipline	73.33%
Relevant Location	Users are targeted through posts about location relevant to them (can be inferred from their social media posts or bio)	60%

Table 8: Taxonomy of Targeting Factors (*Likelihood* refers to the percentage of users that are likely to click on clickbait created using the targeting factor)

PID	Gender	Age Range	Race	Education
P1	Female	30-34 years old	White	Four-year college degree
P2	Female	18-24 years old	White	Four-year college degree
P3	Female	25-29 years old	White	Four-year college degree
P4	Female	25-29 years old	White	Four-year college degree
P5	Female	18-24 years old	White	Graduate degree (MS)
P6	Female	25-29 years old	African American	Graduate degree (Ph.D.)
P7	Female	35-39 years old	White	Two-year college degree
P8	Female	30-34 years old	White	Four-year college degree
P9	Female	30-34 years old	White	Two-year college degree
P10	Male	30-34 years old	North African	Graduate degree (MS)
P11	Male	18-24 years old	White	Four-year college degree
P12	Female	30-34 years old	Asian	Graduate degree (Ph.D.)
P13	Male	18-24 years old	White	Four-year college degree
P14	Female	18-24 years old	White	Four-year college degree
P15	Male	18-24 years old	White	Four-year college degree
P16	Male	40-44 years old	White	Four-year college degree
P17	Male	18-24 years old	White	Four-year college degree
P18	Male	25-29 years old	White	Four-year college degree
P19	Female	18-24 years old	White	Four-year college degree
P20	Male	18-24 years old	Asian	Graduate degree (MS)
P21	Female	18-24 years old	White	Four-year college degree
P22	Male	18-24 years old	White	Four-year college degree
P23	Male	25-29 years old	White	Four-year college degree
P24	Male	18-24 years old	White	Four-year college degree

Table 9: Demographics information of the participants who took part in the participatory design study