User-Centered Risk Communication for Safer Browsing

Sanchari Das[1,2(✉)], Jacob Abbott[1], Shakthidhar Gopavaram[1], Jim Blythe[3], and L. Jean Camp[1]

[1] Indiana University Bloomington, Bloomington, USA
sancdas@iu.edu
[2] University of Denver, Denver, USA
[3] USC Information Sciences Institute, Los Angeles, USA

Abstract. Solutions to phishing have included training users, stand-alone warnings, and automatic blocking. We integrated personalized blocking, filtering, and alerts into a single holistic risk-management tool, which leverages simple metaphorical cartoons that function both as risk communication and controls for browser settings. We tested the tool in two experiments. The first experiment was a four-week naturalistic study where we examined the acceptability and usability of the tool. The experimental group was exposed to fewer risks in that they chose to run fewer scripts, disabled most iFrames, blocked Flash, decreased tracking, and quickly identified each newly encountered website as unfamiliar. Each week participants increased their tool use. Conversely, those in the control group expressed perceptions of lower risk, while enabling more potentially malicious processes. We then tested phishing resilience in the laboratory with newly recruited participants. The results showed that the tool significantly improved participants' ability to distinguish between legitimate and phishing sites.

Keywords: Phishing · Risk-communication · Mental models

1 Introduction

Phishing attacks are one of the most well-known cyber attacks. In 2017 alone, there was a loss of $678 million in the US due to phishing attacks [52]. In 2015, McAfee implemented an (admittedly commercial) study of 19,000 consumers and found that 97% of participants were unable to detect phishing emails [1]. Phishing also plays a critical role in the broader e-crime ecosystem, allowing for technically simple and low-cost intrusions [36]. Thus, defeating phishing remains a significant challenge for human-security interactions [14]. To assist people requiring protection, we created a browser extension that centered the human experience to help internet users incorrectly identifying phishing attacks. We constructed the front-end as an extension using cartoons, as shown in Fig. 1. The design was built on three core components. First, our tool design we assumed

M. Bernhard et al. (Eds.): FC 2020 Workshops, LNCS 12063, pp. 18–35, 2020.
https://doi.org/10.1007/978-3-030-54455-3_2

that participants do not care about the technical source of risk: the site, the ads, or the route. The second component is the recognition that there are trusted web sites for social, professional, and personal reasons that vary between individuals. The third component is that we integrated secure communication as a possible way forward, as a complement to stand-alone indicators. We integrated warnings and controls so that people had to actively and knowingly choose risk-taking. To report on the implications of the toolbar mentioned above, we report on two experiments that used the interaction. For the first experiment, we conducted a naturalistic in-situ study with 44 participants to evaluate the usability and acceptability of the tool. For the second experiment, we conducted an in-lab study with 45 participants to assess efficacy of the tool. Our research questions were:

RQ1: Is the toolbar understandable? Using interviews and qualitative analysis, we evaluated whether individuals accurately describe the implications of the interaction. If it was understood, would it be used? To answer this, we conducted a naturalistic study and interviewed participants about their perceptions of the toolbar.

RQ2: Is the holistic risk management tool acceptable? In the naturalistic setup, we observed how usage and perception of the tool improved during the study.

RQ3: Do participants who were completely unfamiliar with the tool illustrate greater efficacy in detecting phishing sites? We inspected efficacy in mitigating phishing. For this, we conducted a laboratory experiment where participants were interacting with the toolbar for the first time. We evaluated the participants in the experimental group to the same mock phishing sites as a control group.

RQ4: How do stress conditions impact the risk behavior of an individual while interacting with risk-mitigation tools? The final test, of efficacy under stress, was part of the same experiment. Specifically, we evaluated in a real-time in-lab experiment under two stress conditions to better align with the cognitive experience of actual phishing [29].

Our contributions are the use of risk communication to identify and mitigate aggregate risks in a single tool. The tool includes script blocking, embeds phishing indicators, certificate warnings, and provides notification of unencrypted communication. The second contribution is personalized web risk settings based on individual choices and browsing history. In other words, we let each person easily select their own unique set of favorite or most-used websites, deciding to take the risk but knowingly. We complement that by trivially distinguishing the familiar from the unfamiliar through targeted blocking. The third contribution is a design intended to provide security without information sharing, i.e., potential loss of privacy.

Fig. 1. Toolbar showing the low, medium, and high risk tolerance buttons

2 Related Work

In 1996 Zurko [56] coined the phrase "user-centered security". This work was informed by research in user-centered security specifically studies on warnings, usability studies of security tools, and research into user perspectives on security. Cranor and Garfinkle's classic text on usable security, for example, included evaluations on browser warnings [12]. In 2006, a study of five simple commercial toolbars found that none of them had any statistically significant impact [51]. Shortly after this high impact study, the focus moved away from this type of interaction as conventional wisdom began to focus on browser indicators and warnings. A comparison of six indicators across nine browsers resulted in a redesign of security indicators for Chrome [22]. However, despite noting the importance of trust-based risk communication tools and interactive [30] and trust ensuring tools [53], comparatively little work has been done in risk communication with few exceptions [32,39].

2.1 Security as Risk Communication

Risk communication depends on estimates of the underlying risk as well as subjects' mental models of the risks [10,15]. Asgharpour et al. [4] and Wash et al. [48] showed distinct differences in the mental models between experts and non-experts by analyzing simple mental models [9,31]. Mental models and risk perception differ between individuals, and the differences between experts and non-experts is a challenge addressed by security researchers who have collaborated with cognitive science researchers in implementing mental models [6,8,47]. Applying these models requires identifying the model of the specific user, which requires observing user choices and behaviors [43,50] or the inherent natures of the risks [25]. Perceived risk offline is driven by nine characteristics of the hazard [23]: 1) voluntariness, 2) immediacy, 3) knowledge to the exposed, 4) knowledge to experts, 5) control, 6) newness, 7) common-dread, 8) chronic-catastrophic, and 9) severity. Offline, this framework informed four decades of research in risk perception and public policy in a variety of risk domains, e.g., environmental risk [24], health risk [28]. Online, this framework has been used to explain perceptions of technical security risks [25] and insider threats [21]. Mental model research not only focuses on security and privacy but also implements user perception of environmental hazards by incorporating Human-Computer Interaction (HCI) methods [10].

2.2 Browser Warnings and Toolbars

Wu et al. [51] investigated the impact of the three toolbars [11,27,37] and concluded that toolbars do not work. However, it provided no generalized findings for the design of interactions. Felt and Weinberger examined how often a person should be alerted with a warning after the dismissal of an initial warning [49]. Patil et al. [38] recommended providing delayed feedback for non-privacy critical situations. Instead we endeavored to implement real time feedback through risk indicators with the assumption that only the user knows what is privacy critical to them.

2.3 Graphical Usage in Risk Communication

Visual differences including personalized security indicators [33,40] have been proven effective in detecting Phishing websites [34]. Zhang et al. used text, infographics, and a comic to educate participants on why updating anti-virus software is important [55], users expressed that they understood why it was important and while making decisions after the study, referenced the comic example for guidance [54]. Garg et al. explored the difference between the same script when presented as a video and presented as text in educating individuals on how to avoid being victimized by phishing [26]. They used the metaphor of a solicitor impersonating a banking investigator to leverage story-telling to educate older users. Wash found individual stories told by someone users could identify with to be a highly effective form of risk communication [48].

2.4 Usability and Adaptability

Building the tool is not enough, it must also be usable and acceptable [5,16,17,35]. Das et al. found that even technical experts do not adapt simple security tools if risk mitigation techniques are not communicated properly and if the benefits are unclear [13]. Thus, our goal was not only to build a usable and factually useful tool, but also one that communicated the risk mitigated by its use.

3 Prototype Design

Our tool focuses equally on ease of use and effective risk communication. The goal is to allow users to take a security risk only by making informed decisions with proper knowledge of the risk. The toolbar not only works on a local system but also remembers the user's choices and the context in which risks are acceptable, and minimizes risk in other contexts without storing it in the cloud. Our toolbar extension uses very simple metaphorical cartoons to indicate low, medium, and high-risk options. Figure 1 shows how the toolbar's buttons look. We instantiated the illustrations as buttons that control the browser settings while communicating the level of risk for a given connection with the selected parameters. We had three high-level contexts in the architecture (Web, Network, and User). The details of operation are described necessarily in other

publications. Here the focus is on the user experiment and results. To evaluate certificates and generate custom certificate warnings, we used a machine learning approach described by Dong et al. [19], which later expanded with Microsoft [18].

The risk of the network connection was evaluated by reading the network policy and assessing the use of encryption during transmission. Our assessment also included the evaluation of familiarity of Service Set Identifier (SSIDs) and familiarity of the IDs of devices connected to the same SSID for wireless. The assessment of risk above the network level was a combination of the domain name, certificate, and page elements, mainly scripts. Domain names were evaluated based on personal history with an initial default of the top million trusted. The domain name reputation system was grounded in the initial anti-phishing reputation system described in IBM Systems Journal [45]. These visited domains became trusted one week after the first visit or upon explicit user action. That one week window is grounded in reported take-down times from private conversations in the Anti-Phishing Working Group. We evaluated the Certificates using machine learning as detailed in the specific publication on that module [19]. We evaluated the running scripts on familiarity and source. Some familiar scripts were white-listed or grey-listed based on source (e.g., Google Analytics was enabled on Google). Other indicators in our prototype included personal history, checks for common vectors for malware (i.e., Flash, iFrames), and any script that indicated cross-site scripting. This analysis was too burdensome for real-time, and we substituted a lightweight version for the actual experiment reported here. The likelihood of warnings was grounded in the risk setting chosen by the user. The default was a medium risk. The interaction was chosen based on previous work on risk perception, to align user mental model and construct on previous work on cartoons, videos, and images as online risk communication [7,26,44,54].

4 Method: Naturalistic Study

For our experiment, we recruited 82 participants by posting flyers at the university and various places of worship. The outreach to places of worship was grounded in team social connections and could arguably be considered snowball sampling. The goal of this outreach was to have a diverse sample. All stages and work were reviewed and approved by the Institutional Review Board. The first step for participants was completing an initial interview and survey that consisted of basic demographics and expertise questions. Qualitative team members from the College of Arts & Sciences conducted the interviews. We specifically sought non-technical users for this study, so 53 participants were invited to participate in the second portion of the study; the remaining 29 participants were deemed to have too much computer and security knowledge to continue the experiment. We measured the participant's expertise by a combination of knowledge skills and behavior questions from Rajivan et al.'s work [41].

Out of the invited participants, 44 decided to partake in the month-long second phase and were randomly divided into two groups: experimental and control. Both the control and experimental groups brought their personal laptops to our research house. They were assisted in the installation of Mozilla Firefox

if they did not already have it installed, and the experimental extension from our technical team. No use instructions were initially given, excluding a brief installation video. The control group received a version of the extension that was designed not to interfere with their normal browsing and would only run in the background to perform risk calculation and logging usage data. The extension for the control group existed only as a data compilation tool for comparison with the experimental group. We gave the full extention to the experimental group. The default setting for each website, excluding those blacklisted, was set at medium for the experimental group on start. Participants could adjust their default ratings on the menu. Still, each new website visited would load at the selected default level until a participant changed the security rating by clicking on one of the three illustrations. After applying a new security level, the extension remembers the level given for each site and will load that on future visits.

We instructed the participants in both groups to use Firefox for their daily internet browsing. We also asked the participants not to use any other extensions during the experiment. Each participant returned once a week for four weeks for an hour session. They were paid $20 for each session. These sessions consisted of the participant being interviewed in one room while the technical team extracted their log data in another room. At the end of the four weeks, there was an exit interview and survey. We had 44 total participants complete the entire experiment, 23 in control, and 21 in the experimental group. We based the duration of the experiment in part on Anderson et al.'s work on habituation to security warnings [2]. The four week period was more extended than work by Vance et al., which combined recovery periods with functional magnetic resonance imaging (fMRI) examination of responses to warnings [46]. Their work indicated that habituation was evident within a week. Thus, our four-week experimental period should have been sufficient for any habituation effects to be apparent in the results.

5 Results: Naturalistic Study

In this section, we report on a four-week naturalistic study, which includes the interviews and the modifications of the secure browsing behavior of 44 participants. In a four week experiment, we monitored participants' practices as well as self-reported perceptions of their actions. Participants in the experimental group chose fewer online risks than those in the control group.

Interview data and computer logs were collected every week for four weeks from all participants. Crowd workers transcribed the audio files at TranscribeMe![1]. We used the online qualitative data analysis service Dedoose[2] to code the data and provide a first pass at the analysis. A team of researchers developed the original codes by examining the transcribed responses to the most relevant questions for this study. Two researchers coded small sections of transcripts until

[1] https://transcribeme.com/.
[2] http://www.dedoose.com/.

they achieved an inter-rater reliability score above 0.80 and then proceeded to code the remaining 200 transcripts. We asked the participants to use Firefox with the tool enabled for at least six hours per week. Users reported time with the tool fluctuated throughout the study, with 35% reporting that they used the tool for 0–9 h in the first week. By the third week, 33% reported minimal tool use, i.e., 0–9 h. By week 4, 26% reported using the tool 0–9 h; 44% used it 10–14 h, and 22% used it more. Our data collection validated these reports, which proves the tool use increased over time rather than decreasing.

Recall that, the tool accepts the settings for a second-level domain and applies that setting. The result is that there is less interaction with the toolbar over time, as an increasing number of sites will be set according to the user's preference because the websites the user visits will have been increasingly configured. The extension's most visible activity was blocking scripts that could contain malicious content. If participants clicked on the image of the pigs in the brick house, then the tool blocked large sections of advertisements, images, and videos (Low risk, high-security settings). If they clicked on the icon of the pigs in the straw house, then the tool blocked only items on the blacklist (High risk, low-security settings). In practice, this meant that the high risk, straw house, rating blocked almost nothing. Individual participants' answers to "Based on your interaction with the tool last week, what do you think the tool does?" ranged from accurate to erroneous, even in a single session. At some point in the four weeks, 88% of all participants reported accurately that the "tool blocks (removes/hides) things based on the security settings". Over half of this group also incorrectly stated that the tool provided anti-virus protection. Participants expressed their perceptions of convenience versus security and efficiency versus security, as well as wanting particular content and realizing there was a security issue. "I felt like the piggy in the brick wall. My computer was safer thanks to the tool, but there's a battle going on between security and convenience" stated one participant. The same participant then said about the high-risk setting, "The one it's currently on is its easiest setting and allows the website to work very efficiently". It is hard to judge perceptions on 'efficiency' except that the page would appear reasonable to them. Two users did report switching to the lowest security setting to speed up their computer. No participant singled out security versus privacy.

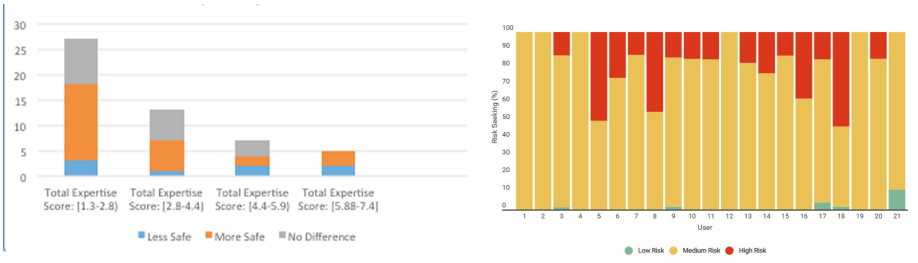

Fig. 2. Increased security perception for participants (left). Level of risk that each user chose during their fourth week of use (right).

Overall, 83% of participant responses indicated that they felt the pictures were effective as a tool for communicating computer security. Only two people said that they would have preferred words to pictures. One of those two felt it was too simple indicated, but that it would work for others: "I think it's good. I think I'm a pretty savvy internet user, it's a big part of my job and so... um, it's very easy, and it makes it very quick to notice, and I kept thinking this would probably be really good for like, my mom, who doesn't know as much". We show a more detailed breakdown of the participants'responses in Fig. 2. This comment may reflect not only ease of use, but also the fact that individuals are better at identifying rational responses to risk for others than to themselves [42]. The primary objection to the tool was that it included warnings, particularly password reuse warnings. The password warning for unsafe use was the only warning difficult to disable and triggered when a password was being sent over an unencrypted link or unprotected wireless connection. There would not be a technical solution at the browser to mitigate such a problem unless Tor or a VPN were integrated into the extension. Every other warning allowed individuals to reset their risk setting before moving forward and indicated that the person could go forward. We also inquired with the control group about the functionality of the tool. For the control group, the extension only logged their browsing activity and calculated the degree of risk for a given page. It was natural for the majority of the control group to respond that the tool gathers/tracks Internet browsing data. Only five people said otherwise, either believing that the tool was designed to track advertisements or that the tool was some form of anti-virus or malware protection. Three people reported that the tool was designed to change the computer speed, as some people reported issues with their computer operating noticeably slower.

5.1 Understanding the Tool

Participants largely understood the meaning of the pictures that conveyed their level of exposure to potential threats on webpages as a function of their own manipulated tool settings. There was some confusion between risk and protection as the lower security level represented a higher risk. The example below portrays a typical response where the confusion is evident; however, the participant's perception is on-point:

Interviewer: "This is Picture B. Can you tell me what this means?"

Participant: "Big bad wolf. This is the medium setting. Again, with what I originally thought the software would do, and these pictures... what they are, what they represent don't really line up to me. Cuz, it's not like anti-virus software. These pictures, to me, make me think, it's going to moderately protect my computer against certain websites that could be dangerous. But that's not really what it does. It just tells me whether it's safe or not, and it blocks some pictures. From what I can discern, ascertain. I don't know".

5.2 Changing Tool Risk Levels

10 of 25 experimental participants reported keeping the security setting on the lowest level the entire time. As the control group, the experimental group perceived their risk as more moderate than it was, as the graph of time spent at each level illustrates in Fig. 2b. 20 of the 25 experimental users reported reducing the settings at some point during the study period. Five said it only once, two in the first week, and three in the third. Reports of reducing the settings were consistent throughout the study. Participants generally wanted to see all of the content on the website or needed to reduce the settings to get the functionality from the site that they desired. There were more changes in risk levels than reported. By the final week, some participants reported not having to change the setting. The design goal was to make the tool highly usable. Therefore part of the customization was storing the participant's choice for a site, so it was not necessary to change the settings on return visits. Participants offered various reasons for changing the risk setting. One decreased security when the default was placed on medium for trusted sites, expressing this as, "Uh, I turned it on no security whenever it automatically bumped itself up to medium". A second participant also explained that decreasing security was needed to access content, "Most of the time, I would keep it on medium setting. That's always good. But if there's something like, if I needed to watch a video, I was like – I would go to Sports Center, and if I wanted to watch a video, I would have to put it on the low setting to watch some of the videos". A third participant explained, "On a site, like Reddit or a news – any site where if I click something and it takes me somewhere else - a site that redirects you - I would tend to put it on medium maybe more because I don't think I'm staying in the same place that I know is safe".

Eight people reported changing the setting to decrease risk, sometimes to hide advertisements (two participants), but the primary reason was playful exploration. Only three participants reported wanting to increase their security with the tool. Two of these three were in the lowest expertise score range. A total of 13 people reported simply playing with the tool. The most often mentioned benefit was ad-blocking functionality. In addition to the perceptions of changes, we examined how often there were changes. We evaluated how often a participant's browsing switched between high, medium, low-risk settings across different websites. We show the results for the last week in Fig. 2. This graph is only for the participants that continued the experiment through the fourth week. While some users chose to be at high risk, most users spent the majority of the time at medium risk. We also noticed that users chose higher risk settings when surfing social media sites; note that the tool at the lowest risk setting blocks almost all functionality of such sites. The extension defaulted to the medium level of risk whenever a user visited a new website, thus introducing protection from potentially malicious scripts and allowing the user to opt for increased or decreased protection. Not shockingly, defaults are powerful even when easy to change. One way of evaluating the graph above is that participants embraced the default setting most of the time.

5.3 Warnings

The following quotes represent how one user felt about password notifications. These findings point to the fact that people not only would not change their passwords but found the notifications about password security to be an annoyance.

Participant Week 1: "With the warnings about the passwords, there's no option to shut those notifications off. As it is with almost every security thing, it's like, 'Well, if you want to keep being reminded of this every time then"'.

The other warnings were click-through and allowed risk level changes. The warning which was explicitly mentioned as problematic was the password warning.

Participant Week 2: "So, when it gives you the, 'You've used this password before,' there's got to be a checkbox for, 'Stop reminding of this.' So, that made it worse. That's pretty much it".

None of the warnings could be disabled, but the other warnings were not subject to complaints.

6 Method: In-Lab Experiment

For the follow-up study, we conducted an in-lab experiment with 45 participants. The second phase of the study was reviewed and approved by the Institutional Review Board as well. For this phase of the study, we partially implemented the study design implemented in an eye-tracking survey for security indicators by Majid et al., where the secure browsing indicators were added to the browsing experience of the users [3]. Out of the 45 participants, nine were female, and 36 were male. Ten participants were within 18–21 (inclusive) years old, 30 were between 22–25 years old, and five were between 26–30. Twenty-four participants were undergraduate students, and 21 participants were graduate students recruited from a non-technical security course at the university. We mainly chose a younger crowd to test the usability, acceptability, and efficacy of the tool. We provided the participants with a verbal recruitment script, which explained the in-lab experiment.

We experimented in the university's computer lab, where the participants used either their personal laptop or the lab's computer where Mozilla Firefox was installed, and it was mandatory for the participants to use Mozilla Firefox for the integration of the experimental toolbar. After providing the verbal recruitment script, we informed the participants that their participation would yield an accepted payment of $2.00. After that, we randomly assigned them to a group that decided their bonus pay, which could be anywhere between $0–$8.00. We provided the students with a link which placed them in a randomly assigned group. The experiment included eight different conditions across two penalty stress conditions and four experimental presentation groups. We based the two different stress conditions by showing a time penalty for incorrect selections or a

deduction in payment. The remaining presentation conditions included the control group which showed sites without our experimental toolbar, the low-risk tolerance group which presented sites through the lens of the low-risk high-security setting, the medium risk tolerance which showed places with the medium risk medium security setting, and finally the high-risk tolerance group that presented sites with the toolbar on the high-risk, low-security environment.

After the assignment of the random conditions, the participants went through the pre-screening questions, where we asked them about their age, nationality, and native language. The experiment included only participants who were more than 18 years old, lived in the US to remove cultural biases, and also to facilitate the location restriction of the in-lab study, and could read and write in English. The experimental setup provided each participant with 26 individual website images, which were randomly sorted into spoofed and non-spoofed versions of the website. If the participants trusted the site, they clicked the login or the sign-in button. If the participants didn't trust the website, then they could click the back button. If the participants clicked login for a bad site, then the error message "Clicked on a Bad Site" popped up. If the participants clicked the back button on a legitimate website, then the error message "Did not click on a good site" popped up. For a successful click, the experiment setup directed the participant to the next website. The participants in the time penalty, for every incorrect click, got a sentence of 15 seconds and could not proceed further until the penalty period ended. The timer was on during their selection of the website. Thus, we ensured the timer created the required stress condition. We penalized the other group with the bonus deduction with \$0.67 from the \$8.00 allotted for the max bonus pay on incorrect selections. Thus, though the time was not a stress condition here, the wrong choice still yielded them the loss of the bonus pay. After explaining the entire procedure, we also asked the participants questions to check their understanding to ensure that they correctly understood the process as the whole. After they correctly answered the whole question set, we directed the participants to their respective set of websites. After they went through the experiment, the participants answered some computer knowledge, expertise, and behavioral questions.

7 Results: In-Lab Experiment

We also report on the in-lab study, with a different set of 45 participants on the usability, acceptability, and efficacy of the implementation of the toolbar extension. We calculated how participants behaved between the toolbar and control groups and found significant improvement ($W = 18, p = 0.005$) in detecting Phishing websites when we compared the Toolbar Low-Risk High-Security Option with that of the control group. We also tested the stress conditions (Money versus Time) in our experiment to analyze how stress creates risk behavior changes. However, we were unable to find any significant differences between the two sets of participants while evaluating their accuracy ratings ($p = 0.8$). Thus we cannot conclude that the difference in stress conditions created much

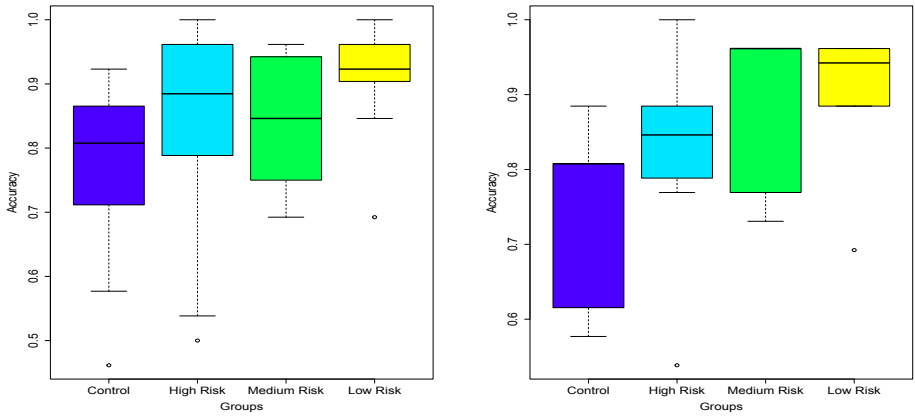

Fig. 3. Box plot of distribution of accuracy (left) and time as stress condition (right).

difference. However, we found significant results when we compared the control with the stress condition as Time and the participants who received toolbar set as low-risk tolerance with the Time condition ($p = 0.04$). Figure 3 shows the significant improvement in the accuracy of the participants who used the toolbar as compared to the participants in the control group. The participants' familiarity with the websites impacts the how the toolbar behaves. The tool modifies the interface based on whether the users are interacting with the toolbar for the first time. Therefore, to evaluate the efficacy of such a mechanism, it was critical to capture the website familiarty score for the participants. We ran a regression where the accuracy was the dependent variable, and the familiarity score was the independent value. We found a positive correlation of the accuracy of the participants with the familiarity ($r = 0.45$), and the correlation of the accuracy with the familiarity was statistically significant ($p = 0.02$). Figure 4 shows the scatter plot of the accuracy of predicting the websites correctly based on their familiarity with the website.

8 Discussion

The results of the four-week test showed that people would change their risk exposure if it is simple to do so. Significant changes in risk exposure online at the individual level, aggregated over users, creates a decrease in exposure. It also illustrated that people did not necessarily feel that they were changing their behaviors. Although the changes in risk level continued over four weeks, the reported changes in risk level decreased. Our optimistic perspective is that this implied that changing the risk level became significantly relaxed as not to be remembered. The result of the in-lab phishing study is that the tool easily distinguished between familiar and unfamiliar sites. Currently, it is trivially easy to implement a phishing site by copying the code and functionality of a legitimate website. The extensive blocking would make such attacks much more difficult.

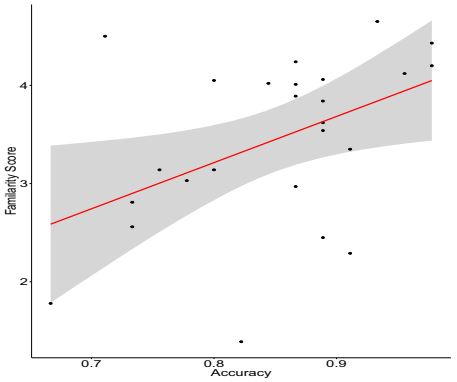

Fig. 4. Scatter plot showing accuracy of the participants in correspondence to their familiarity level with the website

The ease of overriding the blocking makes the defense acceptable. The control group expressed high levels of confidence in their safety and practices. This is not uncommon in computer security and privacy risks. Wash found that vulnerable participants whose mental models of computer security were aligned with unsafe behaviors felt quite safe online [48]. In a study of two-factor authentication, individuals rejected additional security on the basis that they thought their own passwords were secure [20]. Our instrumentation could only measure when Firefox was in use; therefore, if the participants changed browsers, then the data would not be included. If this experimental prototype were to be used in production, the measures of risk would need to be more exacting. The domain names and certificate results were consistently reliable, as shown by the previous studies where the website certificates and domain names were studied [19,45]. The use of PhishTank as a source of blacklisting was feasible only because the use of the toolbar was limited; a dedicated site would be needed in production. Instead of the rough identification for scripts, advanced collaborative script detection should be used. The computational complexity of measuring the risks of scripts is the reason we used white lists and blacklists, despite the relevant preexisting research on that topic.

9 Limitations and Future Work

To preserve the privacy of participants, we did not record the specific addresses of visited sites by a particular person or group. They were used solely on the back-end of the system in the naturalistic study. As a result, specific traffic data by group or person was intentionally not compiled. The in-lab study had a limited number of people tested under two stress conditions; given the lab setting, the generalizability of observed phishing resilience may be restricted. The components of phishing resilience are an area where additional cross-cultural studies would inform our results. This research builds on decades of findings

surrounding risk perception, particularly the perception of online risks. Previous studies reported our work on understanding, estimation, and interaction with privacy risk. The default selected for the naturalistic testing of the prototype was a medium risk. The uncertainty in calculations of the security and privacy risks reifies the importance of defaults. Future work could also include bundling the toolbar with anti-virus programs, mainly as many participants believed this was already the case. Ideally, such an interaction could be bundled with other privacy-protecting systems; Tor could be an ideal candidate. A larger-scale phishing identification experiment with a more diverse population is an additional possibility for future work.

10 Conclusion

As threat detection and information technology become more complex, non-technical people who are already overwhelmed cannot be expected to manage this complexity. These two trends, increasingly sophisticated threats and increasing technical heterogeneity in the user population, have been treated as if they exist in opposition. Through our toolbar implementation and user studies (naturalistic and in-lab study) have shown that these issues can be well-aligned by combining risk communication and usable interactions with a complex, personalized back-end. In the naturalistic experiment, we found that those with the toolbar extension took fewer risks and were more aware of online risk than the control group. Participants in our in-lab experiment showed that using the toolbar extension significantly increased their accuracy in detecting spoofed websites. They determined the acceptability of risk given their history and contexts. Usability, cognitive misalignment, or incentive misalignment have all been presented as underlying the vulnerability to phishing. Among security professionals beyond the usability community, it is common to hear of the "dancing pigs" problem, where *"Given a choice between dancing pigs and security, users will pick dancing pigs every time"*. The challenge to security is framed as security awareness, where users must engage in constant vigilance. Universal constant vigilance and technical excellence is not a reasonable expectation. Our work illustrates that when people are provided clear risk communication and empowered to avoid risks, they do so. Technology needs to provide the right risk communication, at the right time, in the right context, aligned with user mental models and risk perceptions.

Acknowledgement. This paper is dedicated to the memory of programming staff Tom Denning. We want to acknowledge the substantive contributions of Mike D'Arcy as well as Timothy Kelley. We acknowledge the contributions of Jill Minor in substantive editing. This research was sponsored by DHS N66001-12-C-0137, Cisco Research 591000, and Google Privacy & Security Focused Research. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies or views, either expressed or implied, of the DHS, ARL, Google, Cisco, IU, or the US Government. We also want to acknowledge contributors to the experiment itself at Indiana University, including Deborah Taylor, Prashanth Rajivan, and Krishna C. Bathina.

References

1. 97% of people globally unable to correctly identify phishing emails, May 2015. https://www.businesswire.com/news/home/20150512005245/en/97-People-Globally-Unable-Correctly-Identify-Phishing
2. Anderson, B.B., Kirwan, C.B., Jenkins, J.L., Eargle, D., Howard, S., Vance, A.: How polymorphic warnings reduce habituation in the brain: insights from an FMRI study. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 2883–2892. ACM (2015)
3. Arianezhad, M., Camp, L.J., Kelley, T., Stebila, D.: Comparative eye tracking of experts and novices in web single sign-on. In: Proceedings of the Third ACM Conference on Data and Application Security and Privacy, pp. 105–116. ACM (2013)
4. Asgharpour, F., Liu, D., Camp, L.J.: Mental models of security risks. In: Dietrich, S., Dhamija, R. (eds.) FC 2007. LNCS, vol. 4886, pp. 367–377. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-77366-5_34
5. Assal, H., Chiasson, S.: Will this onion make you cry? A usability study of tor-enabled mobile apps. In: Poster presented at the 10th Symposium on Usable Privacy and Security (SOUPS) (2014)
6. Bartsch, S., Volkamer, M., Cased, T.: Effectively communicate risks for diverse users: a mental-models approach for individualized security interventions. In: GI-Jahrestagung, pp. 1971–1984 (2013)
7. Benton, K., Camp, L.J., Garg, V.: Studying the effectiveness of android application permissions requests. In: 2013 IEEE International Conference on Pervasive Computing and Communications Workshops (PERCOM Workshops), pp. 291–296. IEEE (2013)
8. Blythe, J., Camp, L.J.: Implementing mental models. In: 2012 IEEE Symposium on Security and Privacy Workshops (SPW), pp. 86–90. IEEE (2012)
9. Bravo-Lillo, C., Cranor, L.F., Downs, J., Komanduri, S.: Bridging the gap in computer security warnings: a mental model approach. IEEE Secur. Priv. **2**, 18–26 (2010)
10. Camp, L.J.: Mental models of privacy and security. Available at SSRN 922735 (2006)
11. CoreStreet: Spoofstick (2004). http://www.corestreet.com/spoofstick/
12. Cranor, L.F., Garfinkel, S.: Security and Usability: Designing Secure Systems that People can Use. O'Reilly Media, Inc., Sebastopol (2005)
13. Das, S., Dingman, A., Camp, L.J.: Why Johnny doesn't use two factor a two-phase usability study of the FIDO U2F security key. In: Meiklejohn, S., Sako, K. (eds.) FC 2018. LNCS, vol. 10957, pp. 160–179. Springer, Heidelberg (2018). https://doi.org/10.1007/978-3-662-58387-6_9
14. Das, S., Kim, A., Tingle, Z., Nippert-Eng, C.: All about phishing: exploring user research through a systematic literature review. arXiv preprint arXiv:1908.05897 (2019)
15. Das, S., Kim, D., Kelley, T., Camp, L.J.: Grifting in the digital age (2018)
16. Das, S., Wang, B., Camp, L.J.: MFA is a waste of time! understanding negative connotation towards MFA applications via user generated content. In: Proceedings of the Thriteenth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2019) (2019)
17. Das, S., Wang, B., Tingle, Z., Camp, L.J.: Evaluating user perception of multi-factor authentication: a systematic review. arXiv preprint arXiv:1908.05901 (2019)

18. Dong, Z., Kane, K., Camp, L.J.: Detection of rogue certificates from trusted certificate authorities using deep neural networks. ACM Trans. Priv. Secur. (TOPS) **19**(2), 5 (2016)
19. Dong, Z., Kapadia, A., Blythe, J., Camp, L.J.: Beyond the lock icon: real-time detection of phishing websites using public key certificates. In: 2015 APWG Symposium on Electronic Crime Research (eCrime), pp. 1–12. IEEE (2015)
20. Fagan, M., Khan, M.M.H.: Why do they do what they do?: A study of what motivates users to (not) follow computer security advice. In: Twelfth Symposium on Usable Privacy and Security (SOUPS 2016), pp. 59–75 (2016)
21. Farahmand, F., Spafford, E.H.: Understanding insiders: an analysis of risk-taking behavior. Inf. Syst. Front. **15**(1), 5–15 (2013). https://doi.org/10.1007/s10796-010-9265-x
22. Felt, A.P., et al.: Rethinking connection security indicators. In: SOUPS, pp. 1–14 (2016)
23. Fischhoff, B., Slovic, P., Lichtenstein, S., Read, S., Combs, B.: How safe is safe enough? A psychometric study of attitudes towards technological risks and benefits. Policy Sci. **9**(2), 127–152 (1978). https://doi.org/10.1007/BF00143739
24. Flynn, J., Slovic, P., Mertz, C.K.: Gender, race, and perception of environmental health risks. Risk Anal. **14**(6), 1101–1108 (1994)
25. Garg, V., Camp, J.: End user perception of online risk under uncertainty. In: 2012 45th Hawaii International Conference on System Science (HICSS), pp. 3278–3287. IEEE (2012)
26. Garg, V., Camp, L.J., Connelly, K., Lorenzen-Huber, L.: Risk communication design: video vs. text. In: Fischer-Hübner, S., Wright, M. (eds.) PETS 2012. LNCS, vol. 7384, pp. 279–298. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-31680-7_15
27. Herzberg, A., Gbara, A.: Trustbar: protecting (Even Naive) web users from spoofing and phishing attacks. Technical report, Cryptology ePrint Archive, Report 2004/155 (2004). http://eprint.iacr.org/2004/155
28. Johnson, B.B., Slovic, P.: Presenting uncertainty in health risk assessment: initial studies of its effects on risk perception and trust. Risk Anal. **15**(4), 485–494 (1995)
29. Kelley, T., Amon, M.J., Bertenthal, B.I.: Statistical models for predicting threat detection from human behavior. Front. Psychol. **9**, 466 (2018)
30. Likarish, P., Dunbar, D.E., Hourcade, J.P., Jung, E.: Bayeshield: conversational anti-phishing user interface. In: SOUPS, vol. 9, p. 1 (2009)
31. Lin, J., Amini, S., Hong, J.I., Sadeh, N., Lindqvist, J., Zhang, J.: Expectation and purpose: understanding users' mental models of mobile app privacy through crowdsourcing. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing, pp. 501–510. ACM (2012)
32. Marchal, S., Asokan, N.: On designing and evaluating phishing webpage detection techniques for the real world. In: 11th USENIX Workshop on Cyber Security Experimentation and Test (CSET 2018). USENIX Association (2018)
33. Marforio, C., Jayaram Masti, R., Soriente, C., Kostiainen, K., Čapkun, S.: Evaluation of personalized security indicators as an anti-phishing mechanism for smartphone applications. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp. 540–551. ACM (2016)
34. Maurer, M.E., Herzner, D.: Using visual website similarity for phishing detection and reporting. In: CHI 2012 Extended Abstracts on Human Factors in Computing Systems, pp. 1625–1630. ACM (2012)

35. McCune, J.M., Perrig, A., Reiter, M.K.: Bump in the ether: a framework for securing sensitive user input. In: Proceedings of the Annual Conference on USENIX 2006 Annual Technical Conference, p. 17. USENIX Association (2006)
36. Moore, T., Clayton, R.: The impact of public information on phishing attack and defense (2011)
37. Netcraft: Netcraft toolbar (2004). http://toolbar.netcraft.com/
38. Patil, S., Hoyle, R., Schlegel, R., Kapadia, A., Lee, A.J.: Interrupt now or inform later?: comparing immediate and delayed privacy feedback. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pp. 1415–1418. ACM (2015)
39. Patrick, A.: Ecological validity in studies of security and human behaviour. In: SOUPS (2009)
40. Raja, F., Hawkey, K., Hsu, S., Wang, K.L., Beznosov, K.: Promoting a physical security mental model for personal firewall warnings. In: CHI 2011 Extended Abstracts on Human Factors in Computing Systems, pp. 1585–1590. ACM (2011)
41. Rajivan, P., Moriano, P., Kelley, T., Camp, L.J.: Factors in an end-usersecurity expertise instrument. Inf. Comput. Secur. **25**(2), 190–205 (2017)
42. Slovic, P., Finucane, M.L., Peters, E., MacGregor, D.G.: Risk as analysis and risk as feelings: some thoughts about affect, reason, risk, and rationality. Risk Anal. **24**(2), 311–322 (2004)
43. Stanton, J.M., Stam, K.R., Mastrangelo, P., Jolton, J.: Analysis of end user security behaviors. Comput. Secur. **24**(2), 124–133 (2005)
44. Tsai, J.Y., Egelman, S., Cranor, L., Acquisti, A.: The effect of online privacy information on purchasing behavior: an experimental study. Inf. Syst. Res. **22**(2), 254–268 (2011)
45. Tsow, A., Viecco, C., Camp, L.J.: Privacy-aware architecture for sharing web histories. IBM Syst. J. **3**, 5–13 (2007)
46. Vance, A., Kirwan, B., Bjorn, D., Jenkins, J., Anderson, B.B.: What do we really know about how habituation to warnings occurs over time?: A longitudinal FMRI study of habituation and polymorphic warnings. In: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, pp. 2215–2227. ACM (2017)
47. Volkamer, M., Renaud, K.: Mental Models – general introduction and review of their application to human-centred security. In: Fischlin, M., Katzenbeisser, S. (eds.) Number Theory and Cryptography. LNCS, vol. 8260, pp. 255–280. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-42001-6_18
48. Wash, R.: Folk models of home computer security. In: Proceedings of the Sixth Symposium on Usable Privacy and Security, p. 11. ACM (2010)
49. Weinberger, J., Felt, A.P.: A week to remember: the impact of browser warning storage policies. In: Symposium on Usable Privacy and Security (2016)
50. Workman, M., Bommer, W.H., Straub, D.: Security lapses and the omission of information security measures: a threat control model and empirical test. Comput. Hum. Behav. **24**(6), 2799–2816 (2008)
51. Wu, M., Miller, R.C., Garfinkel, S.L.: Do security toolbars actually prevent phishing attacks? In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 601–610. ACM (2006)
52. Yakowicz, W.: The 3 biggest phishing scams of 2018, July 2018. https://www.inc.com/will-yakowicz/biggest-email-phishing-scams-2018.html
53. Yee, K.P.: Designing and evaluating a petname anti-phishing tool. In: Poster presented at Symposium on usable Privacy and Security (SOUPS), pp. 6–8. Citeseer (2005)

54. Zhang-Kennedy, L., Chiasson, S.: Using comics to teach users about mobile online privacy. Technical report, Technical Report TR-14-02, School of Computer Science, Carleton University, Ottawa, Canada (2014)
55. Zhang-Kennedy, L., Chiasson, S., Biddle, R.: Stop clicking on "Update Later": persuading users they need up-to-date antivirus protection. In: Spagnolli, A., Chittaro, L., Gamberini, L. (eds.) PERSUASIVE 2014. LNCS, vol. 8462, pp. 302–322. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-07127-5_27
56. Zurko, M.E., Simon, R.T.: User-centered security. In: Proceedings of the 1996 Workshop on New Security Paradigms, pp. 27–33. ACM (1996)