"I used to live in Florida": Exploring the Impact of Spam Call Warning Accuracy on Callee Decision-Making

Imani N. S. Munyaka
University of California, San Diego
drmunyaka@usd.edu

Daniel A. Delgado, Juan E. Gilbert, Jaime Ruiz
and Patrick Traynor
University of Florida
danieldel@ufl.edu, juan@ufl.edu, jaime.ruiz@ufl.edu, traynor@ufl.edu

***Abstract*—Telephone carriers and third-party developers have created technical solutions to detect and notify consumers of spam calls. The goal of this technology is to help users make decisions about incoming calls and reduce the negative effects of spam calls on finances and daily life. Although useful, this technology has varying accuracy due to technical limitations. In this study, we conduct design interviews, a call response diary study, and an MTurk survey (N=143) to explore the relationship between warning accuracy and callee decision-making for incoming calls. Our results suggest that previous call experience can lead to incomplete mental models of how Caller ID works. Additionally, we find that false alarms and missed detection do not impact call response but can influence user expectations of the call. Since adversaries can use mismatched expectations to their advantage, we recommend using warning design characteristics that align with user expectations under detection accuracy constraints.**

I. Introduction

Robokiller, an anti-spam call company, estimated that 54 billion spam calls were made during the 2020 Coronavirus Pandemic [37]. These reported spam calls resulted in over $3 billion lost, which included $319 million lost to pandemic-related fraud. Spam calls are easy to execute because calls are never end-to-end authenticated in the telecommunications system. Practitioners and researchers have suggested various detection solutions to this problem [4], [14], [22], [24], [26], [27], [32], [33], [38], [39], [43]–[45]. This includes block lists [27] and STIR/SHAKEN [14]. Block lists are heavily used by spam call detection applications and telecommunication carriers, and STIR/SHAKEN will eventually be used by all carriers in the United States of America to authenticate incoming callers.

Although block lists and STIR/SHAKEN can help users avoid spam calls, neither is capable of accurately detecting all calls [27], [50], which means users will be or remain immersed in assisted decision-making with accuracy limited tools. Previously, researchers have investigated what design characteristics users want to experience, how users respond to currently available and user-created warnings, accessible warning designs of incoming calls, and ways to improve warning messaging [10], [40], [41]. We extend prior work by exploring how warning accuracy influences user design preferences, expectations, and responses to incoming calls. Additionally, since research suggests accessibility may be an issue in spam detection apps [1], [40], we take the human rights approach to disability [36] and make an effort to include this community in our exploration. We answer the following research questions:

1) How do current incoming call experiences influence users' design preferences?
2) What effect do false alarms and missed detection have on the end-users decision-making process for incoming calls?
3) How does warning accuracy affect user call expectations and call response?

To answer these questions, we conducted three user studies. First, we interviewed 17 participants to provide feedback on warning designs and to "draw" or describe their ideal warning. Next, the designs created from the interviews were then shown to 27 participants in a diary study. The participants downloaded an Android app that mimicked real calls with user-created warning designs. After responding to a call, participants reflected on why they answered or declined that call in their digital diary. Lastly, we conducted a survey (n=143) using Amazon's Mechanical Turk to determine the effect of warning accuracy on user reaction and call expectations. We make the following contributions:

1) **For users, Caller ID is not a part of Spam Detection:** Participants in our study expressed a willingness to answer spam calls when the Caller ID indicated that the caller was a saved contact. For some participants, the Caller ID information was more important than the spam warning. Thus, viewing the Caller ID functionality as an entity separate from spam detection. This belief is likely amplified by their trust in Caller ID and leads to answering calls identified as spam with saved contacts.
2) **No False Alarm Effect In Spam Call Warnings:** Unlike warnings for other risks, missed detection and false alarms do not significantly change user response to warnings. However, users are likely to change their expectations of caller intent due to warning accuracy.
3) **Names Increase Spam Call Answer Rate:** The MTurk survey tested user responses to spam call warnings with and without a caller name. Participants were more likely to indicate that they would answer these spam calls once

Symposium on Usable Security and Privacy (USEC) 2024
26 February 2024, San Diego, CA, USA
ISBN 979-8-9894372-5-2
https://dx.doi.org/10.14722/usec.2024.23048
www.ndss-symposium.org, https://www.usablesecurity.net/USEC/

a name was provided.

4) **Warning Design and Low Warning Accuracy Helps Scammers:** The results of our study help explain why people answer or decline various calls. The results suggest that while warning design impacts answer rate, warning design and warning accuracy influence user expectations of a call. If an adversary can successfully spoof Caller ID and avoid a spam warning, they are more likely to be successful since users' expectations have changed. We encourage designers to use warning characteristics that communicate the accuracy of the warning (Spam Likely vs Spam Call). We also encourage future work to explore how to improve end-user call expectations under the current limitations of spam detection.

The remainder of this paper is organized as follows: Section II details related work; Sections III, IV, and V present the results of the three user studies; Section VI provides the limitations of the work; Section VII discusses the results found and suggestions for improvement of usability; and Section VIII provides concluding remarks.

II. Related Work

Prior research shows that warnings should be developed with the user's perceived risk and beliefs in mind [5], [48]. They should encompass appropriate signal words, colors, text, and symbols and should be placed where they can be easily seen [7], [48], [49]. Then, warnings should be evaluated based on their ability to change user behavior or nudge users to complete specific actions [5], [48]. Based on these guidelines, researchers have developed and analyzed auditory [11], [15], weather [19], transportation [51], provenance [42], browser [13], email phishing [29] and spam call warnings [10].

In spam call warnings, previous work investigates the implementation of user-designed warnings for incoming calls and evaluates them based on user response [41]. They found that the user-centered warnings that announced a spam call were just as likely to encourage users not to answer spam calls but more likely to encourage users to answer authenticated calls from unfamiliar numbers. Similarly, another approach investigated the potential impact of authenticated call labels on user response to calls during the initial implementation of STIR/SHAKEN [10]. Their results suggest that the use of the "Verified Number" label increased user trust and answer frequency with STIR/SHAKEN, even with a low percentage of validated calls.

Our research complements prior work by exploring the impact of warning accuracy on users' design preferences. While previous work investigates what users want to experience, our work focuses on determining if these preferences are influenced by the warning accuracy of currently available warnings. We also explore the impact of warning accuracy on decision-making strategies, expectations, and answer rates for verified and spam calls. Prior work in warnings suggests that false alarms or missed detection could impact user response in negative ways [2], [3], [46]. However, other studies suggest that the negative impact may have changed over time. For example, recent (2018) research on browser warnings has found that habituation is unlikely a major factor in user decision-making compared to "smaller contextual misunderstandings" [35]. In tornado warnings, a recent study (2019) found that the cry wolf effect was not prevalent in the southeast region of the United States due to user perception of warning accuracy and users reacting to warnings regardless of accuracy perception [21]. In car warnings, a 2016 study suggests that false alarms led to unfavorable reviews and a decrease in warning reliability for auditory-visual warnings [25]. However, false alarms did not impact the reliability of visual warnings or prevent the reduction of safety-critical events. These results motivate the exploration of false alarms in incoming call warnings and their influence on user perceptions and decision-making strategies.

III. Study 1: Design Interviews

We use design interviews to investigate how user experiences with incoming call warnings influence their design preferences and warning interpretations. We use this section to discuss the results of our study. In this section, and the remainder of the paper, we use the term spam call to reflect participant terminology and refer to any call with malicious intent.

A. Methodology

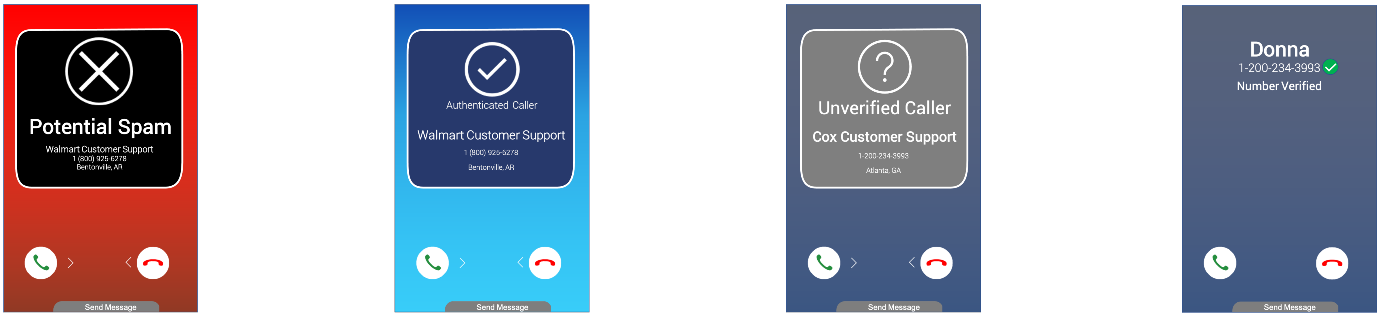
We interviewed 17 participants about their warning design preferences, interpretation of spam call warning designs (sample shown in Figure 1), and their opinion on how they could be improved. We stopped interviewing once new responses did not provide unique data points towards our research focus. Each semi-structured interview started with participants describing their ideal warning for calls identified as spam, verified, or neither. Visually impaired participants would describe the audio they wanted to hear and sighted participants describe or drew the warning visually. Then the researcher would play back the audio for blind participants. Sighted participants who describe their designs could see what the researcher was drawing as they were speaking and were often asked to confirm the drawing's accuracy. Then, participants were asked for their feedback on research-inspired warning designs. All designs had an audible version that mimicked how the warning would be read by a mobile device screen reader. Researchers drew designs for participants with internet connectivity issues. Those without issues were able to make their own designs in a shared PowerPoint.

At the end of the interview, we discussed the participant's designs for a second time to ensure accuracy and make any changes the participant now desired. Following these interviews, we presented designs that reflected participant feedback to an accessibility researcher, human-computer interaction researcher, and a lawyer who is also an accessibility rights activist. They provided additional insight to determine what designs would be used for the diary study.

Due to COVID-19, all interviews were held virtually. We asked participants for an hour of their time, but the majority of interviews lasted 30 minutes. Participants were compensated with a $10 Amazon gift card after the study. Each interview was recorded and transcribed with the participants' consent.

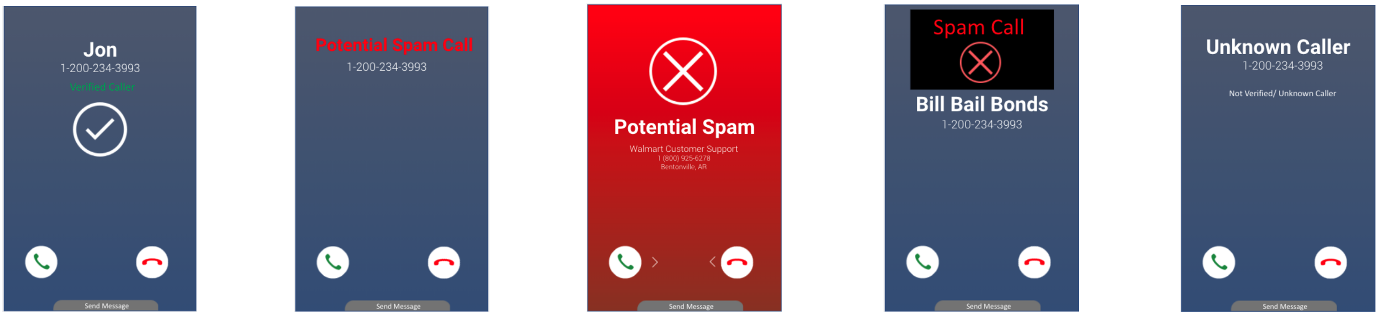
B. Ethics

This study procedure was reviewed and approved by our institution's Internal Review Board. We collected identifying



(a) PSpam-Square (b) Auth (c) UNverf-Square (d) #Verf

Fig. 1: This figure displays four of the warnings shown in the interviews. Figure 1a uses the phrase “Potential Spam” and includes an X mark. Figure 1b uses the phrase “Authenticated Caller”, has a blue background color, and has a checkmark at the top. Figure 1c uses the phrase "Unverified Caller” and has a checkmark icon at the top of the screen. Figure 1d uses the phrase “Number Verified” and has a checkmark icon to the right of the number.



(a) (b) (c) (d) (e)

Fig. 2: This figure displays five warning designs that encapsulate the designs created by the sighted interview study participants.

information from participants for payment purposes. All documentation connecting participant identities to the study were destroyed once all participants were paid and recordings transcribed.

C. Participants

We advertised our study to individuals in a research participant database and those involved in a community organization for the blind. We reached out to both communities to allow for more inclusive feedback around warning design. We interviewed 20 participants. Most of the participants were visually impaired (53%), women (53%), and over the age of 30 (53%). The demographic details are provided in Table VI in the Appendix. After the 17 interviews, three experts were interviewed to help determine which designs would be used for the diary study. Two of the experts were researchers who specialized in accessibility and human-computer interaction and the other was a lawyer and an expert in accessibility rights.

D. Results

Two researchers used thematic analysis to evaluate the interviews [9], [23]. Each researcher reviewed the transcripts and coded 35% of transcripts independently and created a codebook. They discussed results until an agreement was reached for all. The codes were used to identify the themes - Opinion based on Prior Experience, Follow the Warning or Follow Caller ID, and Opinion based on Participant Beliefs. The resulting codebook is in Appendix B.

Opinion based on Prior Experience and Beliefs: Participants were asked to describe how they would prefer to be alerted of different call types. Almost all (9 out of 10) of the visually impaired and blind participants did not have any experience with audible spam warnings but did with other call types. For example, one participant stated:

> “It will either say unknown caller, or it will give a phone number, and in rare cases, it’ll say no caller ID” -P10 (Blind, Female, 40s)

In contrast, the sighted participants had experience with spam warnings and referenced them when expressing a preference in their drawing (see Figure 2a). For example, one participant stated:

> “I sometimes, on spam calls, I get a header or something [and] it says, you know, possible spam or telemarketer or something like that. That would be beneficial [in this design] because what I typically do with those is reject them. ”-P3 (Sighted, Male, 50s)

However, the level of prior spam experience did not prevent participants' from expressing what they wanted to experience and why. Participants referenced past experiences for their design commentary, and similar to prior work [41], they also expressed a strong reliance on Caller ID. This reliance is challenged when Caller ID information is paired with a Spam warning.

During these discussions, we asked participants about the inclusion of Caller ID for each call type. When a call is unverified, every participant wanted the Caller ID information included. However, when discussing spam calls, participants had different opinions on including both the name and number. For example, three participants stated:

> "It's frustrating [if a call says potential spam and you recognize the Caller ID] and you know, how can you give me that, but you're saying it might be spam. You know, if the number is in this system and they know it's an invalid number and registered number, give me that info. " -P10 (Blind,Female,30s)

While this quote represents how Caller ID information is used and the purpose it serves for users, it also showcases the incomplete mental model users have adopted about how Caller ID works. It suggests that some users believe Caller ID and spam detection operate as separate entities. Some participants did not understand why a Caller ID they recognized was also determined by the carrier to likely be a spam call.

E. Resulting User-Inspired Field Study Designs

The designs we showed during the interviews were inspired by the designs used in previous research [10], [41]. Based on participant feedback and designs (shown in Figure 2) as well as the design feedback from the experts, we created three new designs (shown in Figure 3) for the field study. The design choices inspired by expert feedback included using *Verified Caller ID* instead of *Verified Caller* and removing the caller name for spam calls to reduce assumptions.

The results of Study 1 suggest that users notice the limitations of warning accuracy but may not respond to calls based on that accuracy. However, some participant design ideas were structured around providing a solution to improve spam call warning communication.

IV. STUDY 2: DIARY STUDY

Thus, we Once we had a better understanding of user spam call warning preferences, we decided to test these preference-based designs in a diary study.

A. Study Apparatus

We created an Android app to mimic incoming calls to conduct the within-subjects experiment. Once downloaded, participants were asked to provide the contact information of two people they regularly communicate with from their saved contacts in the app. Through the app, users received random pop-ups that mimicked incoming calls from those saved contacts and unsaved contacts. The unsaved contacts included real businesses and unidentified entities. These mock calls were randomly scheduled, similar to calls in real life, shown with the warning designs in Figure 3. Once a participant answered or declined a call, they were asked to explain why and complete a short survey. Since we were unable to provide realistic consequences, the diary entry and survey started by telling them the type of call they declined or answered.

The PSpam and Blue-ID designs in Figure 3b and 3c are based on the warning designs of popular robocall apps at the time of the study, such as Hiya [16] and Truecaller [18], to reflect warnings connected to lower accuracy.[1] The accuracy of spam call app warnings is limited by blocklists [27], [30]. Additionally, some spam call apps used the Blue-ID design to denote that the call was not identified as spam. However, the app did not implement technology to verify Caller-ID. We expressed this by implementing PSpam with a 50% false alarm rate and a 50% missed detection rate through Control and Blue-ID. This means that sometimes users would decline a call with the PSpam design and would find out that the call was not a spam call during the study. The remaining designs in Figure 3d (Unverf), 3e (XSpam), and 3f (Verf-ID) had a 0% false alarm rating, to reflect an ideal experience.

Ethics: Participants were asked to provide the contact information of two people they regularly communicate with. These contacts were saved in a file on the participants' phone and were never sent to our server. The app only sent participant responses to the server. This file was deleted when the participant deleted the app. We repeatedly reminded participants to delete the app once the study ended until the app had zero users. Additionally, since participants would likely receive real calls (spam or otherwise) during the study, mock calls could not be answered via Android's swipe functionality and an icon was added to the top of the screen when mock calls were received to help differentiate them from other calls. This study was reviewed and approved by our institution's Internal Review Board.

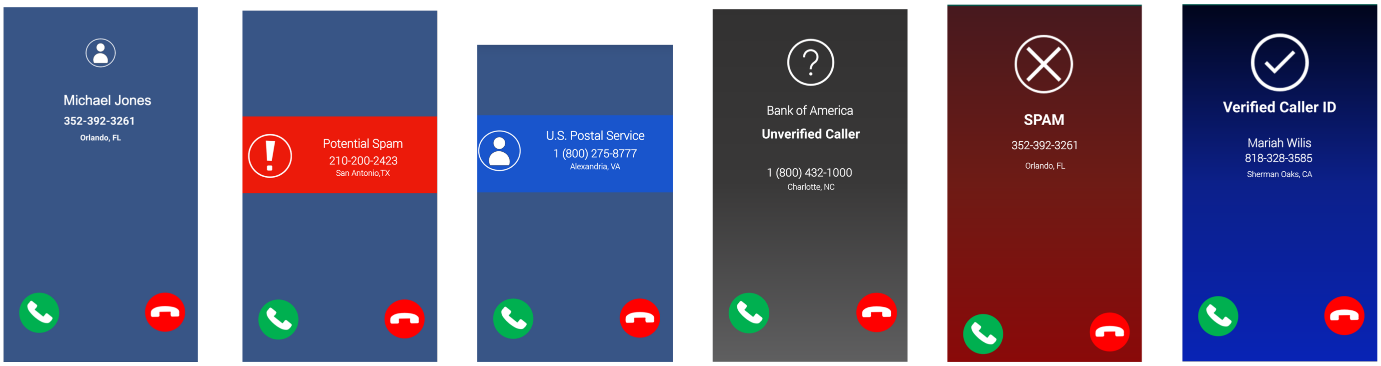
B. Methodology

Participants started the study by completing a consent form, a pre-study survey, and downloading the app. They used the app for two weeks and were asked to react to the 36 calls sent to them during that time frame. Participants received three calls from a saved contact and three calls from an unsaved contact for each of the six warning designs for a total of 36 calls. If the participant actively reacted to the call (accept or decline), they were prompted to complete a one-question survey and optional diary entry to reflect on their experience. If the participant did not react (ignore or not received), the call was rescheduled. Since we were unable to provide a complete call experience that would highlight warning accuracy, the diary entry prompt included text to inform participants if the call they reacted to was indeed a verified or spam call. After their two-week experience with the app, participants completed a post-survey. Each participant was compensated with a $20 Amazon gift card after the study. This study was approved by our local Internal Review Board.

C. Participants

We recruited participants from a research participant database. Although the study started with 30 participants, three

[1] The designs of these apps changed during the implementation of this study



(a) Control (b) PSpam (c) Blue-ID (d) Unverf (e) XSpam (f) Verf-ID

Fig. 3: This figure displays the six warnings shown in the diary study. Figure 3a only displays the Caller ID information, similar to what most users see now when no warning is present. Figure 3b uses the phrase "Potential Spam," a red rectangle behind the text, and includes an exclamation mark. Figure 3c has a blue background color only. Figure 3d uses the phrase "Unverified Caller," which includes a dark gray gradient background and has a checkmark icon at the top of the screen. Figure 3e uses the phrase “Spam”, includes a dark red gradient background, and includes an X mark. Figure 3f uses the phrase "Verified Caller ID," has a blue gradient background color and has a checkmark at the top.

TABLE I: Answer Rate During Diary Study

(n=# of Calls)	**Control** (n=132)	**Unverf** (n=112)	**PSpam** (n=122)	**XSpam** (n=137)	**Blue-ID** (n=106)	**Verf-ID** (n=129)
% Answered	24%	22%	13%	9%	28%	40%
% Answered from Saved #s	26%	18%	11%	8%	19%	22%
% Answered from Unsaved #s	8%	4%	2%	1%	9%	18%

were removed due to emulator use. In total, 27 participants completed this study. The majority of participants were white (48%), identified as female (44%), between the age of 31 and 40 (37%), and were not visually impaired (78%). Complete participant demographics are available in the Appendix in Table VI.

D. Diary Study Quantitative Results

We evaluated the call response data in the diary study and the pre and post-survey results of the participants. The call response data were compared using the Chi-Square test and test of proportions. All other results were compared using the Wilcoxon Signed-Rank Test. We used R, a statistical analysis tool, to conduct the tests [31].

1) Incoming Call Response: Each participant was sent 36 calls. Of the 972 calls sent in total, 234 calls (24%) did not reach their destination due to phones being off or in deep sleep. Participants did not answer (ignored or declined) 574 calls (59%) and answered 164 calls (17%). In Table I, we provide the data from calls that were received. The Chi-Square test showed no significant difference (p>.05) between the answer rates for PSpam (13%) and XSpam (9%) or Blue-ID (28%) and Verf-ID (40%). Additionally, there was no significant difference between the proportion of participants that answered at least one call from their first set of calls and last set of calls when shown Control (n=10 vs n=12, respectively), Unverf (n=11 vs n=8), PSpam (n=8 vs n=3), XSpam (n=5 vs n=2), Blue-ID (n=14 vs n=6), and Verf-ID (n=15 vs n=12). Thus, our results do not demonstrate that false alarms and missed detection impact the real-time answer rate in the dairy study.

2) Pre- and Post-Survey Results: Figure 4 shows the pre- and post-survey results for participant anticipated call response and expectation for each design. The results presented here are limited by the number of pre-and post-survey participant responses we were able to compare for each design, caused by a survey error. Thus, we evaluated the results for each design based on responses from participants who saw the design in both surveys. Participants’ expectation that an incoming call (from an unsaved number) was spam and response to these calls did not change over time for any of the warnings shown (p $>$.05). Thus, we found no evidence for impact of false alarms and missed detection on participant responses or expectations of spam during this study.

E. Diary Results

Two researchers used thematic analysis to code the diary entries and participant answers to the three extended response questions in the post-survey. This method involves deriving categories based on the information provided by the participants or inductive analysis. While we did code before creating themes, the codes and themes are very similar because these entries focused on explaining the participant’s call response and the participants had very similar answers. There were 350 diary entries. Both researchers independently coded 190 (43%) of the diary entries and 45 (50%) of the survey responses. This was followed by a meeting where the independent codebooks

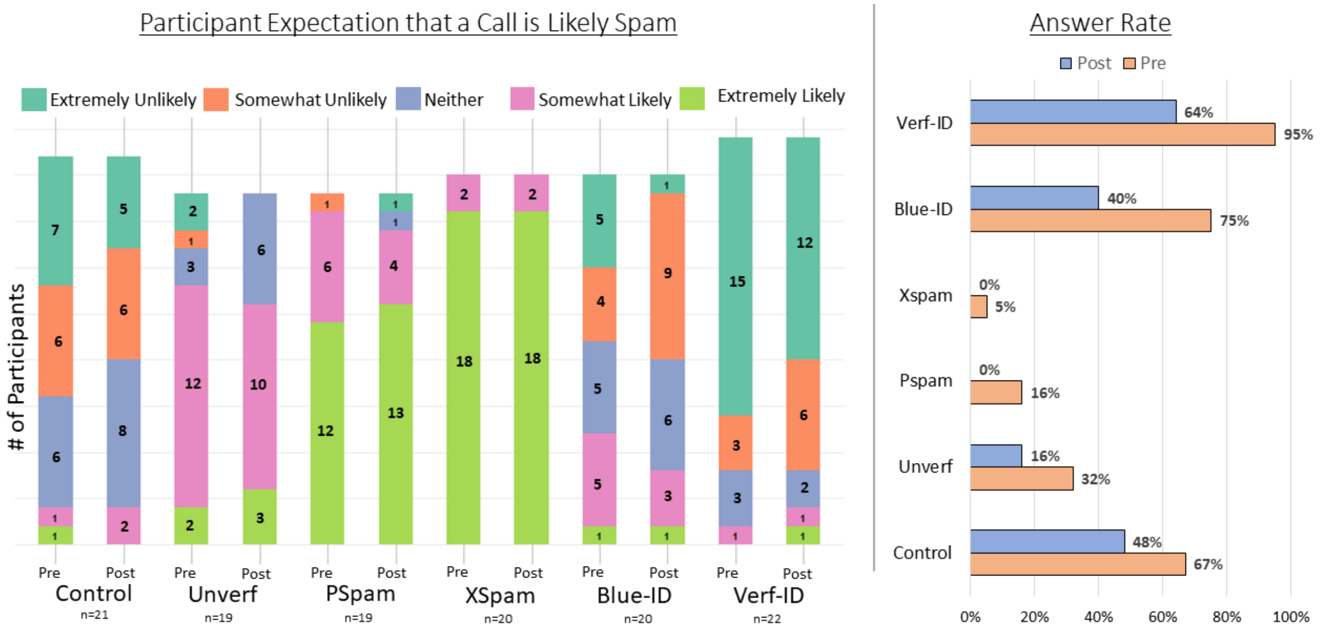


Fig. 4: Perceived Spam Likelihood and Answer Rate Results from the Field Study Pre- and Post-Survey

were combined, and all disagreements in coding the entries were discussed until agreement and the themes were developed. The codebook is shown in Appendix D.

User Response to Control, Blue-ID, and Unverf: Participants relied heavily on Caller ID, and the time they had available to talk to determine if they would answer calls with the Control, Blue-ID, and Unverf design. However, when some participants declined a call from an unknown number with Blue-ID, they stated it was because they did not have business with the organization and thus declined. Additionally, when some participants declined a call from an unknown number with Unverf, they wrote it was because it was unverified or it appeared to be a spam call to them.

Although Unverf and Blue-ID do not explicitly flag whether the call is spam or not, participants' diary responses suggest they still relied on design characteristics to assist them in making their decision when they had no relationship with the caller. However, when they "knew" the caller, some participants answered the unverified call.

User Response to Verf-ID: In response to Verf-ID, many participants answered calls from saved contacts and declined calls from unsaved contacts. However, some participants answered calls from unsaved contacts when they were verified. For example, four participants wrote the following when reflecting on this choice:

> "I realized this may be a spam call but I used to live in Florida" -P21 (Male, 50s, Sighted)
>
> "[M]y kids live in texas and the area code is 210. I felt extremely comfortable with answering." -P23 (Male, 60+, Visually Impaired)

In these cases, the combination of warning design, Caller ID, and false alarm rate likely affected the callee's perception of the caller's intention, thus impacting their decision and expectations. For P21, the caller was not a saved contact. However, it was verified and from a Florida area code. This made the participant feel more comfortable answering the call since they used to live there and still have associates in the area. In the diary entries, when participants described being comfortable answering calls from unsaved contacts, it was always under the Verf-ID condition. Since this was not expressed in the Control or Blue-ID condition, this suggests that the low false alarm rate was likely a factor in participants' decision to answer.

1) User Response to PSpam and XSpam: When participants rejected a call, from saved and unsaved callers with PSpam or XSpam, they noted that it was because the warning said that the call was a spam call. However, some participants still answered those calls when they came from numbers that they recognized. For example, two participants wrote the following when reflecting on why they accepted these calls:

> "I knew the number even though it was labeled as spam." -P24 (Female, 20s, Visually Impaired)
>
> "I thought it was miscategorized" -P18 (Female, 30s, Sighted)

While participants did take the warnings into account, their final decisions were based on the Caller ID. We saw this response for both XSpam and PSpam, which likely means that decisions were based on the user's confidence in their understanding of how detection was being done. But the outcome of this was not always positive. Two participants expressed their frustration when they realized they answered a spam call or declined a legitimate call due to incorrect warnings. The two participants wrote the following:

> "[I]f it's not spam, why did you say it's potential spam????!" -P25 (Female,40s, Sighted)

> "The one part of the calls that threw me was that I would receive calls from my husband's number... But at times, the app told me I picked up a spam call. How would I know this?" - P26 (Female, 20s, Sighted)

These results suggest that warning accuracy, in combination with warning design, can influence user perception of caller intentions during the study when the call is verified. However, users have difficulty correctly interpreting warning phrases for spoofed calls with saved numbers even when the name is removed. The results suggest that Caller ID is a main factor in this confusion.

V. Study 3: Impact of Warning Design and Accuracy

To further test the impact of warning accuracy, we conducted a survey on Amazon's Mechanical Turk to identify participant expectations and reactions to various warnings once accuracy has been made known to them. The methodology and results are discussed below.

A. Methodology

We created a survey that displayed nine designs. The first six warnings were the same designs used in the field study. The other three designs included the two spam call designs from Study 2 (PSpam in Figure 3b and XSpam in Figure 3e) but with full Caller ID information (name, number, location) and #Verf from Study 1 (Figure 1d). We added the spam call designs with the caller name to understand how callers might react to spam warnings with (PSpam+Name and XSpam+Name) and without the full Caller ID information present to evaluate Caller ID in spam warnings.

The first part of the survey asked participants to indicate what they expected from the call and how they might respond to each design. Then they were told how accurate each design was and asked to indicate their response and expectations again. To mimic participants receiving both legitimate and spoofed calls from people or places they recognized, participants were asked to pretend as though they worked for and were friends with the businesses and people named in the survey.[2] They were also prompted to make decisions quickly to mimic the required response time in the real world. We distributed the survey on Amazon's Mechanical Turk. We used this platform because prior works suggests that Mturk worker responses are representative of Americans between 18 and 50 years of age when discussing privacy preferences [34]. MTurk workers had to have a 95% approval rating, be located in th United States, and accurately respond to an attention check where the instructions told them what option to select. We also looked at response time and responses to open ended questions to identify and remove bots. We ended with 143 participants and paid each participant $1 for their participation in the 5-minute survey. This was a repeated measures study as all designs were shown twice to each participant in random order.

B. Ethics

This study was reviewed and approved by our institution's Internal Review Board. We collected MTurk IDs from participants for payment purposes. All documentation connecting participant identities to the study were destroyed once all participants were paid.

C. Participants

All survey participants (n=143) were Amazon Mechanical Turk crowdsource workers. Most participants were white (71%), between 31 and 40 (30%), male (62%), and were not visually impaired (78%). One participant indicated that they were blind, and 30 identified as being visually impaired or having low vision. The participant demographic details are provided in Table VI in the Appendix.

D. Results

Participant responses were analyzed using the Wilcoxon signed rank and Wilcoxon rank-sum with bonferroni correction through R [31]. The Wilcoxon signed-rank Test was used to compare participant call expectations and responses before and after warning accuracy was announced. It was also used to determine if participants' expectations or responses to spam calls changed due to the accuracy of the warning. The Wilcoxon rank-sum test was used to compare the responses based on visual ability to determine if users' expectation or response to incoming calls is influenced by those characteristics, similar to other spam studies [6].

1) Participant Response to Low Accuracy: We informed participants (n=143) that PSpam and PSpam + Name were not always accurate and that Blue-ID was not verifying the calls it displayed. The participant call response and call expectation for Blue-ID did not change significantly with this knowledge, as shown in Table II. However, the lack of accuracy did impact users' call expectations of PSpam (p<.05, r=.202, z=-2.418), PSpam+Name (p<.05, r=.199, z=-2.380), and their call response to PSpam+Name (p<.05, r=.224, z=-2.684). Thus, the accuracy of PSpam and PSpam+Name influenced participant's change in expectations and call response once the caller name was added.

2) Participant Response to High Accuracy: We informed participants (n=143) that XSpam, XSpam + Name, Unverf, Verf-ID, and #Verf were always accurate. This additional information changed the participant's call expectations for Unverf (p <.05, r^2=.245, z=-2.931) and did not change the participant's reaction to XSpam, XSpam + Name, Verf, and #Verf. Thus, the inability to verify an incoming call under the Unverf design decreased participants' belief the call was likely a spam call. In contrast, the high accuracy of the XSpam, XSpam+Name, Verf-ID, and #Verf did not significantly change users responses or expectations of the calls.

3) Participant Response to Spam Caller "Name": We compared participants' call expectations and responses to PSpam, XSpam, PSpam+Name, and XSpam+Name, before their accuracy was announced (see Table III). Participants (n=143) call responses did not significantly change between PSpam and PSpam + Name. However, their expectation of the call being spam significantly changed between PSpam and PSpam+Name

[2] The specific wording used included the names of the organizations and people

TABLE II: Comparing User Response to Non-Spam Warnings Before and After Notification of Accuracy

	Control	Blue-ID		Unverf		Verf-ID		#Verf	
Expectation		*Before*	*After*	*Before*	*After*	*Before*	*After*	*Before*	After
Extremely likely	14%	11%	9%	15%	12%	11%	17%	9%	13%
Somewhat likely	13%	18%	15%	31%	21%	14%	11%	6%	3%
Neither likely	21%	20%	35%	33%	35%	10%	9%	4%	4%
Somewhat unlikely	17%	27%	28%	14%	20%	13%	6%	12%	11%
Extremely unlikely	35%	24%	13%	8%	13%	52%	57%	70%	71%
Response									
Answer	83%	70%	63%	52%	53%	89%	89%	92%	92%
Decline	9%	14%	15%	19%	22%	7%	8%	5%	5%
Other	8%	16%	22%	29%	25%	4%	3%	3%	3%

TABLE III: Comparing User Response to Spam Warnings Before and After Notification of Accuracy

	XSpam		XSpam+Name		PSpam		PSpam+Name	
Expectation	*Before*	*After*	*Before*	*After*	*Before*	*After*	*Before*	*After*
It is extremely likely that this is a spam call	57%	60%	46%	48%	41%	27%	27%	22%
It is somewhat likely that this is a spam call	20%	13%	21%	21%	36%	42%	39%	33%
It is neither likely nor unlikely that this is a spam call	7%	11%	17%	8%	7%	17%	15%	18%
It is somewhat unlikely that this is a spam call	8%	8%	10%	14%	9%	9%	13%	18%
It is extremely unlikely that this is a spam call	8%	8%	5%	8%	6%	6%	6%	9%
Response								
Answer	9%	10%	25%	16%	14%	14%	21%	35%
Decline	61%	65%	49%	53%	55%	49%	48%	40%
Other	30%	25%	26%	31%	31%	37%	31%	25%

TABLE IV: Comparing Visually Impaired and Sighted User Response to Spam Warnings Before Notification of Accuracy

	XSpam		XSpam+Name		PSpam		PSpam+Name	
Expectation	*VI*	*S*	*VI*	*S*	*VI*	*S*	*VI*	*S*
It is extremely likely that this is a spam call	13%	70%	16%	54%	23%	46%	19%	29%
It is somewhat likely that this is a spam call	35%	15%	26%	20%	32%	38%	23%	44%
It is neither likely nor unlikely that this is a spam call	13%	5%	29%	14%	19%	4%	35%	9%
It is somewhat unlikely that this is a spam call	26%	3%	23%	7%	16%	7%	23%	11%
It is extremely unlikely that this is a spam call	13%	7%	6%	4%	10%	5%	0%	8%
Response								
Answer	26%	4%	32%	23%	32%	9%	29%	19%
Decline	55%	63%	39%	52%	32%	62%	55%	46%
Other	19%	33%	29%	25%	35%	29%	16%	36%

VI=Visually Impaired Participants (n=31)
S= Sighted Participants (n=112)

TABLE V: Comparing Survey Responses from Visually Impaired and Sighted Participants to Non-Spam Warnings

	Control		Blue-ID		Unverf		Verf-ID		#Verf	
Expectation	*VI*	*S*	*VI*	*S*	*VI*	*S*	*VI*	*S*	*VI*	*S*
Extremely Likely	29%	10%	29%	6%	29%	11%	26%	7%	32%	9%
Somewhat Likely	32%	8%	39%	13%	26%	32%	39%	7%	26%	6%
Neither	29%	19%	19%	20%	26%	35%	19%	8%	23%	4%
Somewhat Unlikely	10%	19%	6%	32%	10%	15%	13%	13%	16%	12%
Extremely Enlikely	0%	45%	6%	29%	10%	7%	3%	65%	3%	70%
Response										
Answer	65%	88%	74%	69%	68%	47%	77%	92%	77%	92%
Decline	26%	4%	13%	14%	23%	18%	13%	5%	10%	5%
Other	10%	7%	13%	17%	10%	35%	10%	3%	13%	3%

VI = Visually Impaired (n=31)
S=Sighted (n=112)

(p<.05, r^2=.252, z=-3.013). In contrast, participant expectations of a call being spam did not change between XSpam and XSpam + Name. However, their call response significantly changed between XSpam and XSpam+Name (p<.05, r^2=.307, z=-3.677). Thus, the addition of a caller name decreased participants' belief that the call was likely a spam call for PSpam but increased the answer rate for XSpam once the name of the caller was added.

4) Seeing vs. Hearing : Prior work suggests that those with visual impairment are more likely to correctly detect spam emails due to hearing the information instead of seeing the information [6]. Thus, we evaluated if sighted (n=112) and visually impaired (n=31) participants responded to the warnings differently. Table IV shows the call responses and expectations from the two groups.

Before Accuracy Notification: The visually impaired participants expected a higher chance of spam compared to sighted participants for Control (p<.001, r^2=.466, z=5.574), Verf-ID (p<.001, r^2=.546, z=6.539), #Verf (p<.001, r^2=.556, z=6.655) and Blue-ID design (p<.001, r^2=.433, z= 5.178). There was no significant difference between the expectations of the visually impaired and sighted when shown Unverf. Additionally, visually impaired participants were less likely to answer a call with Verf-ID (p<.05, r^2=.187, z=-2.235) and #Verf warning (p<.05, r^2=.183, z=-2.19). However, there was no significant difference between the way each group responded to Control, Blue-ID, and Unverf.

When shown the remaining designs, the sighted participants expected a higher chance of spam for PSpam (p<.05, r^2=.257, z=-3.072), XSpam (p<.001, r^2=.464, z= -5.551), and XSpam+Name (p<.001, r^2=.327, z= -3.909) when compared to the visually impaired participants. There was no difference in the exceptions of the PSpam+Name call when we compared the groups. Lastly, sighted participants were less likely to answer a call with the PSpam warning (p<.001, r^2=.284, z=3.391). However, there was no significant difference between the way each group responded to PSPam+Name, XSpam, and XSpam+Name.

After Accuracy Notification: The visually impaired participants' (n=31) call expectations and call response did not significantly change after accuracy notification for every warning. However, the sighted participants (n=112) call expectations for Unverf (p<.05, r^2=.301, z=-3.603), Blue-ID (p<.05, r^2=.234, z=-2.803), PSpam (p<.01, r^2=.276, z=-3.295), and PSpam+Name (p<.05, r^2=.246, z=-2.603) changed. Their call response changed for PSpam+Name (p<.001, r^2=.774, z=-8.191). Lastly, sighted participants did not change their call expectation or response for Control, Verf-ID, #Verf, XSpam, and XSpam+name.

These results suggest that visually impaired users do not respond to warnings in the same way that sighted users respond. The visually impaired participants were less likely to answer verified calls and more likely to answer spam calls, when compared to sighted participants. They had low expectations of a call being spam when accompanied by a spam warning. Sighted participants' call expectations and responses were more likely to be influenced by warning design and accuracy than those with a visual impairment.

VI. Limitations

The study 1 and 2 results may not represent the experiences and beliefs of the general U.S. population since our participant group is not a representative sample. However, due to the consistency of the results across studies, and prior work, we believe the outcomes hold validity. In study 2, the real-time data in our study was limited by overall participant response. Although various reports suggest that U. S. residents answer 48% or less of all incoming calls [17], it is possible that elements of our study design potentially contributed to that. The use of various warning designs, a small icon to help differentiate calls, and a short study period period, could have contributed to the response rate. Study 3 is limited by ecological validity, similar to other studies that show screenshots of information to observe user behavior [28], [47]. We recognize that users would not be told the accuracy of the warnings when experiencing them in real life. However, we believe our results are still applicable because they represent how users believe they would react to incoming call warnings after the warning accuracy is clarified.

VII. Discussion

In this section, we review the results from the studies and discuss how they apply to our research questions. Then we provide design suggestions based on the results found.

A. How do current incoming call experiences influence users' design preferences?

In the design interviews, we asked users to design incoming call warnings and provide feedback on a set of designs shown during the interview. We found that user design preferences are influenced by their current beliefs and prior experiences with calls. On multiple occasions, participants referenced their incoming call warning experiences to help describe what they wanted to see and why. This suggests that mobile device users are remembering the warnings they are seeing and find all or some of the design characteristics useful.

However, participants were not in agreement about including Caller-ID for spam calls and verified calls. Their varied responses suggest that some users may have incomplete mental models for how Caller ID works and its current relationship to spam detection. Participants' interview responses suggest that some may not be aware of Caller-ID limitations, leading to differences in preference. This also suggests that the nuance in current warning messaging may not be clearly communicated to end users, thus impacting their design preferences.

B. What effect do false alarms and missed detection have on end-users' incoming call decision-making process?

The qualitative analysis in the diary study suggests that participants determined whether or not to accept, decline, or ignore a call based on their assessment of the information in the warning. Some participants answered verified calls from unsaved contacts because they recognized the business and believed their intention was not malicious. Some participants answered "spam" or "potentially spam" calls from a saved contact because they recognized the number and thus assumed the system had simply miscategorized the caller. This suggests that name recognition influences user decision-making and warning accuracy is not a factor in user decision-making for incoming calls. Additionally, while participants did not mention warning or technology accuracy in the interviews or diary study directly, Caller-ID was always associated with the suspicion of warning error.

C. How does warning accuracy affect user call expectations and response?

Our results suggest that warning accuracy does not impact call response. In the diary study, we found that participants' first and last responses to each call type did not change significantly. In the MTurk survey, after participants were provided with additional information regarding the accuracy of each warning, their call response only changed for PSpam+Name with a small effect size. Thus, warning accuracy is unlikely to influence the answer rate. We believe this is due to the framing effect.Warnings reframe the situation callees are responding to. Without a warning, the callee is asked to respond to an incoming call. With a warning, the callee is being asked to respond to a Spam call. Similarly, instead of being asked to respond to an incoming call from a friend, the callee is being asked to respond to a Spam Call from a friend. These are different questions, as evidenced by the difference in answer rate for the Control and XSpam warnings in the field and survey study.

The results also suggest that warning accuracy can influence users' call expectations in the form of spam likelihood. In the diary study, Caller-ID directly related to what participants' expected the call to be about and the warning error beliefs. Additionally, the survey participants changed their expectations of calls with the Unverf, PSpam, and PSPam+Name warning designs, three of the four designs with lower accuracy. These results suggest that warning accuracy can impact user expectations that a call is a likely spam call if they are prompted to consider it. However, this response is not an example of the false alarm effect [8] since it does not necessarily mean credibility or trust is lost. Accuracy influenced expectations but did not change behavior in the study. We encourage future work to focus on evaluating how warning accuracy for spam call detection impacts end-user trust in the technology.

D. Implications

Spammers Will Benefit: The prevalence of spam call victims is more than likely due to successful spoofing techniques, evading current spam detection techniques, and scammers with social engineering skills. For example, participants were shown PSpam and XSpam, with and without the caller's name in the survey. The addition of the caller name decreased the callee's expectation of spam and increased the answer rate. These elements create an ideal environment for spammers since it could assist them in making the scheme more convincing.

Hearing and Seeing Information: We included both sighted and visually impaired participants for every study we conducted. In Study 3, we included these participants to compare the impact of hearing versus viewing the warning information. We recognize warning design is often heavily focused on visual characteristics even though the audible version is equally important for user experience. Our results suggest that call expectations and call responses of the visually impaired are unlikely to be influenced by warning accuracy. We believe this is potentially caused by visually impaired users' reliance on Caller ID, their limited experience with warnings, and the exclusion of visually impaired users in designing spam call warnings. Additionally, the visually impaired participants were more critical of the warnings, but it is not clear if they are more or less susceptible to spam call schemes. We encourage other researchers to explore the design of audible spam call warnings and how susceptible the visually impaired community might be to spam call schemes.

Prioritize Increasing Suspicion: The diary study and MTurk survey suggest that warnings directly impact user decision-making and perception of the caller. Not only were participants more likely to answer a call based on the warning design, but in some cases, the warning influenced the way participants perceived the caller's intention. During the MTurk survey, we saw that the addition of the caller's name to a spam warning changed the participant's expectations and response to the PSpam and XSpam design. Although we don't want users to answer spam calls, we believe that setting their expectations for those calls is a priority. Our study shows that while some users may be more likely to answer an identified spam call with more Caller ID information, they are also likely to answer the call believing the call has some chance of being a spam call. If users are going to answer spam calls, they should do so with some level of suspicion. Unlike other warnings, incoming call warnings are referring to organizations and people, some of whom the callee has a direct relationship with. While HTTPS indicators can refer to similar entities [12], phone calls provide a feeling of connectedness to people that are not replicated in text-based communications [20]. For this reason, we encourage future work to focus on user comprehension to further explore how warnings can prompt incoming call suspicion and determine what impacts the level of caution.

Spam Warning Design: Currently, when a user receives a spam call, they are warned about the call and provided with the name associated with the number presented to them. Since users often reference the name of the caller to determine if they should answer a call, we recommend omitting the name from Caller ID when a call is identified as a spam call. Additionally, designers should consider the accuracy of the technology when choosing warning phrases. The phrase "Spam " and "Potential Spam" should match the accuracy rate of the spam detection methods being used to prevent user confusion. This also means that the phrase "Verified Caller" or "Verified Number" should only be used when verification has a low false-positive rate. Additionally, we encourage future work on phrases such as "Spam Risk" to understand how users interpret the phrase in comparison to what it is intended to be communicated. This phrase was not suggested by our participants but we recognize it is often used to communicate the same message as "Potential Spam".

VIII. Conclusion

In this paper, we investigated how warning accuracy impact design preferences, decision-making, and incoming call expectations and response. First, we interviewed users about their warning preferences. Then a diary study was conducted to capture user expectations and responses to spam calls in real-time. Lastly, an MTurk survey was used to determine how warning accuracy impacts decision-making. The results of these studies suggest that a user's call response and expectation are not influenced by warning accuracy. However, their call response and expectations are influenced by warning characteristics. We encourage warning designers to implement warnings and consider incorporating information that aligns with user comprehension of Caller-ID and other warning characteristics.

References

[1] A. Alshayban, I. Ahmed, and S. Malek, "Accessibility issues in android apps: State of affairs, sentiments, and ways forward," in *Proceedings of the ACM/IEEE 42nd International Conference on Software Engineering*, 2020, pp. 1323–1334.

[2] B. B. Anderson, A. Vance, J. L. Jenkins, C. B. Kirwan, and D. Bjornn, "It all blurs together: How the effects of habituation generalize across system notifications and security warnings," in *Information Systems and Neuroscience: Gmunden Retreat on NeuroIS 2016*. Springer, 2017, pp. 43–49.

[3] H. Azevedo-Sa, S. K. Jayaraman, C. T. Esterwood, X. J. Yang, L. P. Robert Jr, and D. M. Tilbury, "Comparing the effects of false alarms and misses on humans' trust in (semi) autonomous vehicles," in *Companion of the 2020 ACM/IEEE International Conference on Human-Robot Interaction*, 2020, pp. 113–115.

[4] V. A. Balasubramaniyan, A. Poonawalla, M. Ahamad, M. T. Hunter, and P. Traynor, "Pindr0p: using single-ended audio features to determine call provenance," in *Proceedings of the 17th ACM conference on Computer and communications security*. ACM, 2010, pp. 109–120.

[5] L. Bauer, C. Bravo-Lillo, L. Cranor, and E. Fragkaki, "Warning design guidelines," *Carnegie Mellon University, Pittsburgh, PA*, 2013.

[6] M. Blythe, H. Petrie, and J. A. Clark, "F for fake: four studies on how we fall for phish," in *Proceedings of the SIGCHI conference on human factors in computing systems*, 2011, pp. 3469–3478.

[7] C. C. Braun, P. B. Mine, and N. C. Silver, "The influence of color on warning label perceptions," *International journal of industrial ergonomics*, vol. 15, no. 3, pp. 179–187, 1995.

[8] S. Breznitz, *Cry wolf: The psychology of false alarms*. Psychology Press, 2013.

[9] N. Brown and T. Stockman, "Examining the use of thematic analysis as a tool for informing design of new family communication technologies," in *27th International BCS Human Computer Interaction Conference (HCI 2013) 27*, 2013, pp. 1–6.

[10] G. W. Edwards, M. J. Gonzales, and M. A. Sullivan, "Robocalling: STIRRED AND SHAKEN!-An Investigation of Calling Displays on Trust and Answer Rates," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[11] J. Edworthy, S. Loxley, and I. Dennis, "Improving auditory warning design: Relationship between warning sound parameters and perceived urgency," *Human factors*, vol. 33, no. 2, pp. 205–231, 1991.

[12] 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 *SOUPS*, 2016, pp. 1–14.

[13] 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.

[14] A. for Telecommunications Industry Solutions, "Signature-based handling of asserted information using tokens (shaken): Governance model and certificate management," 2017. [Online]. Available: http://www.atis.org/sti-ga/resources/docs/ATIS-1000080.pdf

[15] E. J. Hellier, J. Edworthy, and I. Dennis, "Improving auditory warning design: Quantifying and predicting the effects of different warning parameters on perceived urgency," *Human factors*, vol. 35, no. 4, pp. 693–706, 1993.

[16] hiya, "Robocall radar, spam trends: 2018 report," 2018. [Online]. Available: https://hiya.com/robocall-radar

[17] Hiya, "Hiya state of the call 2022 report," 2022. [Online]. Available: https://www.hiya.com/state-of-the-call

[18] K. F. Kok, "Truecaller insights," 2018. [Online]. Available: https://truecaller.blog/2018/04/26/truecaller-insights-usa-2018/

[19] T. Kox, L. Gerhold, and U. Ulbrich, "Perception and use of uncertainty in severe weather warnings by emergency services in germany," *Atmospheric Research*, vol. 158, pp. 292–301, 2015.

[20] A. Kumar and N. Epley, "It's surprisingly nice to hear you: Misunderstanding the impact of communication media can lead to suboptimal choices of how to connect with others." *Journal of Experimental Psychology: General*, vol. 150, no. 3, p. 595, 2021.

[21] J. R. Lim, B. F. Liu, and M. Egnoto, "Cry wolf effect? evaluating the impact of false alarms on public responses to tornado alerts in the southeastern united states," *Weather, climate, and society*, vol. 11, no. 3, pp. 549–563, 2019.

[22] F. Maggi, "Are the con artists back? a preliminary analysis of modern phone frauds," in *2010 10th IEEE International Conference on Computer and Information Technology (CIT 2010)*. IEEE, 2010, pp. 824–831.

[23] 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.

[24] H. Mustafa, W. Xu, A. R. Sadeghi, and S. Schulz, "You can call but you can't hide: Detecting caller id spoofing attacks," in *2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks*, June 2014, pp. 168–179.

[25] F. Naujoks, A. Kiesel, and A. Neukum, "Cooperative warning systems: The impact of false and unnecessary alarms on drivers' compliance," *Accident Analysis & Prevention*, vol. 97, pp. 162–175, 2016.

[26] S. Pandit, J. Liu, R. Perdisci, and M. Ahamad, "Applying deep learning to combat mass robocalls," in *2021 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2021, pp. 63–70.

[27] S. Pandit, R. Perdisci, M. Ahamad, and P. Gupta, "Towards measuring the effectiveness of telephony blacklists." in *NDSS*, 2018.

[28] K. Parsons, A. McCormac, M. Pattinson, M. Butavicius, and C. Jerram, "The design of phishing studies: Challenges for researchers," *Computers & Security*, vol. 52, pp. 194–206, 2015.

[29] J. Petelka, Y. Zou, and F. Schaub, "Put your warning where your link is: Improving and evaluating email phishing warnings," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–15.

[30] S. Prasad, E. Bouma-Sims, A. K. Mylappan, and B. Reaves, "Who's calling? characterizing robocalls through audio and metadata analysis," in *29th USENIX Security Symposium (USENIX Security 20)*, 2020, pp. 397–414.

[31] 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/

[32] B. Reaves, L. Blue, H. Abdullah, L. Vargas, P. Traynor, and T. Shrimpton, "Authenticall: Efficient identity and content authentication for phone calls," in *26th USENIX Security Symposium (USENIX Security 17)*. Vancouver, BC: USENIX Association, 2017, pp. 575–592. [Online]. Available: https://www.usenix.org/conference/usenixsecurity17/technical-sessions/presentation/reaves

[33] B. Reaves, L. Blue, and P. Traynor, "AuthLoop: End-to-End Cryptographic Authentication for Telephony over Voice Channels," in *Proceedings of the USENIX Security Symposium (SECURITY)*, 2016, (Acceptance Rate: 15.5%).

[34] 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.

[35] 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*. ACM, 2018, p. 512.

[36] M. Retief and R. Letšosa, "Models of disability: A brief overview," *HTS Teologiese Studies/Theological Studies*, vol. 74, no. 1, 2018.

[37] Robokiller, "Robokiller 2020 Robocall Insights Report," *Teltech.co,*, 202.

[38] M. Sahin, M. Relieu, and A. Francillon, "Using chatbots against voice spam: Analyzing lenny's effectiveness," in *Symposium on Usable Privacy and Security*, 2017, pp. 319–337.

[39] A. Sheoran, S. Fahmy, C. Peng, and N. Modi, "Nascent: Tackling caller-id spoofing in 4g networks via efficient network-assisted validation," in *IEEE INFOCOM*, 2019, pp. 676–684.

[40] I. N. Sherman, J. D. Bowers, L.-L. Laborde, J. E. Gilbert, J. Ruiz, and P. G. Traynor, "Truly visual caller id? an analysis of anti-robocall applications and their accessibility to visually impaired users," in *2020 IEEE International Symposium on Technology and Society (ISTAS)*. IEEE, 2020, pp. 266–279.

[41] 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," *The Network and Distributed System Security Symposium (NDSS)*, 2020.

[42] I. N. Sherman, J. W. Stokes, and E. M. Redmiles, "Designing media provenance indicators to combat fake media," in *Proceedings of the 24th International Symposium on Research in Attacks, Intrusions and Defenses*, 2021, pp. 324–339.

[43] A. Tasidou, P. S. Efraimidis, Y. Soupionis, L. Mitrou, and V. Katos, “Privacy-preserving, user-centric VoIP CAPTCHA challenges,” *Information & Computer Security*, 2016.

[44] H. Tu, A. Doupé, Z. Zhao, and G.-J. Ahn, “Toward authenticated caller id transmission: The need for a standardized authentication scheme in q. 731.3 calling line identification presentation,” in *ITU Kaleidoscope: ICTs for a Sustainable World (ITU WT)*. IEEE, 2016, pp. 1–8.

[45] D. Ucci, R. Perdisci, J. Lee, and M. Ahamad, “Towards a Practical Differentially Private Collaborative Phone Blacklisting System,” in *Annual Computer Security Applications Conference*, 2020, pp. 100–115.

[46] A. Vance, “The fog of warnings: how non-essential notifications blur with security warnings,” in *Symposium on Usable Privacy and Security (SOUPS)*, 2019.

[47] J. Wang, T. Herath, R. Chen, A. Vishwanath, and H. R. Rao, “Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email,” *IEEE transactions on professional communication*, vol. 55, no. 4, pp. 345–362, 2012.

[48] M. S. Wogalter, V. C. Conzola, and T. L. Smith-Jackson, “based guidelines for warning design and evaluation,” *Applied ergonomics*, vol. 33, no. 3, pp. 219–230, 2002.

[49] M. S. Wogalter, S. W. Jarrard, and S. N. Simpson, “Influence of warning label signal words on perceived hazard level,” *Human Factors*, vol. 36, no. 3, pp. 547–556, 1994.

[50] J. Yu, “An analysis of applying stir/shaken to prevent robocalls,” in *Advances in Security, Networks, and Internet of Things*. Springer, 2021, pp. 277–290.

[51] H. Yun and J. H. Yang, “Multimodal warning design for take-over request in conditionally automated driving,” *European transport research review*, vol. 12, pp. 1–11, 2020.

Appendix

TABLE VI: Interview Demographics Data

Demographics	Participants					
	Study 1 N=20		Study 2 N=30		Study 3 N=143	
	n	%	n	%	n	%
Vision						
Not Visually Impaired	10	50%	25	83%	112	78%
Visually Impaired	10	50%	5	17%	31	22%
Gender Identity						
Female	10	50%	19	63%	53	37%
Male	10	50%	11	37%	89	62%
Age						
18-30	3	15%	18	27%	45	31%
31-40	4	20%	12	40%	43	30%
41-50	1	5%	7	23%	28	20%
Over 50	11	55%	3	10%	27	19%
Prefer not to answer	1	5%	-	-	-	-
Race						
White	12	60%	21	70%	101	71%
Asian	3	15%	7	23%	13	9%
Hispanic or Latino	1	5%	4	13%	4	3%
Black or African American	5	25%	3	10%	12	8%
Prefer not to answer	1	5%	-	-	-	-
Education						
High School Graduate	1	5%	1	3%	8	6%
Some college but no degree	2	10%	3	10%	13	9%
Associate’s	3	15%	-	-	10	7%
Bachelor’s	6	30%	7	23%	82	57%
Advanced degree	6	30%	9	30%	30	21%
Prefer not to answer	2	10%	10	33%	-	-

A. *Questions from Interview (Study 1)*

1) Imagine you have just downloaded an app to protect you from spam calls. You receive a call and it is a spam call. Please describe to me what you like to see on your screen and I will attempt to draw your vision.
2) Imagine you have just downloaded an app to protect you from spam calls. You receive a call and it tells you the caller id information might be could not be verified. Please describe to me what you like to see on your screen and I will attempt to draw your vision.
3) Imagine you have just downloaded an app to protect you from spam calls. You receive a call and it tells you the call is from a verified caller. Please describe to me what you expect to see on your screen and I will attempt to draw your vision.
4) In the remaining time that we have, I will show (play) you some warnings and I want to get your feedback on them. Please tell me what you like and dislike for each design and how you might react.
5) In your own words, what is the difference between verified, authorized, and authenticated.
6) In your own words, what is the difference between unverified, spam, and unauthenticated.

B. *Codebook from Interviews (Study 1)*

- no changes - no issues with the design
- use of labels - Preferences of Visually Impaired
- current experience - Participant discusses their experience
- voicemail usage - voice mail is spam solution
- call options - Participant detail calls they receive
- too much - The participants say there is to much information in the warning
- dont understand - confused by warning
- use of caller id - Participant prioritizes Caller ID info
- spam def - The participant defines spam,
- (un)auth/ver def - The particpant defines (un)auth/ver
- auth/ver diff - Describes difference between authorization and verification
- design choices - Suggests use of icons or colors
- use of photos - Suggests use of photos in new designs

C. *Questions for Field Study (Study 2)*

1) How would you respond to this call?
2) This warning tells me that.... (likert scale)
3) The call is (definitely, probably, etc) a spam call (not a spam call)
4) Please use the space below to explain your response or provide any comments you may have about the warning.
5) Overall, how difficult or easy was it to understand this warning? (rating)
6) Please use the space below to explain your response or provide any comments you may have about the warning.
7) If you missed or ignored calls received from this study, please use the space below to explain why
8) How did you determine when to answer, ignore or decline a call? What steps did you take?
9) What advice if any do you have for the researchers about how they could improve the warnings you heard or saw during the study?

D. *Codebook from Field Study (Study 2)*

- warning present - Decision based on notification shown
- recognize number - Decision based on Caller ID info
- what user is doing - Decision based on their availability
- Accident - Decision made on accident
- Self-assessment - Participant used their knowledge
- Confused - Participant was not sure how to proceed
- n/a - Participant chose not to respond

E. Questions for MTurk Survey (Study 3)

1) How would you respond to this call? (Accept, Decline, Other)
2) What is the likelihood that this call is a spam call (likert scale)