Designing Indicators to Combat Fake Media

Imani N. Sherman *University of Florida* shermani@ufl.edu

Elissa M. Redmiles *Microsoft Research* eredmiles@gmail.com Jack W. Stokes Microsoft Research jstokes@microsoft.com

Abstract—The growth of misinformation technology necessitates the need to identify fake videos. One approach to preventing the consumption of these fake videos is provenance which allows the user to authenticate media content to its original source. This research designs and investigates the use of provenance indicators to help users identify fake videos. We first interview users regarding their experiences with different misinformation modes (text, image, video) to guide the design of indicators within users' existing perspectives. Then, we conduct a participatory design study to develop and design fake video indicators. Finally, we evaluate participant-designed indicators via both expert evaluations and quantitative surveys with a large group of end-users. Our results provide concrete design guidelines for the emerging issue of fake videos. Our findings also raise concerns regarding users' tendency to overgeneralize from misinformation warning messages, suggesting the need for further research on warning design in the ongoing fight against misinformation.

I. INTRODUCTION

The rise of misinformation has been troubling academics and industry experts alike. Misinformation is associated with increasing political polarization and societal divisiveness and is considered by many as one of the most important problems currently threatening society [1]. Even though propaganda and misinformation have existed long before the Internet, online technologies have accelerated and exacerbated the spread of such content [2].

While the vast majority of prior academic research and industry interventions have focused on text-based misinformation, increasing concern is being raised about broader modes of fake media, including video, images and audio, that have either been subtly manipulated or generated entirely from scratch. Fake media broadly encompasses deep fakes, shallow fakes, and classic signal processing methods. Deep fakes are generated using deep learning algorithms [3], while the more prevalent shallow fakes rely on much simpler approaches such as photoshopping images or dropping video frames [4]. Signal processing methods can also be used to create new or alter existing media and can either used alone or in conjunction with deep fake algorithms [5] to create fake media. Such fake video media has been shown to be both extremely convincing and very difficult for people to detect in the wild, far more so than text-based media [6].

Four primary methods have been proposed to help identify and prevent users from falling for fake video and image content [7]:

• *Education*: Programs which focus on educating users about the problem of fake media and how to identify it [8].

- Manual Investigation: Careful training of media forensic analysts and journalists to identify and prevent the publication of fake media [9].
- Automated Detection: Algorithmic-based (e.g., deep learning methods) detection of fake media [6], [10]–[16].
- Provenance-Based Authentication: Systems that allow for cryptographically authenticating media content to its original source such as a newspaper or broadcaster [17]—[21].

While exciting progress is being made toward automatically detecting and removing deep fake media, such approaches are far from ready to be released at Internet-scale [6]. Thus, the human must remain in the loop to interpret signals regarding the authenticity of video media content. A key question remains: how should we go about communicating to users the authenticity of the media to which they are exposed? While prior work in industry has explored how to communicate the authenticity of images [22], for example by providing information about when and where the image was taken, no prior academic work to our knowledge has examined how to communicate the authenticity of *video* media to end-users.

In this work, we conduct a four-part, mixed-methods study as a first step toward understanding users' perceptions of fake media and developing user experience (UX) best practices to combat it. First, to gain context on people's existing mental models around different modes of fake media, we conduct a semi-structured interview study with 24 Americans from diverse sociodemographic backgrounds in which we probe participants concerns with media - including image and video media as well as text media - as well as their existing strategies for evaluating the truthfulness and authenticity of content. We find that users rely heavily on the source of video and image content in order to evaluate it, far more so than for text-based information. We thus focus on source-related misinformation indicators. Further, provenance systems – which can inform the user whether a piece of content has been transmitted by or derived from the original source – are the closest to being ready for widespread implementation and can provide concrete labels of media authenticity (vs. detection systems which provide a probability of authenticity, see Section II-A for more detail). Thus, we focus our remaining design steps on designing source authenticity (i.e., provenance) indicators for video media in order to align with users' existing mental models and best practice in industry. In the discussion, we review the ways in which our work can generalize to future work using automated

detection mechanisms.

Thus, second, situated in the context gained from our interview study, we conducted a series of participatory design studies with 19 Americans, again from diverse backgrounds, to develop a set of potential UX indicators for alerting users to the provenance of video content. Third, we refined these UX indicators through a series of expert reviews with five academic and journalistic experts from major news organizations and top academic institutions. Fourth, we quantitatively evaluate (n=1,456) a final set of proposed UX indicators (Figure 1 shows those that were found to be most effective), which were developed through participatory design and expert feedback.

In this work, we make the following contributions:

- Differences in Misinformation Detection. To our knowledge, we are the first to identify the differences in techniques that news consumers use to detect misinformation for different media types (text, image, video).
- 2) User-Centered Provenance Warning Design Guidelines. Due to the nuances related to provenance warnings for videos, users and experts heavily emphasized the need for brevity, specificity, and simplicity through their feedback and suggested designs.
- Over-generalization of Warnings. Our quantitative survey results suggest that users often overgeneralize the meaning of misinformation warnings.

Our results offer implications for future research on misinformation, suggest concrete designs and guidelines for warnings of video authenticity, and raise concerns around how users may overgeneralize in their interpretations of misinformation warnings.

II. RELATED WORK

In this section we briefly review prior work on detecting deep fakes, on users' perceptions of media information, and on the design of warnings for misinformation and related technology applications.

A. Deep Fakes and Detection

A face swap is one type of deep fake video where the face of someone in the original video is replaced with the face of a targeted person in the fake video using artificial intelligence [5], [12]. Although this technology is often used by social media users to create unique images and videos, this technology has also been used to create fake sexually explicit videos and has the potential to be used for other malicious activities. Due to this, researchers have begun to create deep fake video detection software that determines authenticity based on blinking [10], face warping artifacts [11], and other techniques [12]–[16]. However, the state-of-the-art fake video detection systems are at best able to detect when content is authentic - for example through provenance-based approaches [17]–[21] – rather than when content is manipulated. Thus, we focus our design studies on developing indicators of authenticity rather than manipulation, as described in Section IV.

B. User Perceptions of Media

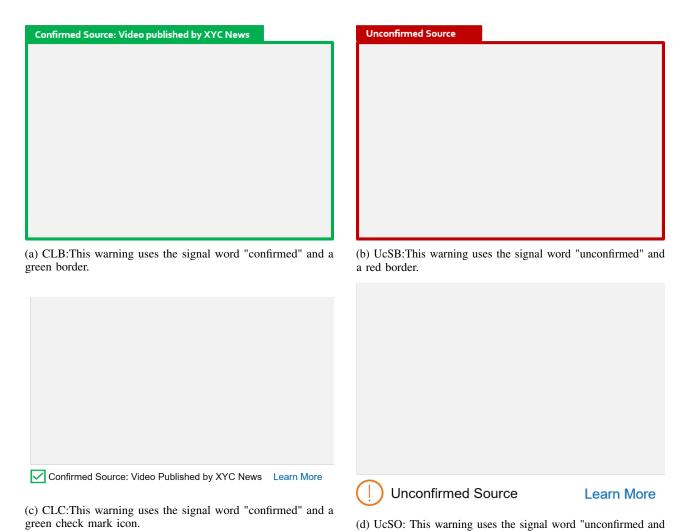
Prior work has found that people may use multiple heuristics to identify misinformation. Multiple studies have found evidence that the source of the information is one such heuristic. This heuristic includes the reader's perception of the trustworthiness of the source itself [23], [24], the appearance of the website on which the information is shown [25], and whether the information confirms or aligns with what the reader already believes [26], [27]. In addition to looking at the source's appearance, the bias in reporting, and alignment with current beliefs, readers also look to see if other sources report the same information or if the information is believed by others [28].

On social media, readers also have additional context that prior work finds they leverage to inform their media perceptions. Specifically, the Media Insight Project found that the credibility of information posted on social media is influenced by who posts the information [29]. This is corroborated by other studies that found that Twitter users rely on the profile picture, name, bio, and the expertise of the person that posted the content to determine the validity of content [30]–[32]. Also, similar to previous work on online news in general [26], Alcott et al. [33] found that people were more likely to believe "fake news" stories on social media that favored the candidate they liked during the 2016 election.

Prior work by Sundar et al. proposes a theoretical framework for understanding end-user perceptions of fake news consisting of four cues - Modality, Agency, Interactivity, and Navigability - that users rely on to determine the credibility of the information [34]. The first is Modality cues, where credibility is determined based on the structure of the medium providing the content. The second is Agency cues where credibility is based on the source of the information. The third is interactivity cues, where the actions one can take with the information is linked to credibility. The last is navigability where the user's ability to navigate the website of information impacts credibility.

Finally, a limited body of prior work has explored these questions with participants in other countries, largely finding similar results. Similar to the findings of Wagner et al. in the U.S. [23], Wasserman and Madrid-Morales [35], [36], who surveyed adults in Kenya and Nigeria, found that participants had a low level of trust in the media and believed they received a high amount of misinformation. Valenzuela et al. [37] surveyed residents in Chile and found that those who were politically engaged were more likely to share misinformation. Duffy et al. [38] interviewed Singaporeans and found that people did not always know when they received fake news, but were open to sharing information if it appeared to be something that would benefit others. On the other hand, the emergence of fake news has made them more cautious about sharing and believing everything they read. Tandoc [39] interviewed college students in Singapore who rated their Facebook friends as more credible than a news source. However, when motivation was high, the opposite behavior was exhibited and students rated the news source as more credible than their Facebook friends.

Thus far, research has shown that people's perceptions of the



an orange exclamation mark.

Fig. 1: The four most effective warning designs developed in this research. The full set of designs are shown in Appendix C in

trustworthiness of media are influenced by their existing beliefs and the source of the media – where source encompasses both elements of the news publisher, the UX of the news itself, and who shared the news. However, all of this prior work of end-user perceptions of media has either not explicitly addressed media mode (e.g., text vs. image vs. video) or have focused specifically on text media [32]. Given that participants may reasonably engage with different types of misinformation in different ways, and due to the growth of video misinformation especially, our interview study explicitly focuses on user experiences with different modes of misinformation. Additionally, prior work has focused exclusively on the truthfulness or trustworthiness of the content rather than provenance, e.g., whether the content itself was manipulated, despite the fact that provenance is

among the most technically feasible methods for detecting fake

news. Thus, in this work we explicitly explore provenance.

Figures 9, 10, and 11.

C. UX Indicators for Misinformation

Prior work in HCI and the social sciences has focused on methods for: actively dissuading end-users from engaging with misinformation typically using warnings [40]–[47] but occasionally also via games and other educational methods [48]; identifying which content should be fact-checked [42], [49]–[52]; and correcting inaccurate beliefs [53], [54]. We focus on the first body of related work as it is most relevant to the work presented here.

The majority of prior work suggests that fact-check warnings can lead to small but significant improvements in the accuracy of users beliefs [41]–[44], [47], [55], [56]. The literature finds that misinformation warnings are most effective when they are immediate [41], [45] (e.g., at the time of first exposure to the information)¹, presented in context (e.g., in close visual proximity to the content) [41], and are specific [43], [44], [56]

¹Earlier prior work argues that immediate correction may, however, increase the risk of backfire effects [46].

(e.g., indicate to the user exactly what is problematic about the content).

Fact-checking is an art of persuasion: fact-checks are typically conducted on politicized content and thus misinformation warnings or direct efforts to correct incorrect beliefs risk backfire effects in which the attempt to correct may lead to users' strengthening their belief in misinformation [46], [57]–[59].

While the vast majority of prior work on user perceptions of misinformation and credibility cues have focused on text content, a small body of prior work has focused on image media. The New York Times' News Provenance Project investigated ways to provide the context of photos and ensure the context remains with a photo as it is shared across the internet [22]. After interviewing 34 individuals with various political leanings about their media needs, their results suggested that users want the context in a simple way so they can form their own opinion. The researchers propose using blockchain technology to provide users with the image source, location, date published, and a short summary on what the photo captures.

No prior work, to our knowledge, has explicitly considered the design of indicators for provenance and especially provenance of video content. Our work takes a first step toward filling this gap.

D. UX for Warnings

The user experience literature also discusses the use of indicators to guide user behavior outside of the misinformation domain. For example, researchers in the usable security community have reviewed the use of warnings for browsers [60]–[70], apps [71]–[73], and physical spaces [74], [75]. Research-based guidelines for warning designs encourage removing visual clutter to improve warning detection, including consequences for non-compliance, use of a signal word, and clearly stating the hazard [76]. Investigations into the effectiveness of SSL warnings, which alert users when websites may not be as they appear or when their connection is unsecured, also found that hiding technical details [77] and providing clear default choices [65] reduce the chances that readers simply click through the warning without reading it [78], [79].

III. INTERVIEW STUDY: USER EXPERIENCES WITH DIFFERENT MODES OF MISINFORMATION

In June 2020, we conducted 24 semi-structured interviews with Americans from diverse sociodemographic backgrounds to understand their experiences with text, image, and video misinformation. These interviews provide context in which we situate our design studies (Section IV). As aforementioned, little prior work has investigated experiences with image and video misinformation and thus it was necessary for us to conduct our own contextual investigation.

In this section, we review our methods and the results of our interviews; limitations of all studies are discussed in Section VI. All methods were approved by our institution's ethics review board.

A. Methods

Our interviews were conducted via telephone or video conference due to COVID-19 restrictions. Interviews lasted for approximately 60 minutes, and all participants were compensated with \$20 Amazon gift cards for their participation.

1) Interview Protocol: Our interview protocol was organized into four sections. The first three sections covered participants' experiences with text, image, and video media, respectively. Each of these sections probed whether participants have ever had concerns about the truthfulness, accuracy or credibility of the text, image, or video information they receive. The interview probed what particular information they were most concerned about and what their concerns were, as well as how they went about assuaging those concerns (e.g., how they evaluated information) including asking the participant to walk through a concrete example of how they evaluated information, if they did so. The interview evaluated whether the participant had particular sources that they always or never trusted, their perceptions of information they saw on social media in particular, as well as whether they had concerns about non-news information (e.g., health or entertainment information). The interview also probed whether the participant had ever had any negative information experiences (e.g., falling for a piece of misinformation or manipulated media).

Finally, the fourth section of the interview asked whether there was anything that the participants' trusted sources could do that would lead to the participant no longer trusting them and whether the participants' approach towards text, video and/or image news had changed in the past 10 years.

We refined our interview protocol for clarity through a pilot session attended by two researchers in addition to the interviewer before proceeding with our interviews.

2) Recruitment: We aimed to recruit a demographically diverse set of participants whose demographics roughly matched those of the U.S. population. To do so we used two methods: we quota sampled 300 responses to a recruitment screening survey that collected respondents' demographics through the survey firm Cint, ensuring that we received a survey sample that was representative of U.S. demographics on age, gender, education, income, race, and geographic region. Respondents were compensated according to their agreement with Cint. We then contacted survey respondents who indicated interest in our full interview on this survey, attempting to balance participant demographics as we scheduled and interviewed respondents.

Through this process, we found that younger respondents, those under 30 years of age, were less likely to respond to our requests to schedule an interview, despite initially indicating interest. Thus, to ensure a balanced interview population we posted an advertisement on Craigslist in our city of residence² using the same screening questionnaire to recruit interview participants and balance diversify our participant pool. See Table III in Appendix A for the final demographics of our participants.

²Craigslist only allows one advertisement for one study to be posted, and thus we could not recruit in multiple cities.

3) Analysis: We analyzed our interview findings through open-coding [80]. Three researchers examined a randomly selected 20% (5) of the interview transcripts to independently develop a codebook. The researchers then met to reach agreement on a final codebook. Two of the researchers then coded a newly selected 20% of the interview transcripts and achieved κ =.67. Given this "substantial" agreement [81], and in line with prior work [82], one researcher proceeded to code the remainder of the transcripts (including the original five used for codebook creation).

Consistent with our qualitative methods and best practice guidelines for qualitative research, we report our results purely qualitatively – describing whether most, many, some, or a few – participants reported a particular code [83], [84].

4) Quantitative followup: In order to supplement this qualitative lens and more broadly evaluate awareness of provenance issues (e.g., deep fakes, manipulated images), we conducted a brief follow up survey on Amazon Mechanical Turk (n=3,552)³ to ask a larger number of participants about: whether they had ever heard of a "deep fake" video (and if so, how they would define that term and whether they had seen one) and whether they had ever heard of (and if so, ever used) reverse image search. We also included a final question "Your answer to this question will not affect your compensation. Did you do an online search (such as using Google) to help answer the questions we just asked you?" See Section VI for a discussion of the use of Amazon Mechanical Turk (MTurk) workers. We analyzed our data through descriptive statistics and report our results below.

B. Results

In this section we report on the results of our interview study and follow up survey. These results provide (a) a context in which to situate the design of our indicators, and (b) the first, to our knowledge, comparative investigation of users' perceptions and experiences with different *modes* of misinformation: video vs. image vs. text.

1) The role of provenance in various media types: Most participants mentioned various fact-checking methodologies reported in prior work [25], [28] such as triangulation (i.e., checking whether multiple sources showed the same information), using their perception of the information source, and using their intuition regarding the validity of the content itself. However, our work is the first, to our knowledge, to identify that participants rely on a different set of methods for each media type. To evaluate text media, most participants used triangulation, source and intuition; while context, intuition and triangulation were used by those who had a method for image evaluation. While few participants had an established method for evaluating video media, those who did relied on triangulation and source. For example, the following statements

are from two participants when asked how they evaluate text and video, respectively:

"The first thing I do is look at the source, the URL and who the author is, that kind of thing. And if that checks out then I'll probably look on something like Reddit which is like a crowd source tool, which you can have people chime in and say, they're listening about it, and whether they think it's true or not. So I think the combination of looking at the source of the article. I think I can get a better picture of whether it's correct or not." -P15 (Male, 20s)

"Obviously I'll supplement [the video] with the actual articles or see if anyone else posted other videos. I can go around the Internet and see, "Oh, does anyone have a different angle of this, or does anyone have any other...? Maybe three people video recording it," or something along those lines. I definitely try to go back and see, is there something I can go do to get a better vantage point or different opinion or something like that." - P16 (Female, 40s)

Although all participants had a predefined process for evaluating text content, some did not have established heuristics for evaluating of images, and many did not have heuristics for evaluating video because they felt these forms of media were too difficult to evaluate. For example, one participant responded quickly with the following when asked how they fact-check text information.

"Google. I Google everything." -P9 (Male, 60s)

Then, this same participant responded with the following when asked about evaluating images.

"Yes. You can really tell if the photo's photoshopped if you really know there's no way in hell that's real. There's no way in hell that's real, that's photoshopped. I get a chuckle out of that." -P9 (Male, 60s)

However, they did not have a method for evaluating video.

"No. I'm not able to. The only way I could check would be to take somebody else's word. You understand what I'm saying? I can't check. If you're going to check something, you've got to physically check it. You. If you're checking on the accuracy of something and you're taking somebody else's word for it ... That's somebody else's word."-P9 (Male, 60s)

Our quantitative results suggest that some participants' struggle to identify heuristics for checking video or image content may be related to a lack of awareness of manipulated video media and standard methods for identifying manipulated image content such as reverse image search. We surveyed 3,552 crowd workers and found that 51% of respondents had heard of deep fake videos and only 39% had heard of reverse image search. When asked if they had searched for answers when completing our questions about these terms 34% of respondents who said they heard of deep fake videos admitted to looking up the term and 26% admitted to doing so regarding reverse

³Survey and all procedures discussed in this paper were IRB approved. Crowd workers were paid \$0.30 for the 2 minute survey task to ensure at least \$8/hr hourly wage. The participant demographics can be found in Table IV in Appendix A.

image search. Accounting for those who looked up the terms to answer our question, 33% of participants had heard of deep fake videos while only 30% had heard of reverse image search.

We also asked participants which media type they were most and least concerned about containing misinformation. Although there was no clear consensus for which type was most or least concerning for participants, the explanations were similar. Most participants were concerned about videos and images being edited or taken out of context and the difficulty associated with evaluating them. Additionally, participants viewed text as necessary for understanding the context and easy to fact-check. For example, participants stated the following:

"Probably [more concerned about misinformation in]the actual text because, like I said, you can't believe everything you see, right?" -P19 (Male, 40s)

"I feel like a lot of people don't necessarily want to read, but if they're scrolling through social media or something, they might be more likely to watch a video...I mean, you might still have false information [in text], but since videos and photos can be altered so much, I would say I'm less concerned about the text of a news article." -P17 (Female, 30s)

"Images are the ones I'm most concerned about, but they just tend to be more accurate than texts articles. I don't know. I think that text articles are easier to check, so even if they are inaccurate but I can find out, whereas images I don't know how often they are accurate or not, just because it's harder to check."-P15 (Male, 20s)

2) Use of Trusted Sources: Most participants referenced the credibility of a source to help determine if the information being presented was trustworthy. Most participants had a list of trusted sources that they often referenced. For example, when we asked a participant about their trusted news sources they stated the following:

"Yeah, Seattle Times is a good one. New York Times. What else? Washington Post is a good one. MSNBC, of course, CNN. Those are all ones that I really trust most of the time. I don't think I've ever had a fake story posted on there." -P19 (Male, 40s)

However, most participants were unable to identify news sources they *did not* trust, they were receptive to information from sources outside their trusted list, as long as they perceived the information as credible. For example, one participant stated the following:

"I don't think it's to a point of never [trusting a particular source]. I just read all types of different sources, and if I find something that is a little funny or interesting, or not believable, I just kind of look at other places see if they're the same, or if something is small, I just let it go. But it's not to a point of never [trusting somewhere]." -P13 (Female, 20s)

3) Social Media Posts: While participants mentioned getting a high volume of visual news information through social media, they were very skeptical about the information they received. Even though participants were willing to read and receive information from unfamiliar news sources, this openness did not extend to social media posts, unless those posts were sharing information from a source they already explicitly trusted. For example, three of our participants stated the following:

"I'm not very trusting [of] anything that's shared over social media. If a video is from a local news outlet that I trust, then I'm more keen or more trusting on that." -P23 (Male, 30s)

"[On] social media, [I'm] probably the most [concerned]...Because you don't know where people are pulling that stuff from." -P11 (Female, 40s)

"Yeah, because I mean, I appreciate [social media] for photographs coming from people who are at the scene without being employed by the [news or] anything like that, but then again you get the same question of whether the image is correct or not. Whereas you don't have that doubt with the news agency." -P15 (Male, 20s)

4) Misinformation Impact: Having experienced misinformation in various forms in the past, news consumers have adopted questioning the authenticity of text content, as found also in prior work [85], [86].

"I used to believe everything. Now it's the complete opposite. First, I don't believe it unless I verify...there's so much stuff that's circling around that you don't even know at this point. It's like you have to make your own judgment now. You can't trust anything you're seeing." -P8 (Male, 30s)

"Honestly, I [started checking after] the last election and all the information that came out about whether or not Russia had something to do with Trump winning the election...seeing news about that is what has made me start to really question things." -P17 (Female, 30s)

Our results suggest that many are applying this same scepticism to image and video media, but may not know how to investigate this content beyond looking for obvious cues. For example, some of our participants stated the following:

"I've always had videos in my life, videos that are online or on TV and things like that. I don't know if that's the case from before, but I mean, it's always been more of the sources that I would trust on TV, and less of a source I would trust anywhere else, just because it's easy to make a video and movie editing has always been a thing." -P15 (Male, 20s)

" I think I'm way more skeptical, just because technology has changed so much. 10 or 20 years ago, I would look at a picture and that's what I'm seeing, where now, I have to kind of stop and think, is this really what I'm seeing?"-P10 (Female, 40s)

"Yeah, I think it's pretty much the same as with articles or images or anything like that. I want to consider the source that the person is sharing from, before I know whether it's... This is going to sound rude, but even worth my time."-P13 (Female, 20s)

IV. DESIGN STUDY: USER-DRIVEN DEVELOPMENT OF VIDEO AUTHENTICITY INDICATORS

In July 2020, we conducted 19 participatory design sessions with Americans from diverse sociodemographic backgrounds to develop a set of potential UX indicators for communicating media provenance for video content (e.g., whether the video being shown to the user was actually from the claimed source). We focus specifically on provenance for two reasons. First, as described in more detail in Section II-A, the most promising technological innovations on combating deep fakes focus on provenance. These are the technologies beginning to be deployed by major news organizations [87], [88], and thus focusing on provenance has the most immediate impact. Second, as evidenced by our interview study results, end-users already rely heavily on the source of content, especially for image and video media. Thus, conducting research on provenance is in alignment with users' pre-existing mental models and practices.

Below, we describe our methods and the results of our design interviews. We also describe our follow up expert design reviews of the participatory design-produced indicators, and the results of those reviews. All methods were approved by our institution's ethics review board.

A. Methods

Our participatory design sessions were conducted via video conference due to COVID-19 restrictions. Design sessions lasted for approximately 60 minutes, and all participants were compensated with \$20 Amazon gift cards for their participation.

1) Participatory Design Protocol: Design sessions consisted of three parts: in the first, participants sketched notification designs while in the second part they gave feedback on other designs (see Figure 4, 5 and 6 in Appendix B). In the third, participants instructed the researcher on how to create their ideal design using PowerPoint. A subset of participant designs are shown in Figure 8 of Appendix B.

We started the design sessions by asking participants to "draw some messages that you think would be able to inform someone looking at a video about whether it really came from the original source / the source stated on the video". We conducted these drawing exercises in the context of a scenario, to improve concreteness and maximize the validity of the sketches: "Imagine you are reading an article on the Blue Bear Times website, and you see a video associated with the article that appears to have been created by Blue Bear Times. Please draw what you would expect to see, if our technology [was unable to verify/was able to verify] that the Blue Bear Times was the source of the video/ created that video." We had a similar set of scenarios to prompt participant sketching in the context of a video appearing in social media.

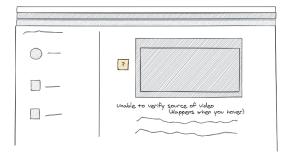


Fig. 2: This figure shows a screenshot of participant P7's unverified source warning design made using the Excalidraw template we provided. This warning includes a yellow box with a question mark to the left of the video. When the user hovers over the box, a message appears that says "Unable to verify source of video."

Participants sketched either within a drawing template that we provided (Figure 2) using Excalidraw [89] or told us what to sketch and watched and revised as we sketched it.⁴ After completing their sketches, participants evaluated a set of potential provenance notification designs that we created based on existing fake news notifications (see Appendix B for the designs). To avoid desirability bias (participants softening their feedback because they knew we created the designs), we told participants that we were not apart of the design process but were hired to get user feedback. For each design, we asked: "What do you immediately notice about this warning/label?", "What do you like about it?", "What do you dislike?", and "How can it be improved?" Finally, after participants viewed all designs, we reviewed which designs, design elements, and phrasings were their favorites among their own designs and the other designs they saw.

- 2) Recruitment: We used the same recruitment approach as described in Section III-A. See Table V for the final demographics of our design study participants.
- 3) Analysis: We chose to do participatory design studies to maximize inclusion of diverse perspectives in the design of our indicators [83], [90]. In order to avoid oppressing these perspectives [91] or reducing our sensitivity to nuances in the participatory design data by too restrictive of a formal coding process, we followed the best practice guidelines of McDonald et al. [83] . To this end, we used a field notes approach to analyze our results and reached consensus on our findings through researcher discussion of the results rather than open-coding with intercoder agreement.
- 4) Expert Design Reviews: We also conducted five expert design reviews with experts who had major mainstream newspaper and/or academic journalism backgrounds. We met with each expert for 30 minutes and solicited their feedback on the set of designs most preferred by our participants. Each expert was compensated with a \$50 Amazon gift card.

⁴Due to COVID19 restrictions we were unable to conduct in-person design studies.

All reviews were attended by three members of the research team, and the research team discussed the results of the session, reaching consensus on the feedback. In the results section below, we qualitatively describe the feedback provided by these experts.

B. Results

- 1) Participant Designs.: Overall, participant's designs consisted of the colors red or yellow for notices about unauthenticated media and green for notices indicating authenticated media (see Appendix B). Participants used either the phrase (un)verified or (un)known source to indicate the authentication status of the media. Participants used either a check mark or green block of color (for authenticated media) and an exclamation point or red block of color (for unauthenticated media). A few participants used interstitial warnings or messages imposed over the video. However, the majority of the participants' designs included a check mark or an exclamation point icon, as appropriate, and a phrase below the video that included the word verified when the video was authenticated and unverified or not verified when the video could not be authenticated. (see Figure 3 and Appendix B).
- 2) Design Feedback: Participants' discussions of their own designs, others designs, and the design probes, as well as experts' feedback on participants' designs can be summarized into four suggestions. We discuss each suggestion below and how it impacted the designs we chose for quantitative testing.

Include a simple and clear message. As shown in Figure 5 in Appendix B, some of the design probes included detailed information, such as date, location and subject, as suggested by prior work on image media [22]. Participants mentioned that while they understood why this information could be useful, they just wanted to read a simple phrase that was easy to understand and would not cause them to have to pause for more than a few seconds to read and react to the message, so as not to interrupt their flow of consuming content. Thus, they preferred that additional contextual information be placed deeper in the UX, such as behind a *Learn More* link. For example, one participant stated:

"I just think that [all the detailed information] would be a bit too much.... I like the learn more [instead of the additional information]... I'd rather just [know that] the video [is] verified or unverified." -P2 (Female, 30s)

"I feel like this [context about the video] almost takes away from the content itself... That's just a lot of information. I would prefer something that I could just quickly glance at to see if something is verified or not." -P15 (Female, 20s)

After receiving this feedback consistently from the participants, we decided to focus on testing warnings with short phrases that included the *Learn More* option. It is however important to note that participants may be unlikely to access the additional information, as shown by recent work examining participants' real practices when exploring news content on social media [32].

Additionally, while most participants used the word *verified* in their warning designs before and after seeing the design probes, perhaps because this is standard language used in social media for authenticated users, experts were concerned about the use of this word. When evaluating a participant design that contained the word *unverified*, Expert A stated:

"I don't know that this is the best word choice. It makes me want to ask... Well who says its unverified?... unconfirmed is a little more exact"-Expert A

We received similar feedback from the other expert reviewers in favor of the word *confirmed*, which prompted us to quantitatively test the word *confirmed* against the signal word *verified*, which experts worried would be over interpreted:

"My main concern is that people will read too much into it and might take it to mean too much more than what it actually means."-Expert C

Tell me why it is unauthenticated. Similarly, participants wanted to be able to find information specifically about why a video might be unverified or unconfirmed. While they did not want this information as part of the main warning, they wanted this information to be available if desired and suggested it be added to some of the designs that did not have it (see Figure 6b in Appendix B).

"Not being able to determine the source is different from we don't think this is a good source... I want to be clear about the intention...Something that is not a verifiable source is based on we don't know the source... Is there a learn more?... Ok, so if someone comes across this, and they'd want to know how come... how do you help answer that?" -P13 (Male, Over 60)

The expert reviewers agreed that having a *Learn More* button would be beneficial. They believed that participants might have questions about the authentication process and adding the *Learn More* button would allow viewers to understand the warning better.

Be specific. In addition to keeping warnings short and clear, but with additional context available through a *Learn More* button, participants and experts wanted to ensure that warnings were as specific as possible; supporting prior work showing that specific warnings are more effective [43], [44], [56].

A few of the designs we showed participants announced when a video source was confirmed but did not clearly state the source itself (see Figure 7b in Appendix B). Participants' responses to these design probes and some of their own designs (see Figure 3) suggest providing explicit information regarding the source of the video to make it clear to the reader what had been confirmed or verified.

"So it's saying the source of this video is known. Ok. Why are you telling me the source of this video is known? Like ,who is it? Like, why wouldn't you just say it? Like, put that in there with it." -P19 (Female, 30s)

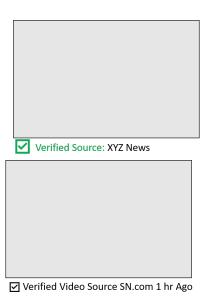


Fig. 3: These are screenshots of two warning designs created by participants during the study. Additional participant created designs can be found in Appendix B.

Our expert reviewers agreed with this sentiment and encouraged making the warnings highly specific. Instead of simply saying "Verified: XYC News", Expert B suggested using "Verified Source: This video was published by XYC". We implemented a shorter version of this statement using "Verified Source: Video published by XYC News" (see Figure 1a) and compared this against the even shorter "Verified: XYC News" in our quantitative tests presented in the next section.

Do not interrupt typical video viewing flow. Finally, participants felt strongly that the warnings should not interrupt their video viewing flow. This is in line with prior work finding that people want to be empowered to evaluate content, but given the autonomy to make their own choices about what to consume [22]. Some design probes showed passive warnings (e.g., a message below the video) while others showed interstitial warnings that required action before the viewer could watch the video. Although our participants understood why these design characteristics might be necessary, they preferred a warning that easily captures the viewer's attention but is easy to ignore when watching a video. Our participants did not want to be distracted by the warning when trying to watch their video of choice. In response to Figure 4d (Appendix B), a participant stated the following:

"Most of these [warnings with colored borders] are drawing way too much attention to themselves. This brings me back to the ideas that I drew which were unobtrusive [see Figure 2]" -P7 (Male, 20s)

Relatedly, participants wanted to avoid requiring additional clicks to watch a video after the warning is displayed. Our participants stated that the goal of the warnings was to notify the user of the authenticity of the media (that it was unmodified from the news source) and not to notify the user of anything about the information itself. Since the information presented

could still be accurate, even though the source was not authenticated, there is no reason to prevent users form watching the video. For example, when we showed the design displayed in Figure 6d (Appendix B), P5 immediately responded by saying:

"Why are you making the customer click multiple times?... I clicked on a video, then I read this, and then I have to click see video... I feel annoyed because there will be multiple time that I have to click." -P5 (Male, 40s)

For these reasons, and since only two participants drew interstitial warnings themselves, we do not evaluate interstitial warnings in our expert evaluation. It is important to note that while interstitial warnings may annoy users, they are known from prior work in other domains to be more effective than passive warnings [65]. However, thus far, few media platforms have adopted such warnings (Twitter is a notable exception).

Similarly, participants gave mixed reviews to the non-interstitial, but the highly visible design probe shown in Figure 4d (Appendix B). Participants appreciated that the design was easy to see, but found the size and text used in the warning disruptive. Experts, however, liked this design probe and suggested decreasing the thickness of the border and changing the text used (see Figure 1a) to accommodate participants' preferences (see Figure 1a).

V. QUANTITATIVE EVALUATION: EVALUATING INTERPRETABILITY OF VIDEO AUTHENTICITY INDICATORS

To evaluate the interpretability of the warnings created through our participatory design and expert evaluation studies, as well as to identify the most understandable set of warnings and/or warning design principles for video media provenance, we conducted a quantitative evaluation of the interpretability of a representative set of video provenance warning designs (n=1,456). The purpose of this work is to confirm the *inter*pretability of our warnings outside of our participatory design sample. However, the evaluation presented here is but a first step: just because users can understand the warnings does not necessarily mean that they will be effective at reducing the rate at which news consumers are misled by fake videos. Future work must be conducted to evaluate these warnings in the field. In this section we review the methods and results from our evaluation of these designs. All methods were approved by our institution's ethics review board.

A. Methods

We quantitatively evaluate a representative set of the designs most preferred by both participants and experts. Specifically, as most participant designs relied heavily on check mark / exclamation point icons, and phrases underneath the video frame, we test warnings that include these elements. As part of these designs, we test the participant-preferred phrase (un)verified against the expert-preferred phrase (un)confirmed. Additionally, since experts worried that such designs would be insufficient to draw participant's attention and thus suggested using a more visible warning design with a border (see Figure

TABLE I: Representative Design Combinations of Icon, Signal Word, and Message Length. * indicates designs ultimately tested in our quantitative evaluation (see Section V-A for more detail).

Design	Signal Word	Message Length	Icon	Phrase Used			
CSC*	Confirmed	Short	green √	Confirmed Source: XYC News			
CLC*	Confirmed	Long	green √	Confirmed Source: Video published by XYC News			
CLB*	Confirmed	Long	green border	Confirmed Source: Video published by XYC News			
UcSO*	Unconfirmed	Short	orange!	Unconfirmed Source			
UcSB*	Unconfirmed	Short	red border	Unconfirmed Source			
VSC	Verified	Short	green √	Verified Source: XYC News			
VSB	Verified	Short	green border	Verified Source: XYC News			
VLC*	Verified	Long	green √	Verified Source: Video published by XYC News			
UvSO*	Unverified	Short	orange!	Unverified Source			
UvSB	Unverified	Short	red border	Unverified Source			
Possible	Possible Survey Pairs: CSC-UcSO			CLB-UcSB VSC-UvSO VSB-UvSB VLC-UvSO			

TABLE II: Quantitative Survey Results. All possible representative design variants are shown for symmetry with Table I but those with – were not tested (see Section V-A)

	Designs										
Selection	CSC (N=287)	CLC (N=285)	CLB (N=299)	VSC	VSB	VLC (N=285)	UcSO (N=285)	UcSB (N=299)	UvSO (N=285)	UvSB	FB (N=300)
1	26%	25%	30%	_	_	29%	34%	36%	34%	_	31%
2	49%*	64%*	57%*	_	_	55%*	64%*	57%*	61%*	_	47%
3	33%	38%	34%	_	_	38%	39%	33%	38%	_	62%*
4	53%	48%	36%	_	_	48%	36%	33%	34%	_	33%
5	17%	13%	16%	_	_	18%	13%	15%	19%	_	17%
6	2%	1%	1%	_	_	1%	1%	2%	1%	_	2%
7	4%	4%	4%	_	_	1%	2%	4%	2%	_	1%
OI	16%	22%	24%	_	_	15%	24%	23%	22%	_	23%

^{*} = Intended answer

1a), we evaluate designs that contain borders against those that use the participant-preferred check mark / exclamation point icons. Finally, we also evaluate one state-of-the-art misinformation warning message, used by Facebook [92] (see Figure 12 in Appendix B), as a baseline for interpreting our results.⁵

Each design we test is comprised of a signal word, icon, and a phrase as shown in Table I. To avoid excessive use of resources, we do not evaluate every single possible combination of signal word, icon and phrase shown in Table I. Instead, we evaluate design attributes sequentially. We first compare signal words in designs that were otherwise identical (we used long,

green icon warnings for this comparison as they are most reflective of participant designs); next, we compared length for the most effective signal word; and finally we compared icons for the most effective signal words and lengths. See Appendix C for the designs ultimately tested and Table II for the results of these tests.

1) Survey Methodology: In August 2020, we tested these designs by distributing five surveys using MTurk ⁶. Four of the surveys tested a pair of designs: one that indicated a positively authenticated media instance and one that indicated a non-authenticated media instance; the baseline survey tested only Facebook's warning, which provides no notification for "true" content, only a warning for potentially false content. The surveys included a digital consent form, design pair interpretation questions, and demographic questions. The

^{1 =} The location and date of the video was confirmed/verified

^{2 =} The news agency that produced the video was unconfirmed/verified

^{3 =} This video provides (in) accurate information

^{4 =} XYC News is (not) a trustworthy source

^{5 =} It is recommended that I (do not) watch this video

^{6 =} Other

^{7 =} I don't know

OI = Participants that only selected the intended interpretation

⁵We use Facebook's 2017 warning design as a baseline because its goal – to raise awareness of potentially suspicious information – is closest to the goal of our work. Facebook's current warning design suggests related articles for user education, rather than aiming to provide users with signal about the present content [93].

⁶See Section VI for a discussion of the use of MTurk.

warnings in each pair were displayed randomly in the surveys to eliminate order effects. The design interpretation questions were multiple selection and asked participants to select the options that best fit their interpretation of each warning they were shown (see Table II for interpretation options). Based on a power analysis, we aimed to recruit at least 250 participants per survey, in order to observe small- to medium-sized differences in the interpretability of the designs.

Each survey took between two and three minutes to complete and participants were compensated \$.40 to complete each one. In order to participate, participants had to be in the United States of America, have a 95% or higher approval rating [94], and could only complete one of the multiple surveys we posted.

- 2) Participants: In total, 1,456 people participated in this study. Additional demographic information can be found in Table VI in Appendix C.
- 3) Analysis: Using a Chi-square test for homogeneity, we compared the proportion of participants that made a particular selection for a design to another design. As comparisons were limited and planned, we did not apply correction [95]. All calculations were completed using R, an open-source software for statistical computing [96].

B. Results

The purpose of each design is to notify users that the video source was authenticated or that it was not. Overall, we find that when the message informed the user that the video source was authenticated, a larger proportion of participants were able to correctly interpret the message as intended when the message included the signal word "confirmed", a longer more specific phrase describing what was confirmed (the source of the video), and a green checkmark (CLC 64%) or green border (CLB 57%) (see Table II). When the source was not authenticated, more participants responded correctly to the signal word unconfirmed than unverified and responded correctly equally as often when this signal word was paired with an exclamation point icon (UcSO 64%) or a red border (UcSB 57%).

1) Effect of Individual Design Choices: When participants were asked to interpret the meaning of the warning design they viewed, they were asked to select all of the options that applied. We tested the significant proportional differences between designs to see how changes in design impact interpretation. In order to do this, we compared warnings that differed in one design choice - phrase length, signal word, and icon. Below, we report on the effect of each design choice.

We first evaluated the effect of **signal word** (e.g., verified versus confirmed) on warnings that were otherwise identical. Comparing VLC to CLC, we find that a larger proportion of participants selected the intended interpretation (p = 0.03) when shown the warning with the "confirmed" phrase (CLC, 64%) than with the "verified" phrase (VLC, 55%).

When examining whether users *only* select the intended interpretation of the warning, we observe that only 22% of respondents selected only the intended interpretation for CLC and 15% did so for VLC (a significant difference with p = 0.031). This suggests that such warnings are highly likely to

be overgeneralized (see Section V-B2 for further discussion); however, the "confirmed" phrasing still outperforms the "verified" phrasing, confirming the concerns our experts raised about overgeneralization from the "verified" term.

Next, we evaluated the impact of **phrase length** using the "confirmed" keyword. When comparing CSC vs CLC, the longer phrase resulted in an increase in the proportion of participants that selected the intended interpretation (49% with CSC vs. 64% with CLC, p = 0.001). Phrase length did not have a significant impact on the degree of overgeneralization (e.g., the proportion of respondents who selected only the intended interpretation).

Finally, we compared the use of **icons** to colored borders by comparing CLC to CLB and UcSO to UcSB. While we observe no significant difference in the proportion of respondents correctly interpreting CLC and CLB, we do observe that the use of a border leads to better specificity in warning interpretation: a significantly higher proportion of participants (p = 0.005) interpreted the warning CLC warning to also mean that XYC news, the confirmed producer of the video, was a trustworthy source. Thus, it appears that the border design may better focus participants' interpretations to the specific content – the video – highlighted by the border. We find no significant differences in users' interpretations of the unconfirmed warnings with a red border vs. with an exclamation point.

2) Performance Against the Baseline & Overgeneralization: To evaluate our warnings against the current state-of-the-art misinformation warnings for text content, we evaluate the interpretability of our best warning designs against Facebook's previously used misinformation warning (Table II). We find that our warnings equally as interpretable as the existing state-of-the-art (64% of respondents selecting intended interpretation for UcSO vs. 62% for FB p > 0.05; 57% for UcSB vs. 62% for FB p > 0.05). Further, we find that participants did not overgeneralize the meaning of the user-created warnings significantly different than they do Facebook warnings (24% of respondents select only the intended interpretation for UcSO vs. 23% for FB, p > 0.05; 23% UcSB vs. 23% for FB, p > 0.05).

VI. LIMITATIONS

The qualitative portions of our work (Sections III and IV) have limitations common to qualitative research. First, the generalizability of our results is limited by the constraint of the qualitative sample size. We took care to recruit a sociodemographically diverse sample to increase the diversity of experiences captured in our work, but it is possible that we may have omitted an important dimension of diversity in our recruitment and thus not included a full spectrum of participant experiences. To partially counteract this risk, we did conduct follow up quantitative studies both to determine the prevalence of certain key interview results (e.g., Americans' awareness of deep fakes) and to validate that the designs produced through our qualitative work were understandable to larger groups of participants. Second, interviews rely on participants' comfort with sharing information with the interviewer and their ability

answer to the interviewer's questions. It is possible that despite our best efforts to follow best practice in our interviews and participatory design sessions, participants did not fully share their experiences in response to our questions.

The quantitative portions of our work are limited by the use of MTurk for recruiting our survey participants. Further, although we added questions to detect bots and limited participants to one of the seven surveys, we recognize that there are ways to circumvent these measures, and our results could include response from bots or repeat survey participants. Although Amazon Mechanical Turk (MTurk) workers are different from the general population in many ways, prior work has found that, for security and privacy related issues, MTurk workers are as representative of the experiences of Americans under 50 as opt-in survey panels with census-quota determined demographics [97]. Given this context, and that we took great care to obtain diverse perspectives in our qualitative work, we proceeded with using MTurk samples for our quantitative confirmations. However, our results should be interpreted with the appropriate sample limitations in mind.

Finally, as aforementioned, our quantitative evaluation evaluates the interpretability of our warnings but does not evaluate the impact of our warnings on user behavior. Future work should seek to evaluate video provenance warnings in the wild, to understand their mitigation effect for fake video media.

VII. DISCUSSION

Our interview results confirm the findings of prior work [25], [28] that users have multiple heuristics for combating misinformation that have been developed over time and exposure to such content. However, while our participants appeared confident and were quick to discuss fact-checking methods for text-based news articles, many were less sure of themselves and lacked heuristics for detecting misinformation depicted in images and videos. In sum, our interview results suggest that there are significant differences in how people experience and respond to text-, image-, and video-based misinformation. Thus, future work on misinformation and its warnings should explicitly consider the media's type.

More specifically, some participants simply expressed that they did not have a method for evaluating image or video content, while others mentioned that "that they would just know" if a video or image had been edited. However, as video editing techniques improve, it will increasingly become more difficult for the untrained eye to detect edited or deep fakes videos, especially when they are presented as real from a trusted source. Due to this, participants expressed concern about not having the necessary context to evaluate videos and images; with some participants even assuming all videos and images are fake unless there is relevant text to provide the context. Our design work takes a first step toward filling this gap, exploring best design practices for the creation of warnings that can provide the user with additional context regarding the authenticity of video media.

In line with prior work [25], [26], [30], we find that source is a particularly critical heuristic for identifying misinformation,

regardless of media mode. Thus, providing warnings that leverage users' existing emphasis on source by providing information on media provenance – whether the media they are seeing is actually unmodified content from the claimed source – can be the first step in the right direction for combating misinformation, especially for video content.

Such warnings must balance avoiding misleading users and reducing user engagement with content by frustrating them with warnings. Our design study results suggest that such warnings should avoid interrupting users' consumption flow – for example, by being brief, easy to interpret, and passive rather than interstitial – and should be highly specific (as also suggested by prior work [56], [65], [76]).

Our quantitative evaluation finds that the most interpretable video provenance warnings are those that use the word *confirm*, along with specific details (such as what the confirmed source of the video is). We found that either a red exclamation point or a red border for unconfirmed videos and a green check mark (for confirmed videos) or a green border around the video were equally interpretable, however the use of the green check mark may be overgeneralized as indicating that the news source is trustworthy as well.

Best practices for future misinformation design work. We leveraged both participatory design with end-users and expert feedback to develop our designs. We find that while participants' designs reflected one of the three most effective design elements – longer and more specific messages, it was the experts who suggested the other two effective elements: the use of a more specific signal word (confirmed vs. verified) and the border design; although the ultimate design of the latter was informed by participants' feedback as well. Thus, our results suggest that future work seeking to design such warnings should conduct *both* participatory design with endusers and expert evaluations to develop the most effective warnings.

Applicability of designs beyond video provenance. In this work, we focus on designs which indicate the authenticity of media generated by a provenance system. When media content can be cryptographically authenticated back to its publisher, these provenance indicators provide a secure guarantee that the media is authentic. On the other hand, automated fake media detection systems produce a score, and possibly a confidence level, that can be used to indicate the probability that the media is "fake". Our study results can also be applied to fake media detection to some extent, but additional research is required to understand the best method to convey to users the notion of probabilistic detection of manipulated media – rather than the binary and definitive notion of authenticated vs. non-authenticated addressed in this work.

Watch out for misinformation warning overgeneralization. Finally, our findings highlight a broad concern for misinformation warnings that address any mode of media: overgeneralization. Both our misinformation warnings and the baseline state-of-the-art warning from Facebook, against which we evaluated our warnings, exhibited high degrees of overgeneralization: while a majority of respondents selected

the intended interpretation of the warning, among other interpretations, just under a quarter selected only the intended warning interpretation. In addition to the intended interpretation of our warnings, users also took the warnings to mean that they should not watch the video because, for example, the content was inaccurate or the source was untrustworthy.

Thus, future work may seek to critically examine the consequences of overgeneralization of misinformation warnings. Evaluations of such warnings should take care to evaluate this property, and warning designers should carefully consider the impact they want to have with the warnings they provide given users tendency to generalize past the intended effect of the warning.

VIII. CONCLUSION

In this work, we conducted a four-part, mixed methods study to investigate Americans' experiences with video misinformation and design warnings to counteract such content. We conducted an interview study with 24 diverse Americans, identifying that the availability of heuristics through which users can identify misinformation and their confidence in applying those heuristics varies by media mode (text vs. image vs. video) and identifying that trust in the source of the information is a unifying misinformation heuristic across media modes.

Building on users' existing use of information source and the emerging viability of technical systems for verifying the provenance (authenticity) of video source as a promising way of combating fake media content, we develop warnings and design guidelines for communicating the provenance of video content to end-users. To this end, we conducted participatory design studies with 19 diverse Americans to develop and qualitatively evaluate designs for warnings to alert users to the authenticity of video media content. We further evaluated these warnings through expert design reviews and larger-scale quantitative evaluations with crowd workers. We find that a combination of user and expert insights are necessary to define the most interpretable warnings, and raise concerns around the potential for users to overgeneralize misinformation warnings regarding video or even text information. Our results offer concrete insights for misinformation warning designers and suggest multiple directions for future research.

REFERENCES

- R. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," 107 California Law Review 1753, 2019.
- [2] J. Anderson and L. Rainie, "The future of truth and misinformation online," 2017.
- [3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, and A. C. Y. Bengio, "Generative adversarial nets," in Conference on Neural Information Processing Systems (NIPS), 2014.
- [4] S. Mervosh, "Distorted videos of nancy pelosi spread on facebook and twitter, helped by trump," https://www.nytimes.com/2019/05/24/ us/politics/pelosi-doctored-video.html, 2019.
- [5] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2face: Real-time face capture and reenactment of rgb videos," in Proc. Computer Vision and Pattern Recognition (CVPR), IEEE, 2016.
- [6] B. Dolhansky, J. Bitton, B. Pflaum, J. Lu, R. Howes, M. Wang, and C. C. Ferrer, "The deepfake detection challenge dataset," 2020.

- [7] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *ICCV* 2019, 2019.
- [8] P. Barnwell, "Are deepfake videos a threat to democracy?" https://www.commonsense.org/education/articles/are-deepfake-videos-a-threat-to-democracy, 2020.
- [9] F. Marconi and T. Daldrup, "How The Wall Street Journal is preparing its journalists to detect deepfakes," 2018, http://www.niemanlab.org/2018/11/ how-the-wall-street-journal-is-preparing-its-journalists-to-detect-deepfakes/.
- [10] Y. Li, M. Chang, and S. Lyu, "In ictu oculi: Exposing ai created fake videos by detecting eye blinking," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.
- [11] Y. Li and S. Lyu, "Exposing deepfake videos by detecting face warping artifacts," arXiv preprint arXiv:1811.00656, 2018.
- [12] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, "Mesonet: a compact facial video forgery detection network," in 2018 IEEE International Workshop on Information Forensics and Security (WIFS), 2018, pp. 1–7.
- [13] D. Güera and E. J. Delp, "Deepfake video detection using recurrent neural networks," in 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). IEEE, 2018, pp. 1–6.
- [14] A. da Silva Pinto, H. Pedrini, W. Schwartz, and A. Rocha, "Video-based face spoofing detection through visual rhythm analysis," in 2012 25th SIBGRAPI Conference on Graphics, Patterns and Images. IEEE, 2012, pp. 221–228.
- [15] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics: A large-scale video dataset for forgery detection in human faces," arXiv preprint arXiv:1803.09179, 2018.
- [16] L.-B. Zhang, F. Peng, L. Qin, and M. Long, "Face spoofing detection based on color texture markov feature and support vector machine recursive feature elimination," *Journal of Visual Communication and Image Representation*, vol. 51, pp. 56–69, 2018.
- [17] Adobe, "Introducing the content authenticity initiative," https://theblog. adobe.com/content-authenticity-initiative/, 2019.
- [18] Amber, "Instilling trust into video," https://app.ambervideo.co/, 2019.
- [19] P. England, H. S. Malvar, E. Horvitz, J. W. Stokes, C. Fournet, R. Burke-Aguero, A. Chamayou, S. Clebsch, M. Costa, J. Deutscher, S. Erfani, M. Gaylor, A. Jenks, K. Kane, E. Redmiles, A. Shamis, I. Sharma, S. Wenker, and A. Zaman, "Amp: Authentication of media via provenance," 2020.
- [20] D. Kayyali, "Proofmode verified visuals enabled!" https://blog.witness. org/2017/04/proofmode-helping-prove-human-rights-abuses-world/, 2017.
- [21] L. H. Newman, "A new tool protects videos from deepfakes and tampering," https://www.wired.com/story/ amber-authenticate-video-validation-blockchain-tampering-deepfakes/, 2019.
- [22] S. Koren, "Introducing the news provenance project," Times Open, July, vol. 23, 2019.
- [23] M. C. Wagner and P. J. Boczkowski, "The reception of fake news: The interpretations and practices that shape the consumption of perceived misinformation," *Digital Journalism*, vol. 7, no. 7, pp. 870–885, 2019.
- [24] M. Flintham, C. Karner, K. Bachour, H. Creswick, N. Gupta, and S. Moran, "Falling for fake news: investigating the consumption of news via social media," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–10.
- [25] B. J. Fogg, C. Soohoo, D. R. Danielson, L. Marable, J. Stanford, and E. R. Tauber, "How do users evaluate the credibility of web sites? a study with over 2,500 participants," in *Proceedings of the 2003 conference on Designing for user experiences*, 2003, pp. 1–15.
- [26] M. J. Metzger, A. J. Flanagin, and R. B. Medders, "Social and heuristic approaches to credibility evaluation online," *Journal of communication*, vol. 60, no. 3, pp. 413–439, 2010.
- [27] A. Guess, J. Nagler, and J. Tucker, "Less than you think: Prevalence and predictors of fake news dissemination on facebook," *Science advances*, vol. 5, no. 1, p. eaau4586, 2019.
- [28] S. Zaryan, "Truth and trust: How audiences are making sense of fake news." 2017.
- [29] D. Sterret, D. Malato, J. Benz, L. Kantor, T. Tompson, T. Rosenstiel, J. Sonderman, K. Loker, and E. Swanson, "Who shared it?: How americans decide what news to trust on social media," 2018.
- [30] J. Y. Lee and S. S. Sundar, "To tweet or to retweet? that is the question for health professionals on twitter," *Health Communication*, vol. 28, no. 5, pp. 509–524, 2013.

- [31] M. R. Morris, S. Counts, A. Roseway, A. Hoff, and J. Schwarz, "Tweeting is believing? understanding microblog credibility perceptions," in *Proceedings of the ACM 2012 conference on computer supported cooperative work*, 2012, pp. 441–450.
- [32] C. Geeng, S. Yee, and F. Roesner, "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*, 2020, pp. 1–14.
- [33] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of economic perspectives*, vol. 31, no. 2, pp. 211–36, 2017.
- [34] S. S. Sundar, The MAIN model: A heuristic approach to understanding technology effects on credibility. MacArthur Foundation Digital Media and Learning Initiative, 2008.
- [35] H. Wasserman and D. Madrid-Morales, "New data suggests african audiences see significantly more misinformation than americans do," 2018.
- [36] —, "An exploratory study of "fake news" and media trust in kenya, nigeria and south africa," *African Journalism Studies*, vol. 40, no. 1, pp. 107–123, 2019.
- [37] S. Valenzuela, D. Halpern, J. E. Katz, and J. P. Miranda, "The paradox of participation versus misinformation: Social media, political engagement, and the spread of misinformation," *Digital Journalism*, vol. 7, no. 6, pp. 802–823, 2019.
- [38] A. Duffy, E. Tandoc, and R. Ling, "Too good to be true, too good not to share: the social utility of fake news," *Information, Communication & Society*, pp. 1–15, 2019.
- [39] E. C. Tandoc Jr, "Tell me who your sources are: Perceptions of news credibility on social media," *Journalism Practice*, vol. 13, no. 2, pp. 178–190, 2019
- [40] N. Dias, G. Pennycook, and D. G. Rand, "Emphasizing publishers does not effectively reduce susceptibility to misinformation on social media," *Harvard Kennedy School Misinformation Review*, vol. 1, no. 1, 2020.
- [41] P. Ozturk, H. Li, and Y. Sakamoto, "Combating rumor spread on social media: The effectiveness of refutation and warning," in 2015 48th Hawaii International Conference on System Sciences. IEEE, 2015, pp. 2406– 2414.
- [42] D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild *et al.*, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018
- [43] U. K. Ecker, S. Lewandowsky, and D. T. Tang, "Explicit warnings reduce but do not eliminate the continued influence of misinformation," *Memory & cognition*, vol. 38, no. 8, pp. 1087–1100, 2010.
- [44] K. Clayton, S. Blair, J. A. Busam, S. Forstner, J. Glance, G. Green, A. Kawata, A. Kovvuri, J. Martin, E. Morgan et al., "Real solutions for fake news? measuring the effectiveness of general warnings and factcheck tags in reducing belief in false stories on social media," *Political Behavior*, pp. 1–23, 2019.
- [45] S. Lewandowsky, U. K. Ecker, C. M. Seifert, N. Schwarz, and J. Cook, "Misinformation and its correction: Continued influence and successful debiasing," *Psychological science in the public interest*, vol. 13, no. 3, pp. 106–131, 2012.
- [46] R. K. Garrett and B. E. Weeks, "The promise and peril of real-time corrections to political misperceptions," in *Proceedings of the 2013* conference on Computer supported cooperative work, 2013, pp. 1047– 1058
- [47] W. Yaqub, O. Kakhidze, M. L. Brockman, N. Memon, and S. Patil, "Effects of credibility indicators on social media news sharing intent," in Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–14.
- [48] L. Grace and B. Hone, "Factitious: Large scale computer game to fight fake news and improve news literacy," in Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–8.
- [49] M. Babaei, A. Chakraborty, J. Kulshrestha, E. M. Redmiles, M. Cha, and K. P. Gummadi, "Analyzing biases in perception of truth in news stories and their implications for fact checking," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 139–139.
- [50] Y. Liu and Y.-F. B. Wu, "Early detection of fake news on social media through propagation path classification with recurrent and convolutional networks," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

- [51] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD explorations newsletter, vol. 19, no. 1, pp. 22–36, 2017.
- [52] A. X. Zhang, A. Ranganathan, S. E. Metz, S. Appling, C. M. Sehat, N. Gilmore, N. B. Adams, E. Vincent, J. Lee, M. Robbins *et al.*, "A structured response to misinformation: Defining and annotating credibility indicators in news articles," in *Companion Proceedings of the The Web Conference 2018*, 2018, pp. 603–612.
- [53] L. Bode and E. K. Vraga, "See something, say something: Correction of global health misinformation on social media," *Health communication*, vol. 33, no. 9, pp. 1131–1140, 2018.
- [54] E. Porter, T. J. Wood, and D. Kirby, "Sex trafficking, russian infiltration, birth certificates, and pedophilia: A survey experiment correcting fake news," *Journal of Experimental Political Science*, vol. 5, no. 2, pp. 159–164, 2018.
- [55] H. Seo, A. Xiong, and D. Lee, "Trust it or not: Effects of machine-learning warnings in helping individuals mitigate misinformation," in Proceedings of the 10th ACM Conference on Web Science, 2019, pp. 265–274
- [56] B. Ross, A. Jung, J. Heisel, and S. Stieglitz, "Fake news on social media: The (in) effectiveness of warning messages," 2018.
- [57] A. Guess and A. Coppock, "Does counter-attitudinal information cause backlash? results from three large survey experiments," *British Journal* of *Political Science*, pp. 1–19, 2018.
- [58] T. Wood and E. Porter, "The elusive backfire effect: Mass attitudes' steadfast factual adherence," *Political Behavior*, vol. 41, no. 1, pp. 135– 163, 2019.
- [59] B. Nyhan, E. Porter, J. Reifler, and T. J. Wood, "Taking fact-checks literally but not seriously? the effects of journalistic fact-checking on factual beliefs and candidate favorability," *Political Behavior*, pp. 1–22, 2019
- [60] A. P. Felt, R. W. Reeder, A. Ainslie, H. Harris, M. Walker, C. Thompson, M. E. Acer, E. Morant, and S. Consolvo, "Rethinking connection security indicators," in *Twelfth Symposium on Usable Privacy and Security* ({SOUPS} 2016), 2016, pp. 1–14.
- [61] J. Clark and P. C. Van Oorschot, "Sok: Ssl and https: Revisiting past challenges and evaluating certificate trust model enhancements," in 2013 IEEE Symposium on Security and Privacy. IEEE, 2013, pp. 511–525.
- [62] C. Thompson, M. Shelton, E. Stark, M. Walker, E. Schechter, and A. P. Felt, "The web's identity crisis: understanding the effectiveness of website identity indicators," in 28th {USENIX} Security Symposium ({USENIX} Security 19), 2019, pp. 1715–1732.
- [63] R. Roberts, Y. Goldschlag, R. Walter, T. Chung, A. Mislove, and D. Levin, "You are who you appear to be: a longitudinal study of domain impersonation in tls certificates," in *Proceedings of the 2019* ACM SIGSAC Conference on Computer and Communications Security, 2019, pp. 2489–2504.
- [64] Y. Wu, P. Gupta, M. Wei, Y. Acar, S. Fahl, and B. Ur, "Your secrets are safe: How browsers' explanations impact misconceptions about private browsing mode," in *Proceedings of the 2018 World Wide Web Conference*, 2018, pp. 217–226.
- [65] S. Egelman, L. F. Cranor, and J. Hong, "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, 2008, pp. 1065–1074.
- [66] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer, "The emperor's new security indicators," in 2007 IEEE Symposium on Security and Privacy (SP'07). IEEE, 2007, pp. 51–65.
- [67] M. Stojmenović and R. Biddle, "Who are they? website authentication: Certificates and identity," in Who are you? Adventures in Authentication Workshop, 2018.
- [68] D.-Y. Yu, E. Stobert, D. Basin, and S. Capkun, "Exploring website location as a security indicator," arXiv preprint arXiv:1610.03647, 2016.
- [69] R. W. Reeder, A. P. Felt, S. Consolvo, N. Malkin, C. Thompson, and S. Egelman, "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*, 2018, pp. 1–13.
- [70] E. M. Redmiles, E. Liu, and M. L. Mazurek, "You want me to do what? a design study of two-factor authentication messages." in SOUPS, 2017.
- [71] R. Abu-Salma, M. A. Sasse, J. Bonneau, A. Danilova, A. Naiakshina, and M. Smith, "Obstacles to the adoption of secure communication tools," in 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017, pp. 137–153.

- [72] V. Stavova, L. Dedkova, V. Matyas, M. Just, D. Smahel, and M. Ukrop, "Experimental large-scale review of attractors for detection of potentially unwanted applications," *Computers & Security*, vol. 76, pp. 92–100, 2018
- [73] I. N. Sherman, J. D. Bowers, K. McNamara Jr, J. E. Gilbert, J. Ruiz, and P. Traynor, "Are you going to answer that? measuring user responses to anti-robocall application indicators."
- [74] N. Scaife, J. Bowers, C. Peeters, G. Hernandez, I. N. Sherman, P. Traynor, and L. Anthony, "Kiss from a rogue: Evaluating detectability of pay-at-the-pump card skimmers," in 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019, pp. 1000–1014.
- [75] R. S. Portnoff, L. N. Lee, S. Egelman, P. Mishra, D. Leung, and D. Wagner, "Somebody's watching me? assessing the effectiveness of webcam indicator lights," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015, pp. 1649– 1658.
- [76] M. S. Wogalter, V. C. Conzola, and T. L. Smith-Jackson, "Research-based guidelines for warning design and evaluation," *Applied ergonomics*, vol. 33, no. 3, pp. 219–230, 2002.
- [77] C. Nodder, "Users and trust: A microsoft case study," Security and Usability, pp. 589–606, 2005.
- [78] A. P. Felt, R. W. Reeder, H. Almuhimedi, and S. Consolvo, "Experimenting at scale with google chrome's ssl warning," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2014, pp. 2667–2670.
- [79] D. Akhawe and A. P. Felt, "Alice in warningland: A large-scale field study of browser security warning effectiveness," in 22nd {USENIX} Security Symposium ({USENIX} Security 13), 2013, pp. 257–272.
- [80] P. Burnard, "A method of analysing interview transcripts in qualitative research," *Nurse education today*, vol. 11, no. 6, pp. 461–466, 1991.
- [81] J. R. Landis and G. G. Koch, "An application of hierarchical kappatype statistics in the assessment of majority agreement among multiple observers," *Biometrics*, pp. 363–374, 1977.
- [82] M. Lombard, J. Snyder-Duch, and C. C. Bracken, "Content analysis in mass communication: Assessment and reporting of intercoder reliability," *Human communication research*, vol. 28, no. 4, pp. 587–604, 2002.
- [83] N. McDonald, S. Schoenebeck, and A. Forte, "Reliability and inter-rater reliability in qualitative research: Norms and guidelines for cscw and hci practice," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–23, 2019.
- [84] M. Sandelowski, "Real qualitative researchers do not count: The use of numbers in qualitative research," *Research in nursing & health*, vol. 24, no. 3, pp. 230–240, 2001.
- [85] M. A. Amazeen, C. J. Vargo, and T. Hopp, "Reinforcing attitudes in a gatewatching news era: Individual-level antecedents to sharing factchecks on social media," *Communication Monographs*, vol. 86, no. 1, pp. 112–132, 2019.
- [86] S. Edgerly, R. R. Mourão, E. Thorson, and S. M. Tham, "When do audiences verify? how perceptions about message and source influence audience verification of news headlines," *Journalism & Mass Communication Quarterly*, vol. 97, no. 1, pp. 52–71, 2020.
- [87] J. Aythora, R. Burke-Agüero, A. Chamayou, S. Clebsch, M. Costa, J. Deutscher, N. Earnshaw, L. Ellis, P. England, C. Fournet, M. Gaylor, C. Halford, E. Horvitz, A. Jenks, K. Kane, M. Lavallee, S. Lowenstein, B. MacCormack, H. Malvar, S. O'Brien, J. Parnall, E. M. Redmiles, A. Shamis, I. Sharma, J. Stokes, S. Wenker, , and A. Zaman, "Multistakeholder media provenance management to counter synthetic media risks in news publishing," in *Proc. International Broadcasting Convention (IBC)*, 2020.
- [88] "Trusted news initiative (tni) steps up global fight against disinformation with new focus on us presidential election," BBC, 2020, (Accessed on 09/13/2020).
- [89] Excalidraw. (2020) Excalidraw. [Online]. Available: https://github.com/ excalidraw/excalidraw
- [90] A. Marwick, C. Fontaine, and D. Boyd, ""nobody sees it, nobody gets mad": Social media, privacy, and personal responsibility among low-ses youth," *Social Media+ Society*, vol. 3, no. 2, p. 2056305117710455, 2017.
- [91] R. Bellini, A. Strohmayer, E. Alabdulqader, A. A. Ahmed, K. Spiel, S. Bardzell, and M. Balaam, "Feminist hci: taking stock, moving forward, and engaging community," in *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–4.
- [92] E. Hunt, "Disputed by multiple fact-checkers': Facebook rolls out new alert to combat fake news," 2017.

- [93] T. Lyons, "Replacing disputed flags with related articles," 2017.
- [94] E. Peer, J. Vosgerau, and A. Acquisti, "Reputation as a sufficient condition for data quality on amazon mechanical turk," *Behavior research methods*, vol. 46, no. 4, pp. 1023–1031, 2014.
- [95] R. A. Armstrong, "When to use the bonferroni correction," Ophthalmic and Physiological Optics, vol. 34, no. 5, pp. 502–508, 2014.
- [96] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2013. [Online]. Available: http://www.R-project.org/
- [97] E. M. Redmiles, S. Kross, and M. L. Mazurek, "How well do my results generalize? comparing security and privacy survey results from mturk, web, and telephone samples," in 2019 IEEE Symposium on Security and Privacy (SP). IEEE, 2019, pp. 1326–1343.
- [98] S. Wang, "Slate's chrome extension helps identify fake news on facebook — and lets readers flag it themselves," 2016.
- [99] Y. Roth and N. Pickles, "Updating our approach to misleading information," 2020.

APPENDIX A INTERVIEW DEMOGRAPHICS

In this appendix, we describe the demographics of the participants of the two parts of our interview study. Table III provides the demographics of the interview participants. The demographics of the MTurk participants who answered questions regarding their knowledge of deep fake videos and reverse image search via a survey are displayed in Table IV.

TABLE III: Demographics for Interview Participants

Demographics	Participants		
(N=24)	n %		
Gender			
Female	13	54%	
Male	11	46%	
Non-binary	0	0%	
Prefer not to answer	0	0%	
Age			
18-30	4	17%	
31-40	9	38%	
41-50	5	21%	
Over 50	6	25%	
Prefer not to answer	0	0%	
Race			
White	12	50%	
Hispanic or Latino	1	4%	
Black or African American	3	13%	
Native American	1	4%	
Asian	5	21%	
Other	2	8%	
Prefer not to answer	0	0%	
Education			
No college	1	4%	
Some college but no degree	5	21%	
Associate's degree	5	21%	
Bachelor's degree	9	38%	
Advanced degree	4	17%	
Prefer not to answer	0	0%	
Employment			
Employed	11	46%	
Self-employed	4	17%	
Not working	5	21%	
A homemaker	3	13%	
A student	1	4%	
Prefer not to answer	0	0%	
Income			
Less than \$30,000	4	17%	
\$30,000 to under \$65,000	6	25%	
\$65,000 or more	13	54%	
	1	4%	

TABLE IV: Deep Fake Survey Demographics

Demographics	Participants		
(N=3,552)	n	%	
Gender			
Female	1488	42%	
Male	2048	58%	
Non-binary	10	0%	
Prefer not to answer	6	0%	
Age			
18-30	1229	35%	
31-40	1182	33%	
41-50	632	18%	
Over 50	498	14%	
Prefer not to answer	11	0%	
Race			
White	2447	69%	
Hispanic or Latino	160	5%	
Black or African American	706	20%	
Native American or American Indian	99	3%	
Asian	231	7%	
Native Hawaiian or Pacific Islander	5	0%	
Other	8	0%	
Prefer not to answer	11	0%	
Education			
No college experience	189	5%	
Some college but no degree	307	9%	
Associate's degree	228	6%	
Bachelor's degree	2153	61%	
Advanced degree	666	19%	
Prefer not to answer	9	0%	
Employment			
Employed	2622	74%	
Self-employed	528	15%	
Out of work	204	6%	
A homemaker	87	2%	
A student	66	2%	
Prefer not to answer	22	1%	
Income			
Less than \$30,000	791	22%	
\$30,000 to under \$40,000	433	12%	
\$40,000 to under \$50,000	598	17%	
\$50,000 to under \$80,000	1066	30%	
\$80,000 or more	664	19%	
Prefer not to answer	42	1%	
1 Telef flot to allswer	72	1 /0	

APPENDIX B PARTICIPATORY DESIGN PROBES

This appendix provides the design probes used in the participatory design study. Figures 4, 5, 6, and 7 depict the screenshots of these design probes. Figure 8 displays a subset of the notification designs created by participants. Finally, Table V displays the demographics of the participants from this study.

TABLE V: Demographics for Participatory Design Study

Native American Asian Asian Asian Asian Asian Asian Asian Asian Asian A Self-employed Not working Self-employed Self-employed Prefer not to answer Not working A student A homemaker Not working Self-employed Sel	Demographics	Participants		
Female 10 53% Male 9 47% Non-binary 0 0% Prefer not to answer 0 0% Age	(N=19)	n	%	
Male 9 47% Non-binary 0 0% Prefer not to answer 0 0% Age	Gender			
Non-binary 0 0%				
Prefer not to answer 0 0%				
Age		-		
18-30	Prefer not to answer	0	0%	
31-40	Age			
Al-50	18-30	5	26%	
No college	31-40	7	37%	
No college		1	5%	
No college	Over 50	6	32%	
Some college but no degree	Education			
Some college but no degree	No college	0	0%	
Associate's degree 2 11% 53% Advanced degree 3 16%		4	21%	
Bachelor's degree 10 53% Advanced degree 3 16% Race		2	11%	
White		10	53%	
White	Advanced degree	3	16%	
Hispanic or Latino	Race			
Black or African American	White	7	37%	
Native American	Hispanic or Latino	1	5%	
Asian Other 2 11%	Black or African American	5	26%	
Other Prefer not to answer 2 0 0% Employment 10 53% Self-employed Not working A student A homemaker 2 11% A student A homemaker 0 0%	Native American	1	5%	
Prefer not to answer 0 0%	Asian	4	21%	
Employment Employed Self-employed Not working A student A homemaker 10 53% 53% 53% 53% 53% 53% 53% 53% 53% 53%	Other	2	11%	
Employed 10 53% Self-employed 2 11% Not working 5 26% A student 2 11% A homemaker 0 0%	Prefer not to answer	0	0%	
Self-employed 2 11% Not working 5 26% A student 2 11% A homemaker 0 0%	Employment			
Not working 5 26% A student 2 11% A homemaker 0 0%	Employed	10	53%	
A student 2 11% A homemaker 0 0%	Self-employed	2	11%	
A homemaker 0 0%	Not working	5	26%	
	A student	2	11%	
Prefer not to answer 0 0%	A homemaker	0	0%	
	Prefer not to answer	0	0%	
Income	Income			
Less than \$30,000 3 16%	Less than \$30,000	3	16%	
\$30,000 to under \$65,000 7 37%		1	37%	
\$65,000 or more 9 47%	\$65,000 or more	9	47%	
Prefer not to answer 0 0%	Prefer not to answer	0	0%	

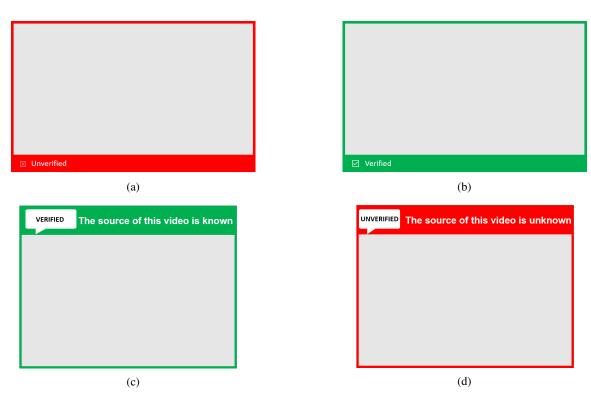


Fig. 4: Warning designs inspired by Slate's fake news detector browser extension [98]

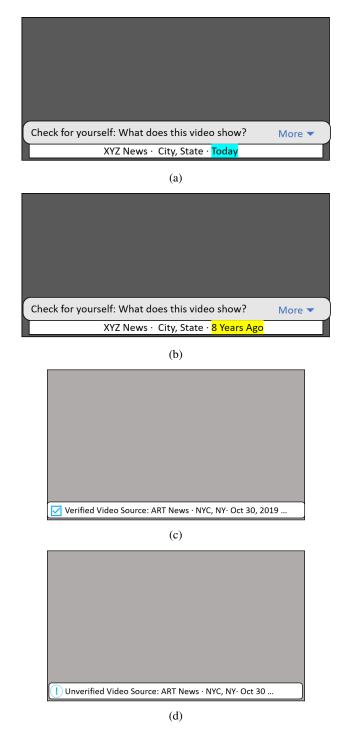


Fig. 5: Warning designs inspired by the News Provenance Project [22]

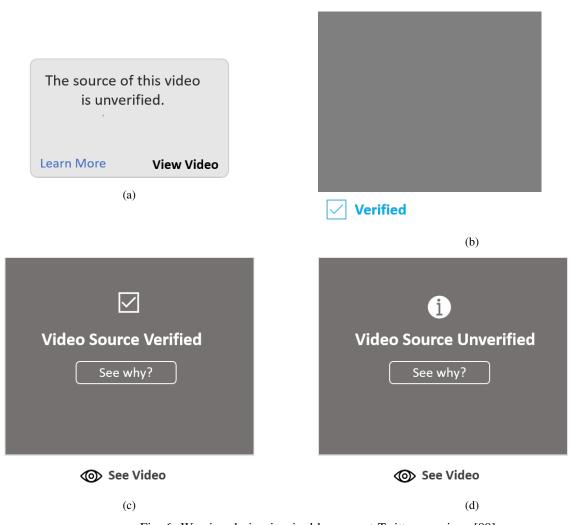


Fig. 6: Warning design inspired by current Twitter warnings [99]

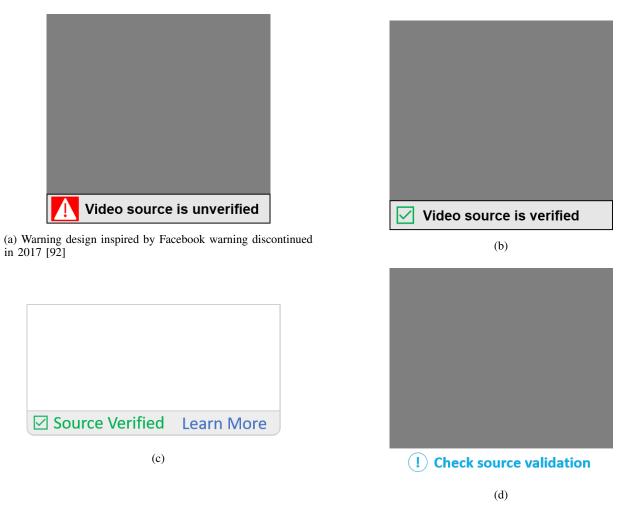


Fig. 7: This figure includes four warning design probes where the warning information is below the video

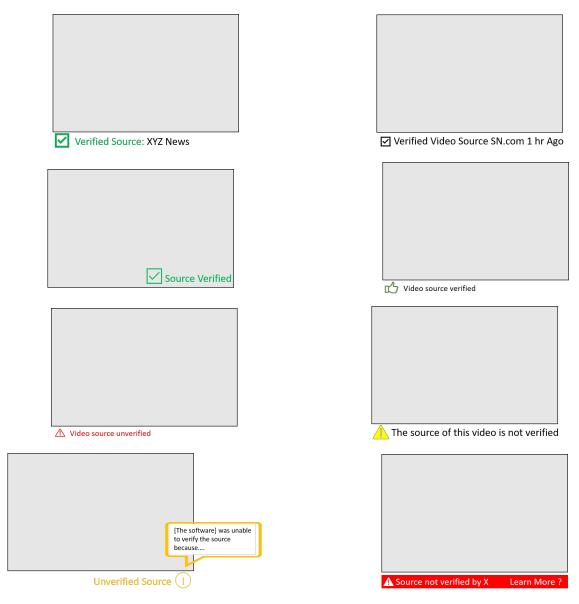


Fig. 8: Subset of Participant Warning Designs

APPENDIX C QUANTITATIVE STUDY

We next provide details of the quantitative study. Figure 12 9, 10, and 11 display screenshots of the designs that were created based on participant and expert feedback. Table VI indicates the demographics of the participants from the quantitative study where participants interpreted the designs.

TABLE VI: Quantitative Study Demographics

Demographics	Demographics Participa				
(N=1,456)	n	%			
Gender					
Female	588	40%			
Male	863	59%			
Non-binary	1	<1%			
Prefer not to answer	4	<1%			
Age					
18-30	540	37%			
31-40	481	33%			
41-50	243	17%			
Over 50	187	13%			
Prefer not to answer	5	<1%			
Race					
White	1049	72%			
Hispanic or Latino	58	4%			
Black or African American	243	17%			
Native American or American Indian	48	3%			
Asian	100	7%			
Native Hawaiian or Pacific Islander	3	<1%			
Other	5	<1%			
Prefer not to answer	6	<1%			
Education					
No college	90	6%			
Some college but no degree	118	8%			
Associate's degree	86	6%			
Bachelor's degree	908	62%			
Advanced degree	250	17%			
Prefer not to answer	4	<1%			
Employment					
Employed	1086	75%			
Self-employed	234	16%			
Not working	75	5%			
A homemaker	32	2%			
A student	22	2%			
Prefer not to answer	7	<1%			
Income					
Less than \$30,000	324	22%			
\$30,000 to under \$40,000	678	47%			
\$65,000 or More	445	31%			
		1%			

Confirmed Source: Video published by XYC News	Unconfirmed Source
(a) CLB:This warning uses the signal word "confirmed" and a	(b) UcSB:This warning uses the signal word "confirmed" and a
green border. Verified Source: XYC News	red border. Unverified Source
Verified Source: ATC News	Onvermed Source
() TI:	(I) II OD TII:
(c) vsc:This warning uses the signal word "verified" and a green border.	(d) UvSB:This warning uses the signal word "verified" and a red border.

Fig. 9: This figure shows the four designs we tested that have a border around the video instead of an icon.

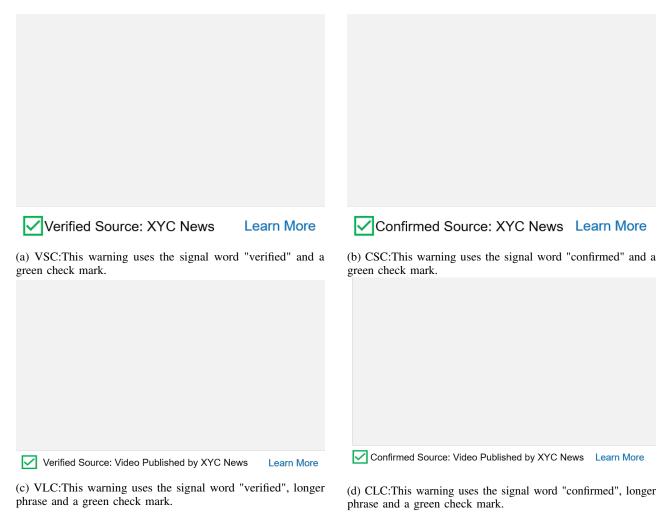


Fig. 10: This figure shows the four designs that included the green check mark in their design to alert users that the source was validated.



Fig. 11: This figure displays the two designs that use the exclamation point icon in the warning to alert users that source was not validated.



Fig. 12: Facebook Misinformation Warning from 2017