Symposium on Digital Behaviour Interventions for Cyber-Security

In conjunction with the 2018 Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour (AISB 2018)

5th April 2018

Preface

This volume contains the papers presented at AISB Symposium on Digital Behaviour Intervention for Cyber Security held on April 5, 2018 in Liverpool. Each submission was reviewed by, on the average, 3 program committee members. The committee decided to accept 6 papers.

This symposium focuses on how digital technology can motivate and influence people to behave more cyber-securely. It brings together researchers, designers, developers and cyber-security experts interested in computers designed to change cyber-security attitudes and behaviours. The symposium covers a wide range of topics on persuasion, from behaviour intervention methods to persuasive argumentation and persuasive user interfaces.

Digital behaviour interventions have a great practical potential. They have been applied in many domains, for instance to improve health (encouraging a reduction in alcohol intake, smoking cessation, an increase in exercise, more healthy eating, and adherence to medical treatment) and to move towards sustainable living (encouraging a reduction in energy consumption, recycling, and use of public transport). There has been much progress in the research community on digital behaviour interventions, as shown for example by the successful Persuasive conference series, a special issue of the UMUAI journal, and a successful series of workshops on Computational Models of Natural Argument (an area overlapping with persuasion). There has also been a lot of interest in cyber-security behaviour, and policy compliance. However, most of this work has not been rooted within the behaviour change literature. There is currently an emergence of work that is beginning to combine these two strands of research, and this symposium helps to further build this community

In addition to presentations by participants, the symposium also includes discussions in smaller groups on topics related to digital behaviour intervention for cyber security, and a keynote presentation by an invited speaker.

March 20, 2018
Aberdeen

Judith Masthoff
Matthew Collinson
JP Vargheese

Organizing Committee

Matthew Collinson	University of Aberdeen
Judith Masthoff	University of Aberdeen
John Paul Vargheese	University of Aberdeen

Program Committee

Kovila Coopamootoo	Newcastle University
Tom Crick	Swansea University
Jaap Ham	Eindhoven University of Technology
Roeland Kegel	University of Twente
Shujun Li	University of Kent
Chenghua Lin	University of Trento
Fabio Massacci	University of Aberdeen
Ganna Pogrebna	University of Birmingham
Karen Renaud	Abertay University
Aneesha Sethi	University of Southampton
Alice Toniolo	University of St Andrews
Manfred Tscheligi	University of Salzburg

Additional Reviewers

Phyllis Nwadike

Table of Contents

Faheem: Explaining URLs to people using a Slack bot

Kholoud Althobaiti[§,†], **Kami Vaniea**[§], and **Serena Zheng**[‡]

k.althobaiti@sms.ed.ac.uk, kvaniea@inf.ed.ac.uk, serenaz@princeton.edu

[§]University of Edinburgh, Edinburgh, UK

[†]Taif University, Taif, KSA

[‡]Princeton University, Princeton, New Jersey, USA

ABSTRACT

Online safety regularly depends on users' ability to know either where a URL is likely to lead or identify when they are on a site other than they expect. Unfortunately, the combination of low URL reading ability in the general population and the use of hard-to-detect approaches like look-alike letters makes the reading of URLs quite challenging for people. We design a Slack bot, named Faheem, which assists users in identifying potentially fraudulent URLs while also teaching them about URL reading and common malicious tactics. In this work, we describe the design of the bot and provide an initial evaluation. We find that Faheem does a good job of interactively helping users identify issues with URLs, but Faheem users show minimal retention of knowledge when they lose access to the tool.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; K.6.5. Management of Computing and Information Systems: Security and Protection

Author Keywords

Phishing; usable privacy and security; real-time learning; security education

INTRODUCTION

Uniform Resource Locators (URLs) are how the majority of internet citizens find information on the world wide web. "Linking" between web pages, chat messages, social media, or even emails is a common method of telling someone else how to find a piece of content. When asked to visit a physical space in the real world using a provided address, most people are able to pull up a map in advance which allows them to answer important questions like: "How far away is it?" or "Does Google Maps think that there really is an Office Depot there?" But with an online URL, people seem to have more difficulty asking and answering basic questions about the location they are visiting, for example: "Is this really the website for Office Depot?" or "Will my password be sent to the website safely so no one else can read it?"

The goal of Faheem is to help people understand the content of URLs so that they can ask and answer questions about the URL, in particular, where it leads.

There are various reasons why understanding URLs can be useful, ranging from avoiding being Rickrolled to being able to identify when personal information is being sent in the URL. Phishing is likely the most financially impactful use case. Phishing attacks involve scammers attempting to obtain users' sensitive information for malicious reasons, with the individuals behind such attacks seeking to deceive users into visiting websites that impersonate legitimate ones [17]. One of the many reasons phishing works is that users cannot accurately read a URL to determine if it really is associated with an organization they interact with or not [8, 25].

Phishing is also quite expensive, costing the United Kingdom (UK) economy as much as £280 million a year [6]. Only about 72% of consumers in the UK even know what "phishing" is even though 92% of organizations report training users to identify and avoid phishing attacks [3]. Which is wise, since 98% of attacks involving a social element use phishing [2].

With the evolution of social media, instant messaging services, such as Slack and WhatsApp messengers, have become the main communication means between friends, relatives and colleagues [13]. These services allow end users to share links and files. However, on the heels of the adoption of such features, phishing on these new channels has become a threat [26]. More specifically, the manipulation of URLs is a popular phishing approach [11] which takes advantages of people's vulnerabilities when interacting with technology, and the characteristics of URLs, which makes it difficult for users to interpret them correctly in order to distinguish legitimate websites from those that are spoofed [21]. For example, URLs are read both left to right (path) and right to left (domain), URLs can be shortened, or URLs can be represented by an IP address, all of which is confusing for users [27].

We present here a new Slack chatbot called Faheem which helps users by parsing URLs for them and explaining the URL elements in a user-friendly way. The goal of Faheem is to both assist a user during a regular chat communication, and help them learn useful URL reading techniques. Ideally, a more polished version of Faheem could be installed on a company public Slack channel to provide contextual information about

AISB'18, April 4–6, 2018, Liverpool, UK

the various URLs being shared and protect employees from erroneously visiting problematic URLs.

We test Faheem against a simplistic URL explanation web page – called URL Explainer – and find that Faheem does a better job of supporting people interactively as well as helping them to retain the knowledge.

BACKGROUND AND RELATED WORK

Uniform Resource Locators (URL)

A URL is a structured description of the location of a digital resource [5] as pictured in Figure 1. Note that URLs do not state where the resource is actually located, merely how to go about locating it, the difference is subtle but a key to understanding some of the design decisions. For example, URLs can contain usernames and passwords, this information is necessary to locate resources behind a login prompt, but strictly speaking, has nothing to do with the actual location of the resource. Similarly, the query string exists so that the requester can pass strings to the host computer and get back the desired resource. Again, query strings help locate the resource in places like databases but do not strictly describe the actual resource location. When reading URLs the distinction becomes important because some URLs are actually the location of a second URL (redirection), such that a basic reading is insufficient to learn the final destination.

At a high level, end-user issues around URL readability focus on: 1) who or what is being communicated with (host, port), 2) what is being said (username, password, path, query string), and 3) how it is being said (protocol). Who is being communicated with is typically a fraudulent communication (phishing) issue where the malicious actor is trying to trick the user into going to the wrong host. What is being said is typically a privacy issue where more information is communicated than the user would like, such as communicating unique marketing IDs via query strings which can include anything the URL creator wants. For example, the HealthCare.gov site which is used by United States citizens up for health care allows users to click on links which take them to different private insurer websites. Those links were found to include information like pregnancy status in the URL query string, effectively sending sensitive data from inside of HealthCare.gov to a private insurer the user had no current relationship with [22]. This behaviour is insecure as the query could be saved in server logs and the browser's history log, which is a potential confidentiality breach [28]. The last issue is about how the information is said which is typically an issue of encryption (http vs. https). In this paper, the primary focus of Faheem is to raise users awareness of the who issue, notably, the phishing techniques.

URL Manipulation Tricks

Phishers will often use URL manipulation approaches to make the URLs they send people look legitimate and deceive the victim into believing they are visiting a trusted website [11]. The following are a set of common tactics used to hide the malicious destination of a URL [14, 27]:

- **Obfuscate:** The company name is not visible in the URL, which could be owing to the use of the IP address in a hostname part, or shortened or redirected links.
- **Mislead:** The expected company name is embedded somewhere in the URL where the user can see it – possibly in the subdomain, pathname or credentials – but that company is not the destination of the URL.
- **Mangle:** The company name has letter substitution, misspelling or non-ASCII characters (similar to English ones), resulting in visually identical web addresses, known as a Homograph attack [16].
- **Camouflage:** The company name contains an extension in the domain name, such as a different top-level domain or deliminator-looking character other than the normal period; this is usually done with the addition of a hyphen. For example, the use of *home-depot.com* instead of *homedepot.com*.

Detecting Malicious URLs

The work related to detecting malicious URLs falls into two main approaches: automated detection and user training.

Automated Phishing Detection

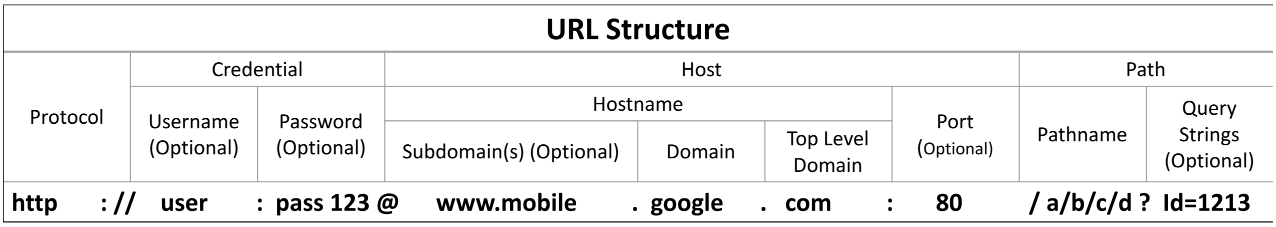
Automated phishing detection uses a combination of many factors to detect phish, which includes the URLs in the communication. These detection tools are used by various groups. Large organizations will use them to scan all incoming communication such as email and proactively remove communications that are known to be fraudulent. Individual users can also download tools for their browsers and other communication clients that will identify fraudulent communication and either remove or warn about it [17]. There are also bots, such as MetaCert, which scans communications in Slack channels [1].

Most phishing identification procedures depend on Blacklists, meaning a list of phishing URLs [20]; however, these tools do not prevent zero-hour attacks, which is the attack before the malicious URL is discovered [17]. The Anti-Phishing Working Group revealed that the normal time taken to discover a phishing URL is 28.75 hours, during which time users are unprotected [9]. These tools can sometimes give false warnings that decrease users' trust in the results and cause them to ignore future warnings; consequently, the effectiveness of these tools relies on users' behaviour [12].

Training Users

While automation is a good idea, and effective, it is currently impossible to completely remove the user from the loop. Communication is an important part of business operations and overly aggressive automatic filters are likely to cost organizations in lost productivity. As a result, some phishing attempts will get through the automated filters, necessitating the training of users as a second complementary line of defence. There are two common types of training: upfront and embedded.

In upfront training, a user will go through a training session where they will learn about phishing in a condensed format. Examples include the Anti-Phishing Phil game [25] and NoPhish app [7] both of which train users to read URLs using concentrated engagement, such as a game. The upfront approach effectiveness relies on the user being able to understand the materials, retain them, and be able to apply them to daily activity. Prior work demonstrates that upfront training is effective when it comes to enhancing users' capability to



URL Structure								
Protocol	Credential		Host				Path	
	Username (Optional)	Password (Optional)	Hostname			Port (Optional)	Pathname	Query Strings (Optional)
			Subdomain(s) (Optional)	Domain	Top Level Domain			
http :: //	**user :**	**pass 123 @**	**www.mobile .**	**google .**	**com :**	**80**	**/ a/b/c/d ?**	**Id=1213**

Figure 1: URL structure and example.

identify phishing URLs; however, the long-term benefit of this approach is uncertain [17]. Importantly, this approach can fail to produce a long-term advantage [18] because of the nature of forgetting [17]; along these lines, Volkamer et. al [27] recommends that users need to integrate training into their daily life. Another issue is that people are unwilling to invest energy in online instructional exercises, particularly given the perceived low risk of being exposed to real danger [10, 15, 27].

Embedded training involves integrating the training into the daily life of the users. The most classic example of which is sending out fake phishing emails to employees and providing contextual training for those who click on the links [19]. Unlike upfront training, embedded training is fairly lite, requiring small amounts of time for most users and more time only for users who click the malicious links. However, due to its lite touch, users may not get the opportunity to build a strong conceptual model of how phishing works; making the lessons harder to apply in different contexts. Because this kind of training is embedded in routine, it is challenging to create consistent security training messages across an organization or worse, between organizations, potentially leaving users with conflicting advice [15, 23].

FAHEEM BOT DESIGN

The objective of this work is to develop and test Faheem: a Slack bot with the capacity to parse URLs posted in a Slack channel and clarify their components. It also warns users about suspicious patterns using friendly explanatory language that users can understand.

Our primary design objective is to create an interactive chat bot which helps average internet users correctly read URLs and identify phishing URLs. In order to accomplish this goal we focus on two features of the bot:

1. Parsing the URLs and identifying common malicious behaviours focusing primarily on the domain issues.
2. Presenting the results to the user in a clear and easy to understand manner.

Platform

We selected Slack as the platform for the bot because Slack is a commonly used communication platform with good support for custom bots. Slack bots can join any group, read and post messages and also contact members in direct messages.

URL Parser

The URL parser uses the Node.js programming language. The detailed processes for the URL analysis is as follows:

1. Listens to all Slack chats in the forum and extracts URLs using *url-regex* package.
2. Identifies and resolves IP address. The *ip-regex* package was used to detect IP address while the *dns* constructor package was used to reverse it to obtain the registered hostname.
3. Checks and resolves redirects and shortened links using the *unfurl-url* package to obtain the destination URL.
4. Parses the final URL into its component parts as shown in Figure 1 by using the built-in *URL* constructor provided by *node.js*.
5. Checks the domain for similarity with domains of the top 500 websites on Alexa Global Sites. Using the Levenshtein distance metric from *clj-fuzzy* package.
6. Checks for non-ASCII characters using the *non-ascii* and *langdetect* packages.

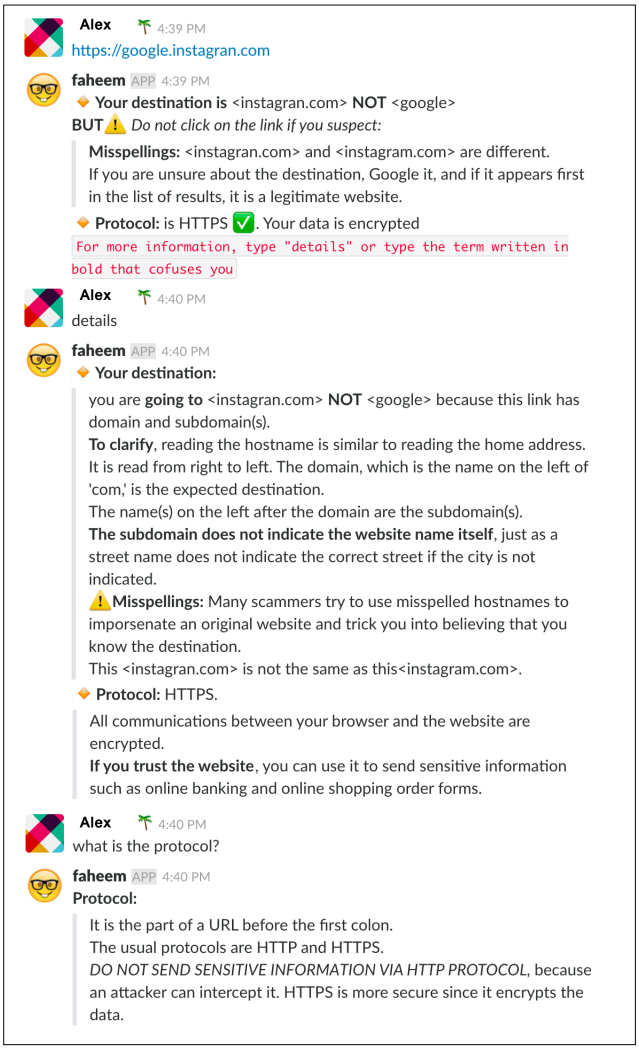
Walkthrough Example

For clarity, we detail here a sample interaction between the Faheem bot and a user Alex also pictured in Figure 2.

The user Alex starts the interaction by posting a URL into a Slack group the bot is listening to, which Faheem then detects. Faheem parses the URL and presents the most important information to Alex first with an offer of further details on request. In this case, Faheem detected that the subdomain is similar to popular domain google and warns the user that this URL will not go to Google. It also detects a small edit distance between the domain 'instaran' and the popular domain 'instagram', which it points out to Alex along with actionable advice on what to do if she is unsure. Finally, it provides positive feedback that the URL uses HTTPS and is therefore encrypted in transit.

Alex wants more details so she replies with "details". Faheem expands each of the previously presented sections and provided general advice for users, such as: "To clarify, the hostname is similar to reading the home address, etc" in order to help them develop conceptual understanding to deal with security risks.

Alex is confused about the Protocol section and asked Faheem about it by typing: "protocol". Faheem responds by explaining what a protocol is, particularly clarifying about HTTP.



Alex 4:39 PM
https://google.instagran.com

faheem APP 4:39 PM
◆ **Your destination is** <instagran.com> **NOT** <google>
BUT ⚠ *Do not click on the link if you suspect:*

> **Misspellings:** <instagran.com> and <instagram.com> are different.
> If you are unsure about the destination, Google it, and if it appears first in the list of results, it is a legitimate website.

◆ **Protocol:** is HTTPS ✅. Your data is encrypted

`For more information, type "details" or type the term written in bold that cofuses you`

Alex 4:40 PM
details

faheem APP 4:40 PM
◆ **Your destination:**

> you are **going to** <instagran.com> **NOT** <google> because this link has domain and subdomain(s).
> **To clarify**, reading the hostname is similar to reading the home address. It is read from right to left. The domain, which is the name on the left of 'com,' is the expected destination.
> The name(s) on the left after the domain are the subdomain(s).
> **The subdomain does not indicate the website name itself**, just as a street name does not indicate the correct street if the city is not indicated.
> ⚠**Misspellings:** Many scammers try to use misspelled hostnames to imporsenate an original website and trick you into believing that you know the destination.
> This <instagran.com> is not the same as this<instagram.com>.

◆ **Protocol:** HTTPS.

> All communications between your browser and the website are encrypted.
> **If you trust the website**, you can use it to send sensitive information such as online banking and online shopping order forms.

Alex 4:40 PM
what is the protocol?

faheem APP 4:40 PM
Protocol:

> It is the part of a URL before the first colon.
> The usual protocols are HTTP and HTTPS.
> *DO NOT SEND SENSITIVE INFORMATION VIA HTTP PROTOCOL*, because an attacker can intercept it. HTTPS is more secure since it encrypts the data.

Figure 2: An example interaction with the Faheem bot.

Explanation of Design Choices

Simple initial information Users need to receive information just at the time they need it, especially since people may struggle to remember information that they have received out of context [4]; therefore, on detection of URLs, Faheem presents a concise summary format of the key information and replacing the technical concepts with terms which can be understood by average users. The key information and format of the summary were agreed by a focus group of security and HCI experts. Domain issues are quite serious security wise as the user may be communicating with someone other than they intend [27]. Hence, Faheem focuses initial information on issues around the domain including warnings for known malicious URL tricks, as shown in Figure 3.

User-lead interaction Faheem gives the users a chance to ask for details in general or for specific information. The goal is to make the interaction user-lead where the user can decide what they are most interested in seeing rather than provide piles of information up-front. Conceptual explanations are also provided to help people build relations between concepts and assist with applying to learned lessons to new situations [24].

Highlight most problematic elements with evidence Where possible, Faheem uses evidence from the URL itself to demonstrate potential issues to users in such a way that they can understand the issue and bring their own expertise to bear. For example, Faheem checks for non-ASCII characters and when found it points out to the user that there are, say, Cyrillic letters in a mostly ASCII URL and shows them which letter is non-ASCII. Other problems like potential misspellings of a common domain are also contextualized by stating both the domain in the URL and the common one so that the user can compare them by themselves. Moreover, where best practices exist, Faheem provides expert advice and positive reinforcement of certain actions. For example. HTTPS is almost always a better choice than HTTP so Faheem puts a green check mark to indicate that having HTTPS is a good feature of the URL.

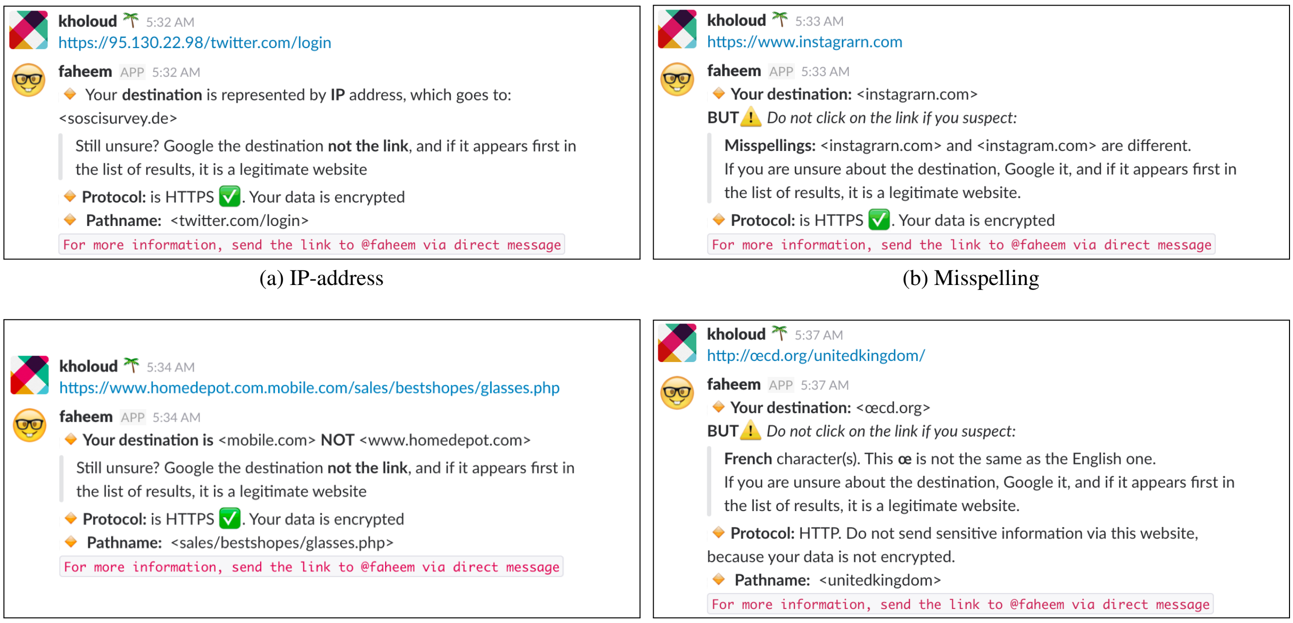
Explanation with advice Faheem provides general and actionable advice. The general advice follows the clarifications to help users to deal with URLs. For example, not to send sensitive data through a HTTP connection. Users are advised to take an action when they doubt a URL (procedural knowledge). This provides them with clear choices and potentially increases their ability to differentiate the original from the spoof URL. For example, Faheem advises them to Google the domain if they are unsure about its safety.

URL EXPLAINER DESIGN

URL Explainer is a website created by one of the authors as a class project when studying abroad. It takes in a URL, feeds it into the URL.js parser, and presents the results on a webpage. Each presented element of the URL is pulled out separately onto different lines where the components are highlighted and a generic explanation provided. URL Explainer also attempts to fetch the URL server-side to get its title and preview. An example can be seen in Figure 4.

A small pilot was run with 14 university students to see if URL Explainer could be used to improve URL reading skills. The study had a simple three-part format, with a pre-test, a test where they could use URL Explainer, and a post-test. We found, unsurprisingly, that participants are bad at identifying the destination of a URL; participants had an average accuracy of just over 50% in the pre-test. When using URL Explainer, participants jumped to 100% accuracy while the control condition which had no assistance stayed at 50% accuracy. Unfortunately, when URL Explainer was taken away, experimental participants dropped to an accuracy of 54% compared to the control which had a post-test accuracy of 34%. The overall take away from the study was that URL Explainer did help people correctly identify the end destination of the URL, but using it did not lead to skill building or retention.

In this paper, we will be using URL Explainer as a control condition to compare Faheem with. We selected URL Explainer as a control condition because it is comprehensive, simplistic, and shown to be effective at helping a user read a



kholoud 5:32 AM
https://95.130.22.98/twitter.com/login

faheem APP 5:32 AM
🔶 Your **destination** is represented by **IP** address, which goes to: <soscisurvey.de>

> Still unsure? Google the destination **not the link**, and if it appears first in the list of results, it is a legitimate website

🔶 **Protocol:** is HTTPS ✅. Your data is encrypted
🔶 **Pathname:** <twitter.com/login>
`For more information, send the link to @faheem via direct message`

(a) IP-address

kholoud 5:33 AM
https://www.instagrarn.com

faheem APP 5:33 AM
🔶 **Your destination:** <instagrarn.com>
BUT⚠️ *Do not click on the link if you suspect:*

> **Misspellings:** <instagrarn.com> and <instagram.com> are different.
> If you are unsure about the destination, Google it, and if it appears first in the list of results, it is a legitimate website.

🔶 **Protocol:** is HTTPS ✅. Your data is encrypted
`For more information, send the link to @faheem via direct message`

(b) Misspelling

kholoud 5:34 AM
https://www.homedepot.com.mobile.com/sales/bestshopes/glasses.php

faheem APP 5:34 AM
🔶 **Your destination is** <mobile.com> **NOT** <www.homedepot.com>

> Still unsure? Google the destination **not the link**, and if it appears first in the list of results, it is a legitimate website

🔶 **Protocol:** is HTTPS ✅. Your data is encrypted
🔶 **Pathname:** <sales/bestshopes/glasses.php>
`For more information, send the link to @faheem via direct message`

(c) Multilevel domain (subdomains)

kholoud 5:37 AM
http://œcd.org/unitedkingdom/

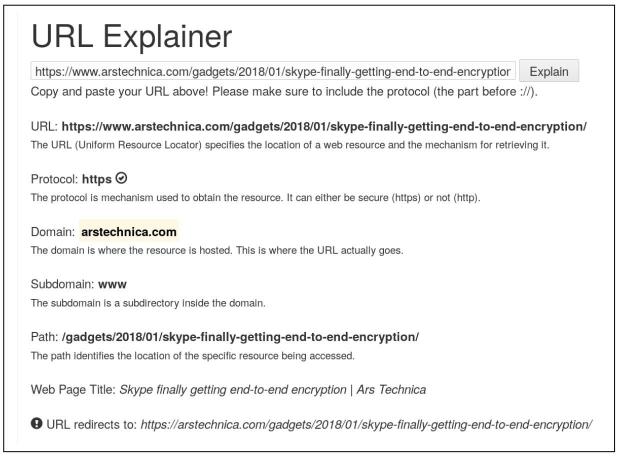
faheem APP 5:37 AM
🔶 **Your destination:** <œcd.org>
BUT⚠️ *Do not click on the link if you suspect:*

> **French** character(s). This **œ** is not the same as the English one.
> If you are unsure about the destination, Google it, and if it appears first in the list of results, it is a legitimate website.

🔶 **Protocol:** HTTP. Do not send sensitive information via this website, because your data is not encrypted.
🔶 **Pathname:** <unitedkingdom>
`For more information, send the link to @faheem via direct message`

(d) Non-ASCII characters in the domain

Figure 3: Sample Faheem messages for different malicious URL patterns.



URL Explainer

https://www.arstechnica.com/gadgets/2018/01/skype-finally-getting-end-to-end-encryptior

Copy and paste your URL above! Please make sure to include the protocol (the part before ://).

URL: **https://www.arstechnica.com/gadgets/2018/01/skype-finally-getting-end-to-end-encryption/**
The URL (Uniform Resource Locator) specifies the location of a web resource and the mechanism for retrieving it.

Protocol: **https** ✓
The protocol is mechanism used to obtain the resource. It can either be secure (https) or not (http).

Domain: **arstechnica.com**
The domain is where the resource is hosted. This is where the URL actually goes.

Subdomain: **www**
The subdomain is a subdirectory inside the domain.

Path: **/gadgets/2018/01/skype-finally-getting-end-to-end-encryption/**
The path identifies the location of the specific resource being accessed.

Web Page Title: *Skype finally getting end-to-end encryption | Ars Technica*

❗ URL redirects to: *https://arstechnica.com/gadgets/2018/01/skype-finally-getting-end-to-end-encryption/*

Figure 4: URL Explainer interface after the user has asked it to parse an Ars Technica news article URL where "www" has been added as a subdomain to show the URL redirect notice.

URL. Unlike Faheem, URL Explainer makes no attempt to identify Phishing indicators or provide contextual information, instead focusing solely on factual presentation of the contents of the URL itself. As such, it is a good choice for a control.

METHODS

An empirical lab study was conducted to investigate the overall effectiveness of Faheem's interactive explanations in raising users awareness of phishing URLs as compared to a basic presentation.

We hypothesise that Faheem users show a greater improvement, compared to URL Explainer users, in their ability to identify phishing URLs in the following two conditions: (i) With the support of the tool (Faheem or URL Explainer). (ii) When access to the tool has been removed.

Participant Recruitment

A request was posted by the lead researcher on their accounts for Twitter, WhatsApp, and Facebook. As a motivation, prospective participants were told that they would be awarded £10 for their participation. Only three people were located locally, the others were from a wide variety of countries including Saudi Arabia and parts of the European Union. The participants were from a variety of sectors including mathematics, business, and management. A total of 40 participants were recruited, 20 for each group, all of whom were aged between 20 and 58 years old with a mean of 28. 60% were female and 40% male.

Study Design

Because of the wide geographic locations of the participants, the study was conducted remotely with the researcher communicating with the participants via email and Slack using a pre-defined script which differed between conditions only in the explanation of the functionality of the systems.

Protocol

Setup: The study purpose was explained to the participants from each group, as well as what phishing is if they were not already aware. They were explicitly told to not visit any of the links, only to read them. They were also asked at the end of the study if they had used any external resources. Participants in the Faheem condition were asked to join a Slack team before the study started in attempt to limit the communication means between researcher and participants and ensure smooth study flow. They were invited to an empty Slack channel to which Faheem was added later in the study.

Demographics: Participants in both groups started the session with a consent form and a demographics survey where they were asked for their Slack username (Faheem) or preferred first name for communication (control), age, gender and topics they have previously studied, with two additional questions incorporated for the experimental group, asking how frequently they use Slack and chatbots to ensure their understanding of Slack and chatbots would not influence the study results.

Pre-test: Participants were given a set of 14 URLs one at a time via a survey and instructed to imagine that they had received each URL during an instant messaging interaction with the text "You want to visit <website name>" associated with the URL. For each URL the participants answered the following questions:

1. Decide whether it is a phishing or an original website.
   Select one: phishing, original
2. Which part of the link does influence your answer?
   Multi-answer: all elements of URL, except the protocol, were provided as choices.
3. Why would you click / not click on the link?
   Free-text answer.

The goal of these questions was to determine their a-priory ability to determine if the URL went to the stated organization or not.

Supported reading: Participants were given access to either a live version of Faheem or screenshots of URL Explainer and asked to use them to answer the same set of questions shown in the pre-test, but with a different set of URLs. For the Faheem group, the participants were given access to Faheem and told that a link would be posed in the group, which Faheem automatically would parse, and questions were then sent in the group chat. For the control group, members were given a survey with screenshots of URL Explainer for the link in question; screenshots were to ensure that easy access to URL Explainer could be revoked during the post-test below. Since URL Explainer produces static output, there is no functional difference between the actual page and an image.

Post-test Similar to the pre-test, participants in both groups were given a new set of URLs and asked to answer the same questions from the pre-test without the support of the tool. Participants were again asked to not type in URLs or use other resources. Access to Faheem was revoked and URL Explainer participants were asked if they had searched for the site online.

Tested URLs

In each of the pre, supported, and post stages the participant is given a set of 14 URLs, which were selected to cover the following phishing techniques: shortened links, redirects, IP addresses, misspellings, multi-level domains, company name located somewhere other than the host position, and non-ASCII characters. The three sets of URLs were selected to be comparable in structure but not identical. Every participant saw the same URLs in the same order. One URL from each stage was excluded from analysis due to a technical issue during data collection. Results are drawn from 13 URLs per stage.

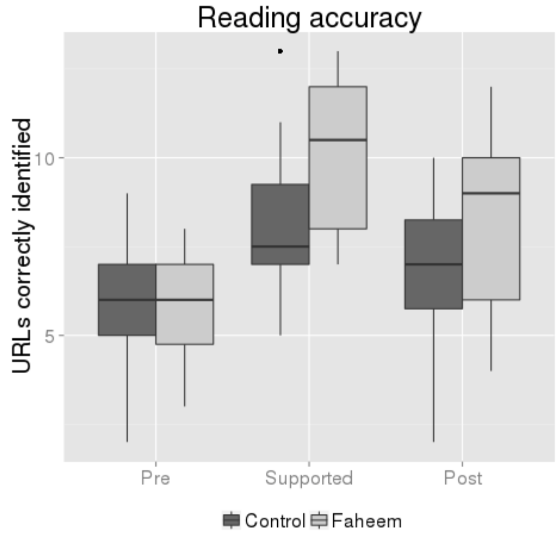


Figure 5: Boxplot of the number of URLs participants correctly identified as phishing for conditions in the pre test, when supported by the tool, and post test.

RESULTS

Our primary concern is if Faheem helped participants to accurately determine if a given URL lead to a particular company or not, both while using Faheem and after. As can be seen in Figure 5, participants were able to use both Faheem and URL Explainer to improve their ability to identify potential phishing URLs with more accuracy than they could without the tools. To determine if the Faheem group experienced a significant improvement compared to URL Explainer, we computed the per-participant change between supported and pre to account for initial skill variation. Then, we ran an independent t-test with an α of .05. We found that the Faheem group (M=4.55) showed statistically significantly more improvement than the control group (M=2.15), (p<0.003), and fairly large effect size (r=0.47).

We also looked at the difference between the pre and post tests. Similar to the prior analysis, we computed the change per participant and then compared using an independent t-test. Faheem (M=2.75) still showed a statistically significant improvement ($p< a0.044$) with (r=0.32) as compared to the control (M=1.05). Though the difference between the conditions narrowed after support was removed.

URL category	Pre		Supported		Post	
	Control	Faheem	Control	Faheem	Control	Faheem
Standard-URLs	43%	28%	83%	70%	58%	62%
IP-Based	80%	85%	70%	85%	85%	95%
Shortened links	15%	15%	30%	35%	_	_
Redirects	25%	60%	65%	100%	_	_
Misspelling	40%	48%	53%	100%	48%	88%
Multi-level domain	50%	40%	50%	80%	49%	50%
Company name not in host position	60%	58%	58%	75%	65%	70%
Non-ASCII characters	55%	55%	45%	100%	45%	90%

Table 1: URL identification accuracy for each condition, stage and type of URL issue presented. Participants were given an organization name and asked if the URL went to that organization or if it was likely phishing. So the top left value should be read as 43% of the standard URLs presented to the control group in the pre-test were correctly identified as the company or phishing.

As a reminder, participants were provided with a company name and asked if the URL lead to that company or if it was likely phishing. Table 1 shows the results of the question for the different conditions, stages, and types of URL manipulations. The pre-test results show that participants in both conditions achieved the lowest scores for standard, shortened, redirects and misspelt URLs.

For the supported stage, both groups scored lower for the shortened link *https://bit.ly/18AOiDE* which redirects to *https://www.facebook.com/unsupportedbrowser*. Participants' justifications were different, with one of them stating "*Bitly* always sends me to advertisement website", and others stating that the link goes to Facebook but 'unsupported browser' in the link is suspicious. Both URL Explainer and Faheem resolve shortened URLs, like the Bitly example above, and tell the user the ultimate destination of the URL. Participants in the supported stage clearly did not understand the feature or it failed to overcome their previous biases as they still do quite poorly at identifying phishing sites. One potential explanation is technical. Both Faheem and URL Explainer make a headless request to resolve the URL server-side. Doing so can trigger behaviours in the host server. In the above example, it caused Facebook to serve back its "unsupported browser" page rather than the actual content, which was then reflected in the two tools.

After using the tools, participants, in both groups, were seen to experience problems when the links containing top-level domains other than `.com`, such as *tagesschau.de*. The Faheem participants who answered this question correctly said that they Googled the domain, suggesting that the Faheem group did benefit from the provided advice. The top-level domain `.de` is the country code top-level domain for the Federal Republic of Germany. Another URL was *https://translate.google.co.uk/*. Participants who are not from the UK did not trust it with the justification provided was that they had never seen a Google website with these characters.

Moreover, the Faheem group of participants were confused between the URL and the recognized brand name for the organization, such as New York Times (*www.nytimes.com/*) whereas the other group's performance was found to be higher because the other tool provided a webpage title containing the full website name.

CONCLUSION AND FUTURE WORK

In this work, we have presented Faheem, a Slack bot which helps users learn about URLs in an interactive format. Faheem assists users who have no understanding of URLs in identifying common URL elements and well known malicious URL tricks. It also assists more experienced URL readers in identifying less user-visible tricks such as non-ASCII letters which are visibly identical to ASCII ones.

To test Faheem we compared it with URL Explainer, a simplistic web page which parses a URL for a user but focuses on a factual clear representation of the URL contents rather than helping the user identify common issues. We find that while using both tools, Faheem is better at helping a user identify URLs which have a destination other than where the user wishes to go. Additionally, we also saw some minimal learning effects with Faheem users showing an improved ability to identify phishing URLs after using the tool.

In conclusion, Faheem is a novel approach to helping users understand the contents of URLs. Our study shows that the approach has some promise, though more comprehensive studies are needed to conclusively determine the effectiveness of Faheem-type solutions.

REFERENCES

1. 2009-2018. MetaCert Security. (2009-2018). `https://slacksecurity.metacert.com/`
2. 2017. *2017 Data Breach Investigations Report*. Technical Report. Verizon.
3. 2017. *State of the Phish 2017*. Technical Report. Wombat security technologies.
4. Lawrence W Barsalou. 1999. Language comprehension: Archival memory or preparation for situated action? (1999).
5. T. Berners-Lee, L. Masinter, and M. McCahill. 1994. RFC1738: Uniform Resource Locators (URL). (December 1994). `https://www.w3.org/Addressing/rfc1738.txt`
6. Mark Button, David Shepherd, Dean Blackbourn, and Martin Tunley. 2016. *Annual Fraud Indicators 2016*. Technical Report. University of Portsmouth Center for Counter Fraud Studies.

7. Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Roland Borza. 2014. NoPhish: an anti-phishing education app. In *International Workshop on Security and Trust Management*. Springer, 188–192.

8. Gamze Canova, Melanie Volkamer, Clemens Bergmann, and Benjamin Reinheimer. 2015. NoPhish app evaluation: lab and retention study. In *NDSS workshop on usable security*.

9. APWG Internet Policy Committee and others. 2013. Global phishing survey: Trends and domain name use in 2h2013. (2013).

10. Nicola Davinson and Elizabeth Sillence. 2010. It won't happen to me: Promoting secure behaviour among internet users. *Computers in Human Behavior* 26, 6 (2010), 1739–1747.

11. Rachna Dhamija, J Doug Tygar, and Marti Hearst. 2006. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. ACM, 581–590.

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

13. Nina Eyrich, Monica L Padman, and Kaye D Sweetser. 2008. PR practitioners' use of social media tools and communication technology. *Public relations review* 34, 4 (2008), 412–414.

14. Sujata Garera, Niels Provos, Monica Chew, and Aviel D. Rubin. 2007. A framework for detection and measurement of phishing attacks. In *Proceedings of the 2007 ACM workshop on Recurring malcode - WORM '07*. ACM Press, New York, New York, USA, 1. DOI: http://dx.doi.org/10.1145/1314389.1314391

15. Cormac Herley. 2009. So Long, And No Thanks for the Externalities: The rational rejection of security advice by users. In *Proceedings of NSPW'09*.

16. Oliver J Hunt and Ivan Krstic. 2017. Preventing URL confusion attacks. (March 21 2017). US Patent 9,602,520.

17. Mahmoud Khonji, Youssef Iraqi, and Andrew Jones. 2013. Phishing detection: a literature survey. *IEEE Communications Surveys & Tutorials* 15, 4 (2013), 2091–2121.

18. Iacovos Kirlappos and M Angela Sasse. 2012. Security education against phishing: A modest proposal for a major rethink. *IEEE Security & Privacy* 10, 2 (2012), 24–32.

19. Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2008. Lessons From a Real World Evaluation of Anti-Phishing Training. *e-Crime Researchers Summit, Anti-Phishing Working Group* (October 2008). http://precog.iiitd.edu.in/Publications_files/eCrime_APWG_08.pdf

20. Alexandra Kunz, Melanie Volkamer, Simon Stockhardt, Sven Palberg, Tessa Lottermann, and Eric Piegert. 2016. Nophish: evaluation of a web application that teaches people being aware of phishing attacks.. In *GI-Jahrestagung*. 509–518.

21. Eric Lin, Saul Greenberg, Eileah Trotter, David Ma, and John Aycock. 2011. Does domain highlighting help people identify phishing sites?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2075–2084.

22. Cooper Quintin. 2015. HealthCare.gov Sends Personal Data to Dozens of Tracking Websites. (20 January 2015). https://www.eff.org/deeplinks/2015/01/healthcare.gov-sends-personal-data

23. R. Reeder, I. Ion, and S. Consolvo. 2017. 152 Simple Steps to Stay Safe Online: Security Advice for Non-tech-savvy Users. *IEEE Security Privacy* PP, 99 (2017), 1–1. DOI: http://dx.doi.org/10.1109/MSP.2017.265093101

24. Bethany Rittle-Johnson and Kenneth R Koedinger. 2002. Comparing Instructional Strategies for Integrating Conceptual and Procedural Knowledge. (2002).

25. Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Anti-Phishing Phil: the design and evaluation of a game that teaches people not to fall for phish. In *Proceedings of the 3rd symposium on Usable privacy and security*. ACM, 88–99.

26. Kurt Thomas, Chris Grier, Justin Ma, Vern Paxson, and Dawn Song. 2011. Design and evaluation of a real-time URL spam filtering service. In *Security and Privacy (SP), 2011 IEEE Symposium on*. IEEE, 447–462.

27. Melanie Volkamer, Karen Renaud, Karen Renaud, Paul Gerber, and Paul Gerber. 2016. Spot the phish by checking the pruned URL. *Information & Computer Security* 24, 4 (2016), 372–385.

28. Andrew G West and Adam J Aviv. 2014. On the Privacy Concerns of URL Query Strings. (2014).

Caught by Phishing Emails? How can Argumentation Schemes be Used to Protect Users?

Rosemary J. Thomas
University of Aberdeen
Aberdeen, United Kingdom
r02rj15@abdn.ac.uk

Matthew Collinson
University of Aberdeen
Aberdeen, United Kingdom
matthew.collinson@abdn.ac.uk

Judith Masthoff
University of Aberdeen
Aberdeen, United Kingdom
j.masthoff@abdn.ac.uk

ABSTRACT

We propose the creation of persuasive messages that would form the basis of an automatic messaging system aimed at protecting potential victims of phishing attacks by changing their behaviour. We use the coupling of the informal logical method of argumentation schemes and persuasive principles [12] as a foundation for generating and structuring the messages.

CCS CONCEPTS

•Security and privacy → Social aspects of security and privacy;

KEYWORDS

Security, Phishing, Email, Behaviour change, Persuasion, Argumentation schemes

1 INTRODUCTION

Phishing emails are designed to trick a user in some way. Often they are circulars, but less commonly they are carefully targeted at smaller groups or even individuals (spear-phishing). Most commonly, there is a fraudulent aspect to these emails, and they claim to originate from a source other than their true source. While security mechanisms do exist to guarantee the authenticity of sources of email, for historical, legacy and usability reasons ordinary email does not use such mechanisms. This makes it easy for a fraudster (phisher) to send emails purporting to come from someone else, the spoofed sender, and to take advantage of the trust that a recipient may have in the spoofed sender in order to persuade the recipient to perform some action. Phishing is often used to solicit or extract confidential information, including authentication credentials from victims, or to get them to authorize some process on a computer. There are a range of techniques used by fraudsters to increase compliance (that is, produce the desired action) in victims, in addition to spoofing the sender [1–3, 5, 8], some of which are discussed below. Phishing can therefore be regarded as a socially negative form of persuasive messaging. Indeed, it is commonly referred to as 'social engineering' in the security community, where there are tool-sets [10] to help with phishing and phishing defence. It should therefore be counteracted by appropriate behaviour interventions, as well as usable security tools.

Several user studies have investigated the personalisation of persuasive messages by adapting the Cialdini principles of persuasion [7, 11]. Table 1 summarises four of Cialdini's principles, excluding the 'Scarcity' and 'Reciprocity' principles which we do not use in this paper.

In appropriate settings, informal logical reasoning in a dialogue or argument can be thought of as a structured, or semi-structured, way of communicating to achieve persuasion. According to Walton, Reed and Macagno [14] (p1) "The most useful and widely used tool so far developed in argumentation theory is the set of argumentation schemes. Argumentation schemes are forms of argument (structures of inference) that represent structures of common types of arguments used in everyday discourse, as well as in special contexts like those of legal argumentation and scientific argumentation." The examples of argumentation schemes are given below.

Recently, two of the present authors, together with a colleague, have attempted to relate Cialdini's principles to argumentation schemes [12], giving what we refer to here as *mapped argumentation schemes*, allowing for automatic generation of persuasive messages. Logical reasoning for behaviour change has been investigated by other authors in other settings [6, 9].

In this paper we propose that the mapped argumentation schemes could be used to provide an effective model of persuasion in the domain of email security, with a specific focus on the problem of phishing. Moreover, we sketch some of the types of messages that could be generated in this way. The refinement of these methods and the experimental evaluation of the resulting messages is future work.

This work is supported by the EPSRC award EP/P011829/1.

2 TRIGGERS FOR ADDITIONAL VIGILANCE WITH PHISHING EMAIL: SPOTTING TYPES

Aside from specific security cues and spam warnings, there are several trigger features that can help users to identify phishing in emails, or at least to raise their level of suspicion or vigilance. We refer to such triggers as 'spotting types'

Principle	Description
Commitments and Consistency (COM)	"It is easier to resist at the beginning than at the end". When a person makes a dedication, he or she will experience individual and social strains to act in accordance with that initial choice.
Social Proof (SOC)	"Where all think alike, no one thinks very much". People confirm what is acceptable, by first learning what others believe to be acceptable.
Liking (LIK)	"The main work of a trial attorney is to make a jury like his client". We are likely to comply with requests put forward by the ones we recognise and like.
Authority (AUT)	"Follow an expert". Symbolic power linked to a person makes people adhere to their advice.

Table 1: Four of Cialdini's principles [4]

in this paper. This includes links, website addresses, source, content, request type and attachments [1–3, 5, 8]. We give a very brief discussion of some of these spotting types below. Suspicion that an email is a phishing attack may involve correlating evidence from consideration of several of these categories.

Source. A user may be able to spot that a message should not be trusted because there is a problem with the source. However, some users may be unaware of the risks associated with unknown or untrusted senders, they rely on the security of others' email accounts, and senders' addresses and identities can be spoofed.

Attachments. A security expert might assert that one should be suspicious of all files attached to emails, particularly when they come from untrusted sources or there are other reasons to be suspicious. However, in practice it is hard for users in the workplace to remain productive while taking special measures to deal with all such attachments, since many workflows progress through email.

Links and website addresses. Phishing emails contain often links that pose some danger to the user (e.g. directing to a spoofed phishing web-site, or initiating a cross-site request forgery or a malware download, or as part of a cross-site scripting attack) [1, 2, 5]. Modern email clients often do not make the true action resulting from clicking a link very obvious to the user, without additional user effort and concentration.

Speech act type, and tone. Phishing emails often contain requests or offers. Phishers often use urgency to increase the possibility of compliance among their victims.

Content. There may be notable mistakes in language and representation [1, 2]; there may be language not expected from the source; there may be semantic content or speech acts not expected from the source; there may be unusual or unexpected formatting of the message.

3 APPLYING ARGUMENTATION SCHEMES TO PREVENT EMAIL PHISHING

We have identified, building on earlier work in another persuasion domain [12], a number of argumentation schemes that appear to apply to persuasion situations around behaviour change to prevent users becoming victims of email phishing. Some of these are:

Argument from commitment with goal (A.C.G.). In this scheme, the proposed action links to help in the achievement of the user goal grounded on previous commitment. In the email and phishing self-protection setting, this scheme can be used to motivate users to commit to a positive email security activity supported by their previous commitment.

Argument from popular practice variant with action (A.P.P.A.). Here the proposed action is a popular practice in a group or majority. This can be used to motivate users to follow a popular email security activity.

Argument from memory variant with goal (A.M.G.). In this scheme, an action-goal link is established that is supported by a person known by the user to have achieved their goal, thereby asserting that the specific action is good. This can be used to motivate users to make them believe a positive email security activity completed by someone they know.

Practical reasoning variant with goal (A.P.R.G.). In this scheme, the proposed action is linked to a user's goal. This can be used to motivate users to perform an email security activity that helps them achieve their goal.

Argument from expert opinion with goal (A.E.O.G.). In this scheme, the action is suggested by an expert in the domain which will help the user achieve their goal. This can be used to motivate users to follow an email security activity that is suggested by the security expert.

An example of an argumentation scheme, using A.C.G., is given in Table 2, along with the scheme giving the generic persuasive message structure. A.C.G. is identified as corresponding to Cialdini's 'Commitments and Consistency'

Major Premise	**Actor A** is committed to **Commitment C** according to **Goal G**.
Minor Premise	When **Actor A** is committed to **Commitment C**, it can be inferred that **Actor A** is also committed to **Action N** which contributes to **Commitment C**.
Conclusion	**Actor A** is committed to **Action N**.
Message Structure	As **Actor A** wants to achieve **Goal G**, **Actor A** is committed to **Commitment C**. So, **Actor A** is also committed to **Action N** as it helps **Actor A** achieve **Commitment C**.

Table 2: Argument from commitment with goal, and corresponding message

Actor A	you
Commitment C	check website links before clicking them
Goal G	keep your email account safe
Action N	preview website links in your email application
User Message	As **you** want to **keep your email account safe**, **you** are committed to **check website links before clicking them**. So, **you** are also committed to **preview website links in your email application** as it helps **you** to **check website links before clicking them**.

Table 3: Instantiation of argument from commitment with goal

principle (COM) from Table 1. Table 3 illustrates an instantiation of the variables in the above scheme along with the corresponding user message in the domain of email security.

Table 4 provides a few examples of the messages that can be produced using the above-mentioned variants of argumentation schemes in the domain of email security.

4 DISCUSSION AND FUTURE WORK

Research Questions

Using developed variants of argumentation schemes, we will investigate the following research questions, which will be positioned in the domain of email security.

(1) Which argumentation schemes are most useful in driving the persuasion process?
(2) How best to deliver these argumentation schemes to people? Which would be the most effective mode of delivery?
(3) How best to combine argumentation schemes with traditional training and education?

These questions have important subsidiary questions. For example, with the first question, how the usefulness varies with context, time, and is specific to the individual. For the second, we may also wonder whether there are situations where argumentation should not be used, and how we can avoid overloading and fatiguing people. For the third, we may suspect that traditional education, training and communications already carry persuasive messages, that there may be interference, and consider how to deal with this.

Research Plans and Methodologies

Our research will use a number of steps, detailed below. These steps are similar to those already proposed in other work to produce healthy eating messages [12].

First, we will adapt the already developed argumentation scheme system and message creation system [13], currently used to produce healthy eating messages, to produce messages for the email security domain. These messages would be validated with argumentation schemes experts, email security experts and lay people.

Second, we need to determine user characteristics such as personality, attitude, behaviour, age and gender that effect the persuasiveness of the messages produced. The validated and revised messages would be presented to lay people to measure the *perceived* persuasiveness.

Third, we need to determine the contexts the messages will be delivered in (e.g. in training or through email clients in live mail processing, at work or at home), and the mechanisms for delivery, since these may impact persuasiveness.

Fourth, we will determine an effective algorithm for message selection which adapts to user characteristics within educational training modules or email processing clients.

Finally, we will evaluate the effectiveness of the message selection algorithm. The evaluation can be implemented using a email simulation system where users receive emails both genuine and phishing, and where persuasive messages could be incorporated with some of the emails. Users will be given tasks which require them to use certain emails (to

Spotting Type	Message	Scheme
Links	As you want to keep your email account safe, you are committed to check website links before clicking them. So, you are also committed to preview website links in your email application as it helps you to check website links before clicking them.	A.C.G.
Source	The majority of people check if the origin of the incoming email is authentic. You should therefore do likewise.	A.P.P.A.
Content	Your aunt recalls that checking the content of the email for discrepancies helped them to keep their email account safe. So, you should believe that checking the content of the email for discrepancies is good.	A.M.G.
Request type	If you don't reply to emails that claim to lose account access by providing personal information, it helps you to safeguard your email account. So, you ought to do this.	A.P.R.G.
Attachment	A renowned IT security expert recommends that you don't open suspicious attachments disguised as authentic that will install malwares to your system to protect email account. So, you should follow their recommendation.	A.E.O.G.

Table 4: Persuasive messages created using argumentation schemes

simulate that in real-life users will sometimes have to open attachments and click on websites). This would help test the *actual* persuasiveness of the messages (and the relative persuasiveness of different message types), but also investigate the extent to which the messages impact on the persuasiveness with which users can do their jobs. We will compare the impact of adaptive message selection to a baseline selection and to the use of no messages at all.

Discussion

The increasing email security and privacy impact of phishing attacks, and the way that they address human vulnerabilities, make it likely that digital behaviour interventions, going beyond traditional usable security, are necessary to support users. In this paper, we have explained the problem of phishing and some methods available to users for identifying it. We have briefly sketched how variants of argumentation schemes, coupled to persuasive principles, could be used to generate messages to help users defend against phishing emails. The detailed experimental validation of this idea is a critical next stage.

REFERENCES

[1] A. Almomani, B.B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani. 2013. A survey of phishing email filtering techniques. *IEEE Communications Surveys and Tutorials* 15, 4 (2013), 2070–2090.

[2] A. Bergholz, J. De Beer, S. Glahn, M.-F. Moens, G. Paa, and S. Strobel. 2010. New filtering approaches for phishing email. *Journal of Computer Security* 18, 1 (2010), 7–35.

[3] T. Caldwell. 2013. Spear-phishing: how to spot and mitigate the menace. *Computer Fraud & Security* 2013, 1 (2013), 11 – 16.

[4] R.B. Cialdini. 2009. *Influence: The Psychology of Persuasion.* Harper-Collins e-books.

[5] I. Fette, N. Sadeh, and A. Tomasic. 2007. Learning to Detect Phishing Emails. In *Proceedings of the 16th International Conference on World Wide Web.* ACM, New York, NY, USA, 649–656.

[6] F. Grasso, A. Cawsey, and R. Jones. 2000. Dialectical argumentation to solve conflicts in advice giving: a case study in the promotion of healthy nutrition. *International Journal of Human-Computer Studies* 53, 6 (2000), 1077–1115.

[7] M. Kaptein, P. Markopoulos, B. de Ruyter, and E. Aarts. 2015. Personalizing persuasive technologies: Explicit and implicit personalization using persuasion profiles. *International Journal of Human-Computer Stud* 77 (2015), 38 – 51.

[8] P. Kumaraguru, J. Cranshaw, A. Acquisti, L. Cranor, J. Hong, M.A. Blair, and T. Pham. 2009. School of phish: a real-world evaluation of anti-phishing training. In *Proc. 5th Symposium on Usable Privacy and Security.* ACM.

[9] I. Mazzotta, F. de Rosis, and V. Carofiglio. 2007. Portia: A User-Adapted Persuasion System in the Healthy-Eating Domain. *Intelligent Systems, IEEE* 22, 6 (Nov 2007), 42–51.

[10] SET . The Social Engineer's Toolkit. ( ). https://www.social-engineer.org/ Accessed 27th December 2017.

[11] R. J. Thomas, J. Masthoff, and N. Oren. 2017. Adapting Healthy Eating Messages to Personality. In *Persuasive Technology. 12th International Conference, Amsterdam, Proceedings.*

[12] R. J. Thomas, J. Masthoff, and N. Oren. 2018. How can Argumentation Schemes be used for Persuasion? Working paper (In progress). (2018).

[13] R. J. Thomas, N. Oren, and J. Masthoff. ArguMessage: A System for Automation of Message Generation using Argumentation Schemes. In *Proceedings of AISB Annual Convention 2018 (18th Workshop on Computational Models of Natural Argument).*

[14] D. Walton, C. Reed, and F. Macagno. 2008. *Argumentation schemes.* Cambridge University Press.

Preparing for GDPR: Helping EU SMEs to Manage Data Breaches

Keshav Kapoor
University of Glasgow
Glasgow, Scotland

Karen Renaud & Jacqueline Archibald
School of Design and Informatics, Abertay University,
Dundee, Scotland
k.renaud,j.archibald@abertay.ac.uk

ABSTRACT

Over the last decade, the number of small and medium (SME) businesses suffering data breaches has risen at an alarming rate. Knowing how to respond to inevitable data breaches is critically important. A number of guidelines exist to advise organisations on the steps necessary to ensure an effective incident response. These guidelines tend to be unsuitable for SMEs, who generally have limited resources to expend on security and incident responses.

Qualitative interviews were conducted with SMEs to probe current data breach response practice and to gather best-practice advice from SMEs themselves. The interviews revealed no widespread *de facto* approach, with a variety of practices being reported. A number of prevalent unhelpful-practice themes emerged from the responses, which we propose specific mitigation techniques to address.

We therefore propose a SME-specific incident response framework that is simple yet powerful enough to inform and guide SME responses to data breach incidents.

1 INTRODUCTION

Cyber security has been identified as one of six Tier 1 threats to national security. It is estimated that it will cost the UK up to £27 bn per year [29]. Annually, it is reported by industry white papers [21, 55] that the number of data breaches is rising [23]. Attacks are also becoming increasingly sophisticated.

Small organisations (SMEs) are not immune to being targeted by hackers [20]. SMEs cannot expect to avoid detection or attacks due to their small size. Indeed, Krebs [35] reports that they are increasingly the prime target. It is essential that they plan for, respond to, recover and learn from hacking attacks [48, p. 131].

On the other hand, it is infeasible for SMEs to follow advice given to larger organisations. Incident response advice is rarely tailored to an organisation's needs, nor does it acknowledge organisation size and resources [17, 41]. A number of general standards and guidelines have been published to inform business incident responses [13, 22, 32, 43] but they are extensive and attempt to cover all bases. For example, the Experian guide to dealing with data breaches [22] has 31 pages and makes frequent reference to "upper management". The Data Breach Response Checklist, published by the US government [43], refers to the role of Human Resources and having a 'response team'. Their advice is also extensive and comprehensive, spanning 8 pages. The ICO data security breach management guide [31] also makes reference to Human Resources and IT teams, and having 'technical' and 'non-technical' staff to assist in the development of recovery plans. These data breach response guides, while certainly exemplary and helpful to large organisations, are not practical for SMEs to follow, especially those at the smaller end in terms of numbers of employees.

The situation for SMEs, at present, is that the consequences of a data breach could be bankruptcy [11]. Yet they are simply not in a position to hire and maintain security staff to take care of this the way big organisations can, nor, in many cases, can they afford to outsource their data breach response.

The European Union's GDPR regulation is coming into force in May 2018 and applies to all organisations regardless of size [16]. This is going to force organisations to contemplate their data breach response strategies [10]. SMEs need simple and clear guidelines for responding and meeting the requirements of the new law. Not doing so could make them go out of business or risk breaking the law.

We carried out research to develop an SME-specific incident response framework that was simple enough for SMEs to follow, yet powerful enough to be fit for purpose. The framework we developed is less comprehensive than the general guidelines published by respected bodies, but still covers the legally mandated aspects of a breach response. Moreover, it has been deliberately simplified for use by a non-expert and/or non-technical SME owner. Elements have been incorporated specifically to address typical panicked reactions such as overly technical and unthinking responses, and to encourage the development and maintenance of an organisational memory to ensure that SMEs develop personal best practice in terms of breach responses.

2 BEST PRACTICE INCIDENT RESPONSE

We commenced with a literature review in order to inform the formulation of the interview questions we were going to ask our SME respondents.

Academic literature has seen some focus on incident handling within SMEs [26]. Despite this step in the right direction, concrete research has yet to come up with a widely-agreed SME-specific incident response framework.

Several incident response frameworks and guides have emerged from industry [13, 32], government [28, 40] and academia [38, 44]. American standards bodies, such as NIST [13] and CREST [14], also provide helpful guidance. Most demarcate the following distinct incident response stages:

AISB'18, April 4–6, 2018, Liverpool, UK

Preparation: the first step and involves organisations readying themselves for an incident through establishment or preparation of an internal information security incident response team (CSIRT).

Verification: 'detection or reporting of security incidents within an organisation' [25, p. 32].

Containment: this might involve isolating the systems, changing passwords and disabling accounts, depending on the vulnerability that was exploited.

Eradication: organisations seek to eliminate the components of an incident with a focus on the restoration of systems [34] through actions including; identification of affected hosts, conducting malware and forensics analysis.

Recovery: taking actions such as: continuous testing and verification coupled with using back-ups to restore systems to normal business operations.

Follow-up: activity such as holding a post-incident meeting to explore actions that were taken and considering how effective they were. What could be done differently next time? Answers should feed back into organisational practices through the establishment of new controls, procedures and policies [37, 59]. However, as Jaatun *et al.*'s [33] study of the petroleum industry shows, in practice, challenges often arise which makes learning lessons difficult.

3 INTERVIEWING SMES

Ethical approval was sought and granted for interviewing SMEs, see below. We crafted a set of questions to explore SMEs' understanding of data breaches and their extant practice with respect to incident responses. We decided to carry out semi-structured interviews so that we had the flexibility to explore their responses further and elicit valuable insights from them.

We thus commenced with a pre-defined set of questions (see Appendix A) then explored their responses, changing ordering and exploring particular issues they raised [42].

We targeted SMEs, and specifically SME employees who were responsible for Information Security. Participants were recruited via convenience sampling and word-of-mouth. Due to difficulties in recruiting SMEs to participate, we switched to asking them to participate in order to recommend best responses to breaches. This helped us in terms of recruiting, since they were no longer concerned about admitting to being breached themselves.

3.1 Carrying out the Interviews

In Summer 2017, we carried out semi-structured interviews with organisations to explore current practice and to gather advice from 11 respondents [30]. We explored three particular topics:

(1) Understanding: of what the term "data breach" meant [52, 53],

(2) Current Practice: what they currently did with respect to preparing for, and responding to, data breaches [25, 38], and

(3) Best Practice Advice what they would advise other organisations like themselves to do with respect to managing data breach incidents (to allow them to feel that they were contributing to compiling 'good practice' for the benefit of other SMEs).

3.2 Results

(1) Data Breach Understanding. When asked to define the term "data breach", participants used phrases similar to: '*it involves unintended disclosure or access of information*'. Some also highlighted the difference between a security incident and data breach saying '*security incidents are sort of a wider ranging term, for instance you could have an incident with cloud technologies or networks, breaches would kind of be more concerned with data*'.

(2) Current Practice.

Preparation:

All participants explained that preparation before a data breach was crucial. Only two participants had no formal preparation in place because '*we don't know how and what to plan for.... plus, we have an IT supplier!*' When asked which actions should be prioritised during preparation one explained, '*establishing and preparing; the CISO*' others proposed, '*establishing who to contact*'. Participants who reported to having a CSIRT explained that preparation should focus on '*fully equipping the CSIRT*'.

Participants 11 and 5 outlined the importance of rehearsal stating, '*plans should be war gamed annually*'. Levels of rehearsal varied, with participant 5 explaining that their plan is practiced '*in anger on a daily basis*'. On the other hand, Participant 2 reported '*I don't think the plan is practiced at all!*'. A number of participants reported having plans in place but also admitted that '*we have a plan ... whether we follow that every time (we respond) I am not sure*'.

Verification:

All interviewees explained that an immediate first step when responding to breaches is acting on manual and automatic reports. Participant 2 explained '*after the report, the first thing we did was to take steps to verify that it was indeed a data breach*' and Participant 4 said that, they also '*identified [the] nature of the attack by trying to understand the potential impact and damage caused*'.

When asked who should react first to these alerts, those handling externally, such as Participant 3, explained that the external party needs to react first as '*information security incidents are handled, verified and coordinated through an external source*'. Participants handling internally echoed the views of Participant 6 who explained that verification is conducted by '*the person or team most appropriate to respond*'.

Interviewees also explained that during verification, forensic and technical tools were used, with Participant 6 stating '*identify the incident and detect possibly via SIEM tools*'. Some participants explained that once a data breach had been verified, forensic triage was to be completed. At this point Participant 10 added that, '*if you've verified a data breach, identify whether stolen data was encrypted or non-encrypted*'.

Containment:

Respondents explained that actions to contain breaches came next, with Participant 9 explaining actions focused on, '*ensuring that the outflow of data has been stemmed*'. Participant 7, further stated '*containment depends on the type of incident that has occurred ... let's say a phishing email came through, you could take some quick steps ... However, with 'WannaCry' we had to take more serious action*'. During containment, common actions included isolation by '*Taking the* *server offline cut all ties to isolate it from the network.*' '*changing all passwords*'. Participant 10 even mentioned, '*Carry on! Some people will say isolate until you've fixed it — well the answer is, don't*'.

Participants also emphasised the importance of communication which some referred to '*contacting senior management to let them know what had happened … to get their permission to take servers offline.*' and for others referred to getting, '*Legal and PR involved*'. For participant 3, external communication was key as they explained '*we have contractual arrangements with many external suppliers*'. Interviewees also echoed the views of Participant 4 who said, '*dependent on the type of attack, we may have to contact authorities (ICO)*'.

Recovery:
Participants described a recovery stage where, similar to Participant 9, actions focused on, '*restoring the integrity of the hacked system*' through the prioritisation of technical actions because as Participant 6 explained, '*technical changes are the quickest to implement for recovery. They may be a blunt instrument to address the problem but may be necessary*'. Participant 5 also suggested if a breach '*results in a loss of system operation, then ideally recovery of systems from a recent backup*'. Meanwhile, Participant 7 pointed out that, '*sometimes recovery from these things can be much bigger*' and others such as Participant 2 even said, '*I have no idea the IT and technical guys took care of the recovery*'.

Follow-up:
All interviewees pointed out that learning lessons during the follow up stage and implementing these lessons back into practice after a breach would lead to improvements in the response process and prevent future attacks from happening. Despite this, there was an inconsistent and unreliable execution of lessons learnt in practice. Participant 7, explained that, their organisation prioritised '*continuous application of lessons learnt and proactively apply best practice*'. Others said implementing lessons learnt is difficult because '*events are rapidly forgotten, as business priorities change security concerns drift back towards the bottom of the pile*'.

Participants expressed 'lessons should be learnt through organisational changes including, '*security policy revision*', '*security culture change*', '*user training*' and, '*changing passwords*'. A large number of interviewees also focused on technical changes such as '*changing firewall rules*', '*deploying canary/honeypot devices*'.

(3) Best Practice Advice

Locate assets:
Interviewees advised that each individual organisation needed to understand what it's 'crown jewels' are by asking questions such as; 'What issues are created by the compromise of data in an organisation?' so that they can use this information '*to potentially work out plans of what to do when people actually come after them*'. Participant 7 provided an example of this and explained, '*if we lost 50 email addresses and phone numbers it's not ideal but may be catastrophic for others. However, if we lost 2 million customer records —- that's catastrophic for us*'.

Prioritise security:
Participant 2 expressed views held by others, stating, '*first things first: you need to establish a mind-set where you expect to be breached*'. Interviewees stated that, by so doing, organisations could prioritise security concerns and '*engender a culture of security within the workforce: both at work and within private lives*'.

Simplify:
Participant 9 and others advised organisations to, '*boil security down to the simple things*'. Participant 3 gave examples of simple solutions stating, '*ensure the IT you are using is protected by the latest versions of hardware and software —- saving by using old technology is false economy*'.

People are important:
Participants said: '*cyber-threats are often seen too narrowly as a technical issue. In fact cyber-security depends on the right approaches to technology, but also personnel*' .

Participants advised, '*It is essential to establish open and clear communication networks with staff, senior management and third parties*'. Others also advised open and constant communication with '*affected customers*', '*external bodies*' and '*external incident handling parties*'.

Need for Measured Action:
Respondents explained that it was essential for organisations to document every action. Participant 2 said this was because '*you need to have evidence of all the actions you've taken*'. Participant 6 advised that, to comply with this, '*organisations should have tools in place to gather information, before the incident takes place*'.

Interviewees encouraged organisations to seek external support before a breach from 'The National Cyber Security Centre' and by '*hiring security professionals*'. Others stated that seeking help during the incident was vital in order to contain the breach, with Participant 10 explaining '*if a breach is verified bring in expertise very quickly to act upon the problem*'.

Interviewees urged organisations to '*report breaches of personal data to the ICO and to become aware of important regulation such as The General Data Protection Regulation (GDPR)*' [31].

3.3 Limitations

A limitation of this research is that the sample is not large enough to be fully representative of the views of the general population. This is due to the fact that many organisations do not want to speak about such a sensitive topic. Only eleven of the 100+ organisations we contacted were willing to be interviewed. It was only when we switched the focus, from speaking about their own data breach responses, to eliciting advice for others that we were able speak to eleven SMEs. Still, we have to acknowledge this as a limitation and we hope to be able to find a better way to recruit participants in the future.

4 REFLECTION

It is clear that the SMEs we spoke to define and interpret data breaches similarly and are aware of what the data breach meant. As a consequence, they could outline which actions needed to be prioritised during the preparation and follow-up stage. Compared to the unawareness reported by Line *et al.* [53] and Tan *et al.* [52], these results are somewhat unexpected, but gratifying. It could be a consequence of the intervening years having raised the prominence of data breaches in the media. This is further evidenced by the Marsh Report [36, p. 2] which reports that levels of basic or complete organisational understanding of cyber-risk rose from 60.8% in 2015 to 83.8% in 2016.

A number of insights emerged from our interviews, which we highlight here in order to inform our development of a framework to help SMEs to respond to breaches.

(1) Over-Emphasis on Technical Measures. Results from the interviews show that throughout the breach response process organisations displayed a disposition for implementing technical measures over non-technical ones. Moreover, in scenarios where organisations could not execute the measures themselves, technical expertise and solutions were prioritised. These results are unsurprising as they lend weight to the arguments presented by Shedden *et al.* [48] that organisations prioritise the use of technical measures because they believe data breaches are, by definition, a technical problem which demands a technical solution.

One participant said: '*technical changes are the quickest to implement for recovery. They may be a blunt instrument to address the problem but may be necessary.*' There was little reference to staff training or to the role of the human element in improving resilience to future attacks. Only in the follow-up section was this mentioned but this seemed more of a wish list than something that was actually implemented.

Best practice advice also makes a specific recommendation about technical aspects, but says little about awareness training for staff. They refer to a culture of security but don't say how this ought to be achieved.

This confirms reports from other researchers about incident responses appearing to place an unrealistic 'emphasis on technical competence in responding to incidents' [48, p. 133].

(2) Unthinking Responses. The best practice responses highlight the fact that preparation and a realistic expectation of being breached is important, in terms of knowing where assets are, and prioritising security.

Yet, in the responses about extant practice, there was some evidence that people would respond without really checking that the response would address the source of the breach. For example, requiring all staff to change their passwords before it has been confirmed that the attack vector involved a leaked password imposes significant burdens on staff without necessarily addressing the source of the breach.

The SMEs find it difficult to execute what they propose in theory throughout the entire data breach event. For example, during the preparation and follow up stages, organisations outline the importance of having plans in place and of learning lessons. However, in reality, organisations reported not using or ignoring existing plans during a breach and found learning lessons difficult after because of: a lack of expertise, lack of resources and skills to implement ideas in practice. This resulted in actions being designated to more skilled individuals, and the organisation's failure to prioritise cyber-security.

These results confirm research by Hove *et al.* [30] and Jaatun *et al.* [33] who found that organisations have incident response plans in place but that, in practice, these procedures were not well established. However, whilst current studies outline how organisations experience difficulties implementing theory into practice during each individual step of incident response, this research confirms that these trends are still evident throughout the entire breach response process.

This all points to a hasty and unmeasured response to data breaches, which means SMEs run the risk of carrying out the wrong actions and not dealing appropriately with breaches. Ineffective responses can have negative consequences. A prime example is UK telecoms company TalkTalk which lost an estimated 157000 customers' personal data [19]. The BBC reports that breaches at TalkTalk have cost the company up to £35 million in damages [8].

(3) Lessons are not Learned. Our interviews revealed an inconsistent and unreliable execution of lessons learnt. Some participants were indeed aware of the value of such an activity but pointed out the difficulties of doing this in the general melee of business life.

Researchers have highlighted the importance of a follow-up stage where lessons are learned to be commonplace within organisations [37, 59]. However, as Jaatun *et al.*'s [33] study of the petroleum industry shows, and we confirm, in practice challenges often arise that makes learning lessons difficult, and this deters their ability to respond to future incidents more effectively [3, p. 651].

5 MITIGATIONS

In proposing the mitigations we were mindful of the fact that SMEs have limited resources. In a more resource-rich organisation, these problems could be solved by hiring extra staff, or by contracting an external company to deal with any breaches that do occur. SMEs often do not have the luxury of these solutions. Hence we proposed mitigations here that would not require major expense and would essentially simplify the process. The main aim was to make it more manageable for solo responders who were not necessarily information security experts.

(1) Over-Emphasis on Technical Responses. Organisations believe that paying more attention to the human element i.e. having the right people in place before the breach, and working with individuals after the breach, is vitally important. These findings are interesting because, in practice, organisations prioritised technical measures, but when giving best practice advice there was an emphasis on prioritising measures addressing users. These findings lend support to research by Adams and Sasse [1] which promoted focus on the increasing importance of human elements within cyber-security research.

Incident response has to be holistic, addressing technical, managerial, legal and human aspects of information security [18]. The emphasis on technical responses is probably due to a measure of panic. Incident response is a stressful experience and Von Lubitz *et al.* [56] explain that, 'stress has a demonstrable negative effect on human information processing and interactions with chaotic environments'.

In helping SMEs to mitigate this tendency we are suggesting the use of checklists, commonly used in the medical field. Checklists providing easy-to-follow instructions to manage complex processes

[24, p. 120]. This technique, we believe, will be useful because the medical environment is also stressful and checklists might well benefit information security as much as it has been shown to benefit medical procedures by preventing omissions and thereby saving lives.

Gawande [24, p. 49] argues that checklists are an effective tool in these kinds of situations because they 'do not try to spell out everything' but instead act as a guide by providing reminders of only 'the most critical and important steps' [24, p. 120].

The Alien Vault's incident response guide [5, p. 21] argues that emergency contact checklists are valuable for maintaining communication with all the relevant stakeholders. Furthermore, checklists can also help maintain a paper trail during the breach because [34].

For SMEs, having checklists that encode essential incident response plans in an easy-to-process format constitutes an inexpensive way to provide valuable, structured and easy-to-understand guidance. This 'can prove highly beneficial as they can help ensure that personnel take prompt, consistent and holistic action under less than ideal conditions' [39]. The core plans that need to be encoded into checklists are [2, 49]: (1) Disaster Recovery Plan, (2) Crisis Communication Plan, and (3) Business Continuity Plan.

(2) Unthinking Responses. When an organisation has been breached, both 'co-ordination and timing' become serious concerns [24, p. 49]. To mitigate this, John Boyd's OODA loop, used extensively in the military, can prove beneficial because it 'provides the essential framework for knowledge-based multidimensional critical thinking and rapid decision-making' [56].

OODA has four stages: *Observe, Orient, Decide, Act.* **Observe** refers to actively absorbing the entire environment and changes that 'identify anomalous behaviour that may require investigation' [5, p. 18]. In the context of data breach response, responders will ask themselves key questions such as; 'What's normal activity on my network?' the better to understand the attack [46].

Orient is when information and knowledge gathered during *Observe* is broken down and assessed to introduce 'the first steps needed to re-organise it into the pre-disaster configuration' [56, p. 571].

The third stage is **Decide**, which refers to the responder defining 'the nature and characteristics of the action(s) to be taken' [56, p. 343]. In the context of incident response, responders assess different options obtained during the orientation stage to hypothesise the best course of action which ensures the 'fastest recovery' [5, p. 19].

The fourth stage is **Act** and refers to testing the proposed hypothesis made in the previous stage, to remediate and recover [5, p. 20] back online. The Act section is not the final part of the loop because the feedback from the action taken will form the basis of the next cycle of the loop.

For SMEs, using the OODA loop when responding to data breaches requires them to observe first, then orient and decide before they act. This, together with the checklists, ought to ensure a measured and more effective response to the breach.

(3) Lessons are not Learned. Current incident response frameworks, both in industry and academia [13, 53], perceive incident response cyclically i.e. a feedback-enabled lessons-learned loop feeding into the next incident response in order to improve the effectiveness of the responses.

The best way to do this is firstly to maintain a "lessons learned" database. The *lessons learned, or known errors, database* is a commonly used measure in organisations [54]. Sharif *et al.* [47] argue that it is critical for tacit knowledge be shared within organisations. In the context of incident response, the SME-specific incident response framework needs deliberately to incorporate a feedback loop, as originally proposed by Beer [9], to keep such a database current and helpful.

Other researchers in information security have argued for the need to learn lessons from data breach incident responses [3, 12, 27, 45]. Making such a feedback loop explicit in the framework will help to remind SMEs of the need to examine and learn from incident responses after the event.

This loop, together with the use of checklists to encode essential actions, make it easier to incorporate lessons learned into a simple, usable, and systematic form [24].

6 AN SME-SPECIFIC INCIDENT RESPONSE FRAMEWORK

SMEs need to develop a mind-set whereby a breach is expected at any time, and plan accordingly. The framework we suggest here incorporates the essential requirements of the GDPR, incorporated into checklists, and moderated by applying OODA instead of leaping in, in a panic.

Those actions that are required by **GDPR** are marked as such. The other items have been added specifically to help SMEs, mitigating their resource limitations. Although they have been marked as **SME-Specific** they would be helpful for large organisations too, but might not be necessary

Before the Breach

- *GDPR — Identify Business-Critical Resources and Sensitive Information:* Identify the 'crown jewels' (business-critical systems and personal customer information) to establish which areas need focused attention.
- *GDPR — Be Aware of Regulations:* The new GDPR regulations have to be complied with. Organisations have to ensure that they are aware of their responsibilities before any incident occurs. SMEs must familiarise themselves with the relevant notification regulations.
- ***SME-Specific*** *— Seek External Advice & Support:* Seek external support and knowledge from government initiatives and freely available advice guides. Boil security down to the simplest things. For example, implement the H.M Government 'Cyber Essentials' as a starting point [15].
- *GDPR — Assign Response Roles:* Decide if it is beneficial for cyber-security matters to be handled in-house or externally. By doing this, roles and responsibilities are clearly defined.
- ***SME-Specific*** *— Develop Checklists to guide Incident Response:* Compile checklists to help responders recall essential information regarding organisational processes. It is vital that three plans are encoded into checklists [2, 49]: (1) Disaster Recovery Plan, (2) Crisis Communication Plan, and (3) Business Continuity Plan.
- *GDPR — Carry out Security Awareness Training:* The employees of an organisation are an essential link in the information

security chain. Conducting regular awareness training is the only way to increase their resilience.

During the Breach Response

- ***SME-Specific*** *— (OODA) Observe, Orient, Decide, then Act:* ***First***, *Observe:* Responders gather information from the incident environment. ***Second***, *Orient:* Responders use information gathered to prioritise response actions. ***Third***, *Decide:* Responders use knowledge to hypothesise the best course of action to effectively respond to a breach. ***Finally***, *Follow the Disaster Recovery Plan:* Now, follow the checklist developed during the preparation phase.
- *GDPR — Document all Actions, with Timeline:* Checklists should be used to prompt responders regarding key processes and also to document every action taken. A paper trail is crucial.
- *GDPR — Report the Breach to the Supervisory Authority:* This must be done within 72 hours.
- *GDPR — Follow the Crisis Communication Plan:* Maintain communication with important internal stakeholders, regardless of whether the breach is being handled internally or externally. Employees are also stakeholders [7].
- ***SME-Specific*** *— Summon External Incident Response Support if Required:* If a breach is being handled internally and overwhelms resources use emergency external professional support.

After the Breach Response

- *GDPR — Reflect on Lessons Learned:* Irrespective of internal or external handling of breaches, evaluate the experience and ask questions such as: "What could be done better?"
- *GDPR — Feed Lessons back into Checklists:* Transform the main takeaways from the evaluation to refine the: (1) Disaster Recovery Plan, (2) Crisis Communication Plan, and (3) Business Continuity Plan.
- *GDPR — Boost Security Awareness:* Use the breach to boost security awareness and encourage individuals to learn lessons proactively through open forums.
- ***SME-Specific*** *— Do not Neglect the Humans:* Security is not just about technical measures. Work with and educate employees across the organisation regarding security using free advice e.g. H.M Government 'Cyber Essentials'.

6.1 Expert Review Feedback

Following ethical approval, see below, we sent the previous list of recommendations to four security experts. We asked them for feedback so that we could refine the recommendations

(1) *'I love where you are going with this. You need to get organisations into a state of preparedness by asking: Is it important to your business if someone can get into your computer and steal your customer information?*

(2) *'These are implementable, and scalable, you could strengthen them further by encouraging organisations to devote some effort to situational awareness – Is somebody responsible for understanding what sorts of threats are out there? The recent NHS malware incident is, a good example. Did the average organisation cotton on to that, and take precautionary steps?'*

(3) *'There is very little improvement, just more expansion is needed. The term 'crown jewels' refers to any business-critical systems the organisation relies on that would have a significant detrimental impact should they be unavailable. Look at other regulation such as: The NIS Directive and PECR.* Organisations should seek to incorporate lessons learnt into the organisation's security awareness programmes, for key incidents such as WannaCry, conduct 'ask me anything' type awareness-raising sessions allowing employees to understand more about these types of incident.

(4) *'It looks good, you could put a bit about policies and feeding lessons learned back into policy'*

Based on this feedback, the final incident response framework is presented in Figure 1, and Figure 2.

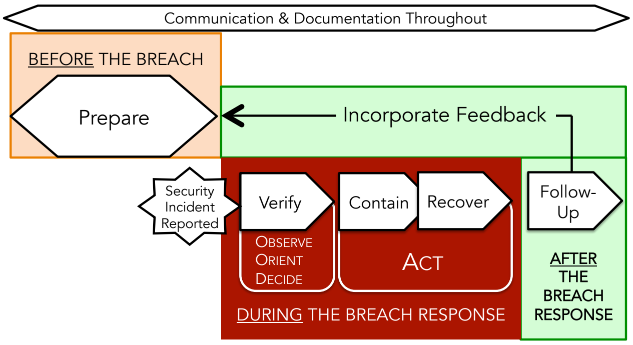

Communication & Documentation Throughout
BEFORE THE BREACH
Prepare
Incorporate Feedback
Security Incident Reported
Verify
Contain
Recover
Follow-Up
OBSERVE
ORIENT
DECIDE
ACT
DURING THE BREACH RESPONSE
AFTER THE BREACH RESPONSE

Figure 1: Final Incident Response Plan

7 DISCUSSION

The SMEs who participated in our study were well aware of data breaches, an improvement since 2003 [52]. Many were also aware of the fact that they ought to have some kind of plan in place to deal with any data breaches that did occur.

There was no broad agreement with respect to what the response to incidents ought to be. For example, during Containment, organisations discussed a variety of different actions with some suggesting isolating systems and others disagreeing with this approach. This confirms Grispos [25] assertions about the variability of incident responses. Morever, despite a general awareness that something ought to be done, and plans to lay down what the reaction should be, it did not seem that they tested or followed their plans when the need arose. This disconnect between awareness and action has already been commented on by other researchers [4].

We identified three particular themes that seemed to be getting in the way of SMEs responding as effectively as possible to data breaches, quite apart from their size and limited resources.

The *first*, an over-emphasis on technical security measures, was raised by Von Solms and Von Solms [57] in 2004, more than a decade ago and confirmed by [3]. It is disappointing to find that this kind of myopic focus is still prevalent in industry in 2017, when an increasing focus on the human's role in information security is becoming accepted by industry [1, 6, 51].

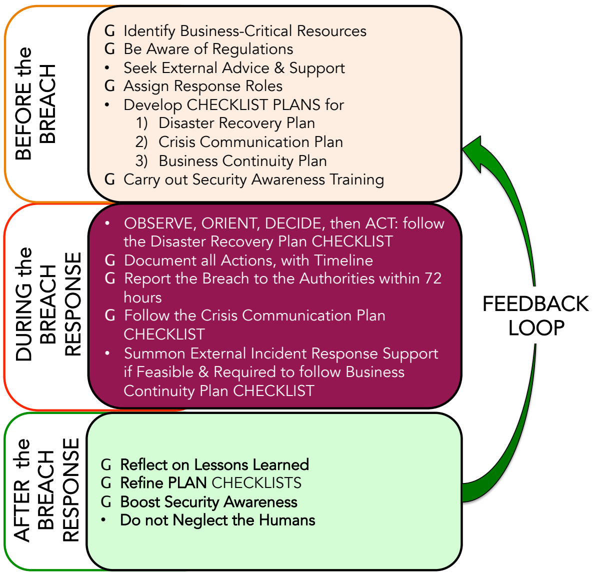

BEFORE the BREACH
G Identify Business-Critical Resources
G Be Aware of Regulations
• Seek External Advice & Support
G Assign Response Roles
• Develop CHECKLIST PLANS for
1) Disaster Recovery Plan
2) Crisis Communication Plan
3) Business Continuity Plan
G Carry out Security Awareness Training
DURING the BREACH RESPONSE
• OBSERVE, ORIENT, DECIDE, then ACT: follow the Disaster Recovery Plan CHECKLIST
G Document all Actions, with Timeline
G Report the Breach to the Authorities within 72 hours
G Follow the Crisis Communication Plan CHECKLIST
• Summon External Incident Response Support if Feasible & Required to follow Business Continuity Plan CHECKLIST
FEEDBACK LOOP
AFTER the BREACH RESPONSE
G Reflect on Lessons Learned
G Refine PLAN CHECKLISTS
G Boost Security Awareness
• Do not Neglect the Humans

Figure 2: SME-Specific Incident Response Framework (G=GDPR-Required Response)

The *second* problem is that organizations struggle to respond in a measured way. In our discussions with participants about containment it was obvious that they did not really respond in a consistent way. This is understandable because people are going to be stressed by the event [50, 58].

The *third* problem is that SMEs did not seem to have a mechanism for learning lessons from previous data breach events and responses.

The framework we propose specifically addresses each of these problems, using techniques that have proven successful in other disciplines: checklists (medicine) [24], OODA (the USA Military) [56] and an institutional lessons-learned archive resource (business knowledge management) [47, 54].

8 CONCLUSION

This paper set out to propose a feasible yet helpful framework to inform SME incident responses to data breach responses. We carried out a series of semi-structured interviews in order to inform the development of this data breach response framework. This SME-specific framework is different from others because it incorporates successful techniques from medicine (checklists) and the military (OODA), and explicitly incorporates a feedback loop to ensure that lessons are learned over the lifetime of an organisation. It is also relatively simple and not as heavy-weight as other best practice recommendations aimed at more resource-rich organisations.

This framework is not intended to be the final version. It clearly needs to be used "in anger" by SMEs and refined and improved based on their experiences. We present it here in order to gain feedback from other researchers. We hope to find SMEs who are willing to trial the framework and we hope thereby to refine it until it starts becoming a helpful resource. We believe that our deployment of proven techniques from other disciplines will prove helpful in incident responses too.

Our long-term aim is to support SMEs more effectively in coping with data breaches in the face of the coming GDPR legislation.

ETHICS

Ethical concerns centred on the need to maintain the anonymity of interviewees and the confidentiality of information they revealed. To address this plain language statements and consent forms were given to participants and offered participants anonymity and confidentiality. The ethics form was reviewed on 20 June 2017 (Application number: SP S/2017 SOCIAL SCIENCE/859) and approved subject to minor amendments which were adhered to.

REFERENCES

[1] Anne Adams and Martina Angela Sasse. 1999. Users are not the enemy. *Commun. ACM* 42, 12 (1999), 40–46.

[2] Steve Adubato. 2008. *What were they thinking?: Crisis communication: the good, the bad, and the totally clueless.* Rutgers University Press.

[3] Atif Ahmad, Justin Hadgkiss, and Anthonie B Ruighaver. 2012. Incident response teams–Challenges in supporting the organisational security function. *Computers & Security* 31, 5 (2012), 643–652.

[4] Eirik Albrechtsen. 2007. A qualitative study of users' view on information security. *Computers & Security* 26, 4 (2007), 276–289.

[5] Alien Vault. 2017. Insider's Guide to Incident Response – Expert Tips. (2017). https://www.alienvault.com/resource-center/ebook/insider-guide-to-incident-response (Accessed on: 18/05/2017).

[6] Debi Ashenden. 2008. Information security management: A human challenge? *Information Security Technical Report* 13, 4 (2008), 195–201.

[7] Debi Ashenden and Angela Sasse. 2013. CISOs and organisational culture: Their own worst enemy? *Computers & Security* 39 (2013), 396–405.

[8] BBC. 2015. TalkTalk hack to cost up to £35m. (2015). http://www.bbc.co.uk/news/uk-34784980 (Accessed on: 14/07/2017).

[9] Randall D Beer. 2000. Dynamical approaches to cognitive science. *Trends in Cognitive Sciences* 4, 3 (2000), 91–99.

[10] C Rustici. 2016. Don't think that Brexit will save you from the EU data protection rules. (2016). http://www.computerweekly.com/opinion/Dont-think-that-Brexit-will-save-you-from-the-EU-data-protection-rules (Accessed on: 23/08/2017).

[11] Cabinet Office. 2015. Cyber security 'myths' putting a third of SME revenue at risk. (2015). https://www.gov.uk/government/news/cyber-security-myths-putting-a-third-of-sme-revenue-at-risk (Accessed on: 15/07/2017).

[12] Julia S Cheney. 2010. Heartland Payment Systems: lessons learned from a data breach. (2010). FRB of Philadelphia - Payment Cards Center Discussion Paper No. 10-1 https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1540143.

[13] Paul Cichonski, Tom Millar, Tim Grance, and Karen Scarfone. 2012. Computer security incident handling guide. *NIST Special Publication* 800 (2012), 61.

[14] CREST. 2017. Cyber Security Incident Response Guide. (2017). https://www.crest-approved.org/wp-content/uploads/2014/11/CSIR-Procurement-Guide.pdf (Accessed on 26/05/2017).

[15] Cyber Essentials. 2018. (2018). https://www.cyberessentials.ncsc.gov.uk/ Accessed 20 February 2018.

[16] Deloitte. 2016. Incident Response: We've had a privacy breach – now what? (2016). https://www2.deloitte.com/content/dam/Deloitte/za/Documents/risk/ZA_CIC_Incident_Response_09062016.pdf (Accessed on: 23/08/2017).

[17] Department for Digital, Culture, Media and Sport. 2016. Cyber Security breaches survey. (2016). Available at: https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2017 (Accessed on: 06/05/2017).

[18] Sneza Dojkovski, Sharman Lichtenstein, and Matthew Warren. 2006. Challenges in fostering an information security culture in Australian small and medium sized enterprises. In *ECIW2006: proceedings of the 5th European conference on Information Warfare and Security.* Academic Conferences Limited, 31–40.

[19] Experian. 2015. Nearly 157,000 had data breached in TalkTalk cyber-attack. (2015). https://www.theguardian.com/business/2015/nov/06/nearly-157000-had-data-breached-in-talktalk-cyber-attack (Accessed on: 14/07/2017).

[20] Experian. 2016. SMEs under threat - The crippling consequences for unprepared small to medium sized businesses. (2016). http://www.experian.co.uk/assets/identity-and-fraud/smes-under-threat.pdf (Accessed on: 26/06/2017).

[21] Experian. 2017. Fourth annual 2017 Data Breach Industry Forecast. (2017). http://www.experian.com/assets/data-breach/white-papers/2017-experian-data-breach-industry-forecast.pdf (Accessed on: 16/07/2017).

[22] Experian Data Breach Resolution. 2014. Data Breach Response Guide. (2014). http://www.verizonenterprise.com/products/security/incident-response/ Accessed 27/12/2017.

[23] Kevvie Fowler. 2016. *Data Breach Preparation and Response: Breaches are Certain, Impact is Not.* Syngress, Cambridge, MA.

[24] Atul Gawande. 2011. *The Checklist Maniufesto.* Profile, London.

[25] George Grispos. 2016. *On the enhancement of data quality in security incident response investigations.* Ph.D. Dissertation. University of Glasgow.

[26] Alexander Harsch, Steffen Idler, and Simon Thurner. 2014. Assuming a state of compromise: A best practise approach for SMEs on incident response management. In *IT Security Incident Management & IT Forensics (IMF), 2014 Eighth International Conference on.* IEEE, 76–84.

[27] Ying He, Chris Johnson, Karen Renaud, Yu Lu, and Salem Jebriel. 2014. An empirical study on the use of the generic security template for structuring the lessons from information security incidents. In *Computer Science and Information Technology (CSIT), 2014 6th International Conference on.* IEEE, 178–188.

[28] H.M. Government. 2015. Cyber security: advice for small businesses. (2015). https://www.gov.uk/government/publications/cyber-security-what-small-businesses-need-to-know Accessed 3/1/2018.

[29] Houses of Parliament. 2017. Cyber Security of UK Infrastructure. (2017). Number 554, May. http://researchbriefings.files.parliament.uk/documents/POST-PN-0554/POST-PN-0554.pdf.

[30] Cathrine Hove, Marte Tarnes, Maria B Line, and Karin Bernsmed. 2014. Information security incident management: identified practice in large organizations. In *IT Security Incident Management & IT Forensics (IMF), 2014 Eighth International Conference on.* IEEE, 27–46.

[31] Information Commissioner's Office. 2017. Data security incident trends. (2017). https://ico.org.uk/action-weve-taken/data-security-incident-trends/ (Accessed on: 16/07/2017).

[32] International Standards Organisations. 2011. ISO/ IEC 27035, Information technology – Security techniques – Information security incident management. (2011). https://www.iso.org/standard/44379.html (Accessed on: 05/07/2017).

[33] Martin Gilje Jaatun, Eirik Albrechtsen, Maria B Line, Inger Anne Tøndel, and Odd Helge Longva. 2009. A framework for incident response management in the petroleum industry. *International Journal of Critical Infrastructure Protection* 2, 1 (2009), 26–37.

[34] Patrick Kral. 2011. SANS Institute The Incident Handlers Handbook. (2011). https://uk.sans.org/reading-room/whitepapers/incident/incident-handlers-handbook-33901 (Accessed on: 05/07/2017).

[35] Brian Krebs. 2017. 4 Years After Target, the Little Guy is the Target. (2017). 17 Dec https://krebsonsecurity.com/2017/12/4-years-after-target-the-little-guy-is-the-target/.

[36] Marsh. 2017. UK Cyber Risk Survey Report: 2016. (2017). https://www.marsh.com/content/dam/marsh/Documents/PDF/UK-en/UK%20Cyber%20Risk%20Survey%20Report%202016.pdf (Accessed on: 04/09/2017).

[37] Stefan Metzger, Wolfgang Hommel, and Helmut Reiser. 2011. Integrated Security Incident Management–Concepts and Real-World Experiences. In *IT Security Incident Management and IT Forensics (IMF), 2011 Sixth International Conference On.* IEEE, 107–121.

[38] Sarandis Mitropoulos, Dimitrios Patsos, and Christos Douligeris. 2006. On Incident Handling and Response: A state-of-the-art approach. *Computers & Security* 25, 5 (2006), 351–370.

[39] Emilio F Moran. 2016. *People and Nature: An Introduction to Human Ecological Relations.* Vol. 1. John Wiley & Sons.

[40] NCSC. 2017. Cyber Security: Small Business Guide. (2017). https://www.ncsc.gov.uk/smallbusiness Accessed 3/1/2018.

[41] NCSC. 2017. Professional service scheme Cyber Incidents. (2017). https://www.ncsc.gov.uk/scheme/cyber-incidents Accessed 3/1/2018.

[42] Briony J Oates. 2005. *Researching Information Systems and Computing.* Sage, London.

[43] Privacy Technical Assistance Centre. 2012. Data Breach Response Checklist. (2012). http://ptac.ed.gov/sites/default/files/checklist_data_breach_response_092012.pdf Accessed 17/12/2017.

[44] Chris Prosise, Kevin Mandia, and Matt Pepe. 2003. *Incident Response & Computer Forensics.* McGraw-Hill/Osborne New York.

[45] Karen Scarfone, Tim Grance, and Kelly Masone. 2008. Computer security incident handling guide. *NIST Special Publication* 800, 61 (2008), 38.

[46] Bruce Schneier. 2014. The future of incident response. *IEEE Security & Privacy* 12, 5 (2014), 96–96.

[47] Mohammad Nazir Ahmad Sharif, Nor Hidayati Zakaria, Lim Shu Ching, and Low Soh Fung. 2005. Facilitating knowledge sharing through lessons learned system. *Journal of Knowledge Management Practice* 12 (2005), 117–124.

[48] Piya Shedden, Atif Ahmad, and AB Ruighaver. 2010. *Organisational learning and incident response: promoting effective learning through the incident response process.* School of Computer and Information Science, Edith Cowan University, Perth, Western Australia.

[49] Susan Snedaker. 2013. *Business Continuity and Disaster Recovery Planning for IT Professionals.* Newnes, Amsterdam.

[50] Daniel J Solove and Danielle Citron. 2017. Risk and Anxiety: A Theory of Data Breach Harms. (2017). https://scholarship.law.gwu.edu/faculty_publications/1244/.

[51] Zahoor Ahmed Soomro, Mahmood Hussain Shah, and Javed Ahmed. 2016. Information security management needs more holistic approach: A literature review. *International Journal of Information Management* 36, 2 (2016), 215–225.

[52] Terence Tan, AB Ruighaver, and Atif Ahmad. 2003. Incident Handling: Where the need for planning is often not recognised. In *1st Australian Computer, Network & Information Forensics Conference.*

[53] Inger Anne Tøndel, Maria B Line, and Martin Gilje Jaatun. 2014. Information security incident management: Current practice as reported in the literature. *Computers & Security* 45 (2014), 42–57.

[54] Gertjan Van Heijst, Rob van der Spek, and Eelco Kruizinga. 1998. The lessons learned cycle. In *Information Technology for Knowledge Management.* Springer, 17–34.

[55] Verizon. 2017. Data Breach Digest. (2017). http://www.verizonenterprise.com/resources/reports/rp_data-breach-digest-2017-perspective-is-reality_xg_en.pdf (Accessed on: 16/07/2017).

[56] Dag Von Lubitz and Nilmini Wickramasinghe. 2006. Dynamic leadership in unstable and unpredictable environments. *International Journal of Management and Enterprise Development* 3, 4 (2006), 339–350.

[57] Basie Von Solms and Rossouw Von Solms. 2004. The 10 deadly sins of information security management. *Computers & Security* 23, 5 (2004), 371–376.

[58] Bryan Watkins. 2014. The impact of cyber attacks on the private sector. *Briefing Paper, Association for International Affair* (2014), 12.

[59] Rodrigo Werlinger, Kasia Muldner, Kirstie Hawkey, and Konstantin Beznosov. 2010. Preparation, detection, and analysis: the diagnostic work of IT security incident response. *Information Management & Computer Security* 18, 1 (2010), 26–42.

APPENDIX A

1. What company do you work for? (if you prefer not to be identified then anonymity will be respected)
2. What is your role in the company?
3. What does the term data breach mean to you?
4. Have you had any security breaches within your organisation? (if you prefer not answer then skip to question 6)
a. If yes, can you describe the breach? (i.e. DDOS) ....
b. How did you detect the breach? – what methods were used? (automatic or manual?)
c. How did you follow up the event, was the breach investigated? If so
d. Does your organisation have an incident response plan to use in case you get hacked?
If you have a plan –is it rehearsed,
If yes, how often, and what sort of vulnerabilities does it cover?)

5. Do you have a CSIRT team –
If yes, how does the team prepare?
if no team or plan – then why not?
6. Say you experience a hacking event. Could you say how you think you should respond?
7. Who should react first?
8. What actions should be taken to recover from the breach?
9. What actions should be prioritised?

What General Advice would you give to other companies?

1. What would be your top three incident response tips for an organisation that has suffered a breach?
2. How would you simplify the process of incident response?
3. How should lessons be learnt?
4. How could we ensure that companies can learn from attacks?
5. How do you think events could change security attitudes within your organisation?
6. How important do you think preparation and having a pre-determined plan is in terms of being able to deal with a breach?

How to Design Browser Security and Privacy Alerts

Lynsay A. Shepherd
School of Design and Informatics
Abertay University
Dundee, United Kingdom
lynsay.shepherd@abertay.ac.uk

Karen Renaud
School of Design and Informatics
Abertay University
Dundee, United Kingdom
k.renaud@abertay.ac.uk

ABSTRACT

It is important to design browser security and privacy alerts so as to maximise their value to the end user, and their efficacy in terms of communicating risk. We derived a list of design guidelines from the research literature by carrying out a systematic review. We analysed the papers both quantitatively and qualitatively to arrive at a comprehensive set of guidelines. Our findings aim to to provide designers and developers with guidance as to how to construct privacy and security alerts. We conclude by providing an alert template, highlighting its adherence to the derived guidelines.

1 INTRODUCTION

It is non-trivial to design effective alerts in the security and privacy domain.

Browser designers do their best to inform users about security-related aspects as they surf the web. Owing to the number of potential pitfalls, this means end users can be bombarded with security alerts [2], and users often ignore them [8, 33]. Developers sometimes make unfounded assumptions about the background knowledge of alert recipients [30], which can make the alerts incomprehensible.

Privacy alerts also have flaws, both in terms of content and delivery [35]. Users are often overwhelmed by these alerts because there are too many [21], or because they do not know what actions to take as a consequence [56]. This diminishes the impact of alerts, and leaves users vulnerable to unknowingly carrying out actions which will compromise their privacy or security.

Traditional usability guidelines cannot necessarily be used "as-is" in the security and privacy context. This is because neither privacy nor security are the end user's primary task [11, 32]. Alerts interrupt the user's pursuit of their primary goal and are thus often perceived to be a nuisance [5]. We therefore need specific guidelines to inform alert design in the security and privacy context.

Much has been written about alert design, highlighted by the following sections. Yet one can hardly expect busy deadline-driven software engineers, the very people who are producing these alerts, to keep up with the latest research.

We therefore performed a systematic literature review to consolidate all the published guidelines into one coherent list (Section 3). Previously, Bauer *et al.* [13] presented a list of warning design guidelines in 2013. Our work provides an updated, more comprehensive, list of guidelines, specifically tailored towards browser-based alerts.

Having derived a comprehensive set of guidelines, we realised that merely providing a list of guidelines is not an optimal way of supporting designers. Luger and Rodden [38] argue that such lists of guidelines are unlikely to be followed in the pressured environment of software development and design. Moreover, some of the published guidelines conflict [39], which is unhelpful.

To make our consolidated guidelines as helpful as possible, we decided to convey the *spirit*, rather than the *letter*, of the guidelines in the form of an example alert template (Section 4). This conveys the "how" of alert design, rather than the "what", encapsulated in a linear set of alert guidelines. Future work is explored in Section 5, and we conclude in Section 6.

2 INFORMING END USERS

First we clarify the nomenclature used in this paper. We then provide an overview of the human in the loop model of human information processing. We conclude by explaining the difference between the foundational security and privacy concepts.

AISB'18, April 4–6, 2018, Liverpool, UK

2.1 Nomenclature

We investigated guidelines that inform the design of warnings, alerts, notifications, prompts or provision of feedback. The underlying concept is the same: provision of important information to an end user that the system considers he or she should be apprised of. We shall use the term 'alert' as a unifying term to represent all the terminologies used by papers cited in this paper.

2.2 Human Information Processing

Wogalter and Mayhorn [63] explain that warnings (what we call alerts) are a type of risk communication. Wogalter [64] explains that warnings have two purposes: (1) communicate information, and (2) reduce unwise behaviours. To achieve these aims the warnings have to be designed carefully. The Web Content Accessibility Guidelines[1] can also be applied to alerts [7] i.e. that they should be perceivable, operable, understandable and robust.

Shannon [57] and Lasswell [36] both proposed models of human communication which help us to understand how humans process alerts.

Wogalter, DeJoy, and Laughery [65] developed the C-HIP model in the context of warning research. Their model builds on the work of Shannon and Lasswell and can be considered to be somewhat unrealistic because it does not include a noise component. In a world of noisy communication, such a model is incomplete. Cranor [17] proposed a 'human-in-the-loop' framework which is more comprehensive and reflects the factors impacting communications in the context of privacy and security alerts.

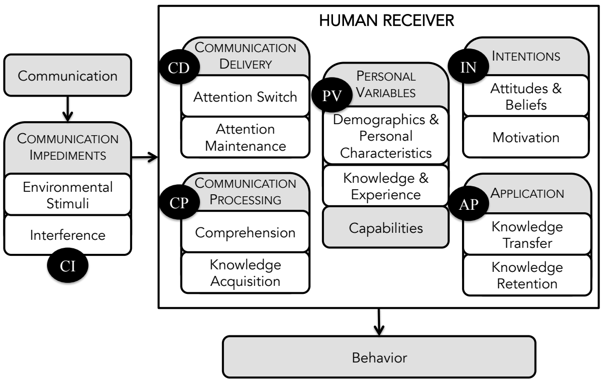


Figure 1: Cranor's Human in the Loop Security Framework [17] (Layout tweaked due to space constraints, and acronyms added for later reference).

2.3 Security vs. Privacy

It is important to realise that security and privacy are fundamentally different concepts. Skinner *et al.* [59] argue that a secure information system does not necessarily imply that privacy will be preserved in the system. Gritzalis and Lambrinoudakis [24] make the distinction between privacy and security as follows: "*a piece of information is secure when its content is protected, it is private when the identity of the owner is protected*". As an example, they refer to a company that collects customer information, and stores it in an encrypted format. This ensures that the information is secured. Yet the same company may sell the information to another company, thereby violating the owners' privacy.

Bambauer [12, p. 667] explains: "*Privacy discourse involves difficult normative decisions about competing claims to legitimate access to, use of and alteration of information.*" Security, on the other hand, is "*the protection of information and information systems from unauthorised access, use, disclosure, disruption, modification, or destruction in order to provide confidentiality, integrity, and availability.*" [20].

Privacy and security are clearly distinct concepts, but their alerts still share some common characteristics in that they exist to tell the end user to something important. We therefore present three lists of guidelines: (1) generic, (2) privacy-specific, and (3) security-specific.

3 CONTEMPLATING THE ALERT LITERATURE

We decided to focus on browser alerts firstly because of the popularity of web applications [41] such as email, claimed to be the most popular application in use [6] and video streaming [31]. The second reason is that browsers run on all devices, ranging from Desktops to Smartphones. We felt that our guidelines could be maximally useful to developers if we focused on guidelines for browser alerts.

We carrried out a systematic literature review, as advised by [50]. The literature search was carried out between November and December 2017 as follows:

Databases: ACM, Springer, Web of Science, Scopus, IEEE, and then Google Scholar to identify publications that did not appear in the other databases.

Keywords: 'design guidelines' *and* 'browser' *and* ('security' *or* 'privacy') *and* ('feedback' *or* 'warnings' *or* 'alert' *or* 'notification')

Time Range: 2007—2017

Exclusion Criteria: Patents, citations, non-peer reviewed, not English or unobtainable.

3.1 Quantitative analysis

One particular measure of activity in a research field is the number of papers published over the decade in question.

[1] https://www.w3.org/TR/WCAG21/

Database	Papers Returned	Papers Excluded	Papers Analysed
Scopus	2	1	1
ACM	12	9	3
Springer	214	199	15
Web of Science	0	0	0
Google Scholar	181	134	47
IEEE	79	73	6
Total			72

Figure 2 shows the number of papers, and also how many times the papers have been cited up to the date we carried our our literature review.

It is interesting to note that 25 of the 72 papers had no citations at all. The average number of citations is 7.38, the mode is 0, and the median is 2. Only four of the papers had been cited by more than 50 other publications. The top two most-cited publications appeared in conferences and the third most-cited publication appeared in a journal.

Figure 3 shows the citations for papers in each of the paper focus areas. The top cited paper is a security paper, with the next two most-cited papers being in the privacy area.

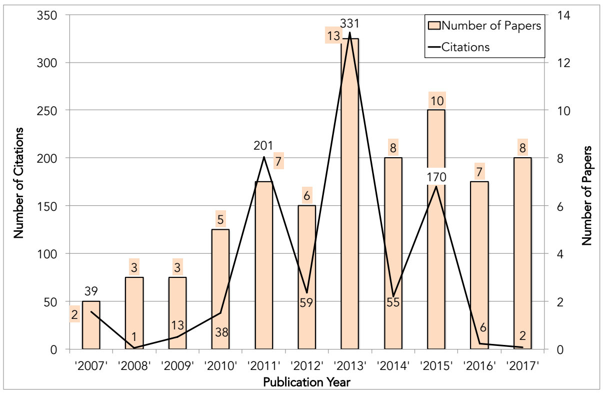


Figure 2: Number of Papers and Citation Numbers per Year

It is interesting that so many of the guidelines appear in Masters and PhD theses (18). Of these, 10 were never cited. The most-cited thesis, a PhD, was cited 13 times. Eight of the 10 PhDs had never been cited. The average number of citations across all theses was 2.47, but the mode and median are both 0. This suggests that guidelines published in these formats have not made a significant impact on the field.

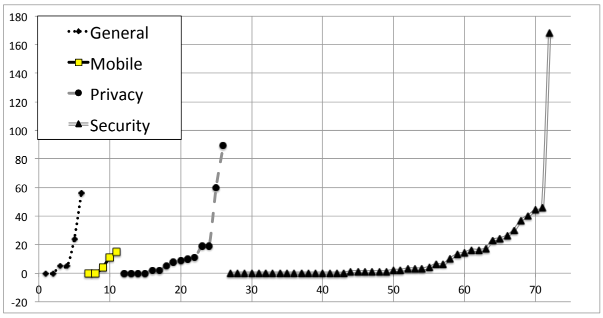


Figure 3: Number of Citations per Paper (by paper focus)

3.2 Qualitative analysis

We analysed the guidelines using Thematic Analysis [25]. This approach supports pinpointing, examining, and recording themes that emerge from the papers. We commenced by familiarising ourselves with the papers. We then generated initial codes and searched for themes as we collated these codes. We then reviewed the themes, defining and naming them.

Some of the guidelines applied equally to privacy and security, but others were clearly specific to either privacy or security. This is not unexpected because, as argued earlier, privacy and security are fundamentally different concepts.

We shall thus present generic guidelines first, then concept-specific guidelines separately, reflecting the fundamental differences between the two concepts.

3.3 Generic Guidelines

We report first on the generic themes that coincide with Cranor's framework [17], depicted in Figure 1.

Communication Impediments (CI). Here mitigations to ameliorate the effects of impediments should be included. For example, provide users with the means to recover from hasty decisions [29].

Personal Variables & Capabilities (PV). Some users may have low numeracy levels. Instead of providing them with figures regarding risk, perhaps utilise emotions or mood. Similarly, users may have different understanding of visuals [45]. Only by testing can the efficacy of these be confirmed.

Communication Delivery (CD). Human attention is a finite resource [16]. Do not squander it, and do not expect the recipient to give you any as a matter of course.

CD1: Modality — Murphy-Hill & Murphy [42, 62] suggest that pictures be used to ease communication. When delivering warning alerts, users prefer text and graphical-based information, rather than auditory information[15].

On the other hand, Goldberg [23] suggests that text should be used exclusively. This might be to maximise accessibility, and the middle road would be to ensure that alt text is provided for all images.

Work conducted by Anderson *et al.* [9, 10] notes the use of polymorphism in warning alerts to reduce habituation.

CD2: Timing — If alerts appear too often, the recipients may get annoyed and start ignoring them [4, 42]. Alerts should be issued only when necessary, to avoid irritating the user [61].

Westermann [62] found that people were most annoyed by alerts if they are busy with a task, especially when the task is complex, but less annoyed in between tasks or when they were waiting for something.

It is important to prioritise the warnings so that only the most important ones merit immediate interruption [4, 62].

CD3: Location — Westermann [62] considers alert location to be important. Many browsers, for example, display alerts either in the address bar (padlock and the word 'Secure' in Chrome), or at the bottom of the screen. These are easily missed by users. If we want people to notice the alerts it ought to be displayed where they are currently focusing their attention. In particular [4] point out that passive toolbar-located warnings are less effective than full page warnings. Pala and Wang [48] also suggest alerts should be placed where the users are focusing their attention. In the study conducted by Chen, Zahedi, and Abbasi, users preferred alerts to be placed in the centre of the screen [15].

CD4: Appearance — Kelley [32] provides a number of recommendations: (1) the alert should be surrounded by a box to clearly demarcate it; (2) provide a title to facilitate speedy recognition. Be careful with colour use so as not to disadvantage those with colour deficiencies [23]. A neutral grey colour can be used for the background of alerts, as it is unlikely to annoy the user [61].

Communication Processing (CP).

CP1: Make Essential Information Pertinent — Lin [37] suggests highlighting the most important information. Keep initial details about the risk to a minimum [45, 46]. Only the most important information should be displayed to the user immediately with links to more information should they want it [61]. The granularity of information is important. Wordy information will not be read, and information that is too condensed can be obscure. In providing alerts, a balance must be found [11].

CP2: Maximise Understandability and Consistency — Alerts should maximise understandability [32, 42, 55, 62], and should be presented in a consistent format [7, 42]. Provide concrete explanations [47]. The importance of this aspect is confirmed by [44]. Keep explanations simple [37]. Acronyms and jargon should be avoided and the use of meaningful terminology encouraged [11, 32, 55, 58]. Separate semantically different kinds of information [32, 60].

Presented text should be easy for users to comprehend [61]. Short, simple sentences, devoid of complex grammatical structures, should be used. The use of technical words should be avoided (i.e. words listed in the indexes of IT security books) [18, 26, 27, 48]. Unclear alerts are more likely to be ignored, and consideration should be given to the exact meanings of words used [45].

Work by Bravo-Lillo *et al.* [14] investigated the use of redesigned warning alerts. Longer warning alerts performed poorly in user testing, suggesting users may have become confused. Although existing work highlights that shorter alerts are most effective at communicating security warnings to the user, the challenge of delivering such alerts whilst providing the user with enough text to foster understanding has been acknowledged [22].

Application (AP).

AP1: Be Specific — Bravo-Lillo *et al.* [14] state that *"to be successful, warnings should both motivate a user to respond, and help users understand the risk, in that order"*. Always tell the users what actions to take, if indeed they should take action [7, 29].

AP2: End Goal — Consider the way in which you want to communicate a risk to the user e.g. is the alert to draw them away from a risky situation, or is the alert to help them to understand the risk [45]?

AP3: Effort Does Not Deter — Akhawe and Felt [4] explain that designers should not use the number of clicks required to bypass a warning to deter users. Their study showed that users were not sensitive to the number of clicks once they had made a decision.

Intentions (IN).

It is important to note that delivering warnings is worthwhile. Silic *et al.* [58] found that people took note of displayed warning messages, suggesting that they thought about the information before making the decision to proceed. If people are reading and thinking about messages, these messages have a chance of changing attitudes and beliefs.

Vasalou [60] says alerts should give recipients "space for interpretation", so that they can interpret the information as it applies to themselves personally.

Phrasing of alerts could be personalised, depending on the user's skill level, and experience [15, 45, 46]. Personalised alerts were said to be successful when used to inform users about two-factor authentication, and bullet-points can be used to aid clarity of information presented [52].

It is important for the user to retain a level of control [29, 60, 66]. Schaub *et al.* [54] distinguish between three levels of user control: (1) blocking, (2) non-blocking and (3) decoupled. A designer has to decide whether the user has to acknowledge the message (blocking) or not (non-blocking), whether they can defer it (decoupled), or whether it will expire after displaying for a certain period of time [42].

Users should be provided with the option to respond to a risk they have been alerted to, and helped to visualise potential consequences [45]. Work by Volkamer *et al.* [61] concurs that the potential consequences of a risk should be conveyed to the user, along with potential recommendations.

Make sure the user can easily get in touch with someone to ask about warning-type alerts [23]. Contact information should be conspicuously placed so worried users will be able to get help [29].

3.4 Privacy-Specific Guidelines

Allow users to make privacy choices that are (1) meaningful, (2) informed, (3) timely [16].

P-CI: Inspire Trust — Trust should be deliberately built and maintained [42] by framing the privacy alert very carefully [3]. Rather counter-intuitively, privacy alerts should not provide justifications for information requests. Researchers report that justifications potentially reduce the end-user's trust in the system [1, 3, 34, 51].

P-PV: Privacy Expectations — Lin [37] points out that users have different privacy expectations, and that an alert interface should reflect this reality.

P-CD: Specificity — Ensure that the sensitivity of the data is communicated to the user [43].

P-IN: Enhance Control — Ensure that control resides with the user [43]. Do not merely report that some privacy invasion has occurred: allow the user to control disclosure. It is necessary to balance the number of interruptions with ensuring that the user retains a sense of control [16].

People have different levels of privacy concerns, and the alerts should afford users the level of control matching their personal privacy concerns.

3.5 Security-Specific Guidelines

Herzog and Shahmehri [28] highlight the importance of security features in applications, stressing that *"security is rarely the primary user task"*.

S-CI: Context-Sensitive Help — Constantly visible context sensitive help may prove useful in helping the user understand security. Help may be provided via the use of an agent [28]. The user should be provided with the option to find further information in a contextually-aware setting [46, 48].

S-CD1: Provide Justification — The user needs to know why the alert is being provided [42].

Provide information as to whether a component is secure or insecure. By displaying this information in either case, this provides a consistent interface for the user [48]. Ensure that the current state of the system is displayed to the user [46].

S-CD2: Colour — Research regarding two-factor authentication suggests the use of blue as a peaceful colour. Red might indicate an incident has occurred [15, 52]. Felt *et al.* [22] suggest utilising *"opinionated design"*. For example, make the "correct" response the more visually appealing option e.g. the button should have a high contrast level against the background. Others have utilised green as colour, noting that it is seen as safe. Whilst users should be given options regarding how to proceed with their tasks, it has been suggested that placing the "correct" option in green serves to guide users towards the safe choice [61].

Where colourblind users may have issues with warnings, the use of secondary information (icons) can convey the same message [61].

S-CD3: Graphics — In one study, participants felt the inclusion of graphics in an alert about two-factor authentication conveyed a tone which was less serious, and suspicious [52]. Conversely, other studies conclude graphics are required in alerts, to convey reassurance, draw attention, and to reduce cognitive effort [61]. This is a prime example of conflicting guidelines.

Eargle [19] suggests that facial expressions could be used to convey threat levels in security alerts but this has not been confirmed by any other studies in our studied group.

S-IN: Control Level — If a security issue is detected on a page, users would prefer the security alert to block them from visiting a malicious website [15]. Other research stated the final security decisions should be left to the user, though users should be provided with alternative options on how to proceed with their task [61].

4 INFORMING DESIGNERS

The previous section provided a list of recommendations for designing alerts. In Figure 4, the solid black boxes denote privacy guidelines (prefixed with 'P'), whilst the grey-scale boxes denote security guidelines (prefixed with 'S'). Privacy and security guidelines include communication impediments (P-CI, S-CI), communication delivery (P-CD, S-CD), personal variables (P-PV), and intentions (P-IN, S-IN).

However, as pointed out by [53], and confirmed by [38], designers have difficulty benefiting from these kinds of flat lists of guidelines.

Luger and Rodden's [38] designers spoke about the value of examples in encapsulating the lessons of design guidelines in a more palatable format. We thus present a template in Figure 5, and explain how it satisfies the guidelines,extending the initial template produced by Bauer *et al.* [13].

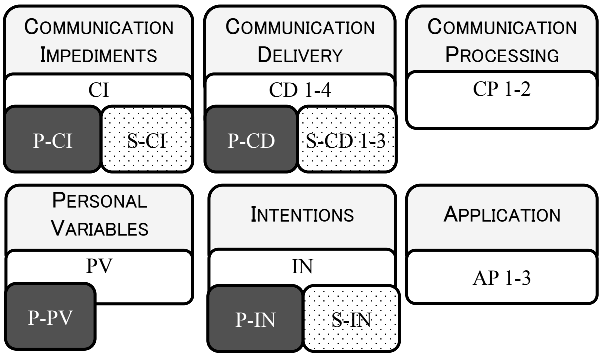
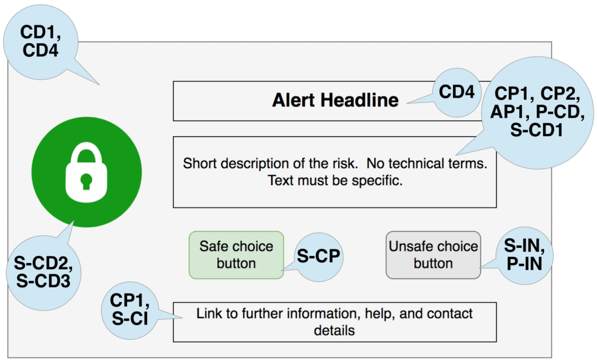


Figure 4: Consolidated Guidelines



Short description of the risk. No technical terms. Text must be specific.

Link to further information, help, and contact details

Figure 5: Example alert template

Generic Guidelines: The template contains both an icon, and text to communicate the contents of the alert (CD1). The alert has a border and a headline title, along with the use of colours and icons. The background colour of the template is neutral (CD4). The text explaining the alert should be clear, specific, and easy to understand, requiring minimal cognitive effort (CP1, CP2, AP1). If a user requires more information, they should be presented with the opportunity to access this, along with relevant contact details.

Privacy Guidelines: If the alert is being used to notify the user about privacy, the sensitivity of the information being dealt with should be clearly communicated to the user (P-CD). Users are provided with a choice in the alert, ensuring they remain in control (P-IN).

Security Guidelines: The text explaining the alert should justify why it is being displayed (S-CD1). The safe choice button on the template alert is more visually appealing than the unsafe choice, and it clearly contrasts with the background of the alert (S-CP). Users are provided with a choice in the alert, ensuring they remain in control (S-IN). Users should be presented with the option to access context-sensitive help. Colour and graphics should be used to aid in communicating the role of the alert, ensuring colourblind users are not placed at a disadvantage (S-CD2, S-CD3).

Template Summary: Graphics and text are used to communicate the nature of the alert (CD1, CD4, S-CD2, S-CD3). A headline title and a neutral background are used (CD4). The text explaining the alert should be clear, specific, and easy to understand, requiring minimal cognitive effort (CP1, CP2, AP1, PC-D, S-CD1). Users should have the opportunity to access further information, and relevant contact details (CP1), along with context-sensitive help (S-CI). Users are provided with a choice in the alert, ensuring they remain in control (P-IN, S-IN). The safe choice button is more visually appealing than the unsafe choice, clearly contrasting with the background of the alert (S-CP).

4.1 Development Good Practice

Creating a well-designed environment can aid in establishing trust [40]. Moreover, it is important to ensure that people are receptive to alerts [62]. The best way to confirm both trustworthiness and alert receptiveness is by means of thorough testing [29]. Options are A/B testing in the wild, controlled experiments, field studies [4], or case studies post-deployment [42].

5 FUTURE WORK

The systematic literature review identified a lack of research surrounding the optimal placement of security and privacy alerts within the web browser. Whilst work carried out by Chen, Zahedi, and Abbasi [15] showed users preferred alerts in the centre of the screen, usability studies have shown there are a variety of patterns users exhibit when browsing web content [49]. This suggests that further research is required into the optimal placement of security and privacy alerts.

It is also interesting to note from Figure 4 that there are no security or privacy-specific guidelines in terms of Communication Processing or Application. Furthermore, few of the papers returned in the systematic literature review mentioned consistency of alert design. These are certainly areas for further investigation.

Several guidelines gathered from literature conflict, and this issue has been highlighted by other guideline papers [53]. Previous research has acknowledged that *"Not all best practices can be simultaneously satisfied"*; therefore, trade-offs must occur [22]. Masip *et al.* [39] have investigated the development of a design process to assist with design choices when there are potentially conflicting user interface guidelines. In the future, we plan to develop a methodology for prioritising the guidelines to support security and privacy alert design.

6 CONCLUSION

The systematic review process highlighted a large proportion of the work found online relating to alerts were sourced from student theses (both at Masters and PhD level). Whilst conducting the analysis process, it became clear that some alert guidelines were developed for security, and others were developed for privacy. These seemed, in many cases, to be fundamentally different, suggesting that different guidelines are required for these two distinct areas.

We publish this work as a first attempt to provide guidance to designers and developers who need to incorporate alerts into their systems. In the future, we seek to prioritise the guidelines, addressing the issue of potential conflicts, and with feedback from practitioners, iteratively refine the guideline list.

REFERENCES

[1] M Aagaard. 2013. How Privacy Policy Affects Sign-Ups — Surprising Data From 4 A/B Tests. (2013). ContentVerve.com.

[2] Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. 2016. You are not your developer, either: A research agenda for usable security and privacy research beyond end users. In *Cybersecurity Development (SecDev), IEEE.* IEEE, 3–8.

[3] Idris Adjerid, Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2013. Sleights of privacy: Framing, disclosures, and the limits of transparency. In *Proceedings of the Ninth Symposium on Usable Privacy and Security.* ACM, 9.

[4] Devdatta Akhawe and Adrienne Porter Felt. 2013. Alice in Warningland: A Large-Scale Field Study of Browser Security Warning Effectiveness.. In *USENIX Security Symposium*, Vol. 13.

[5] Eirik Albrechtsen. 2007. A qualitative study of users' view on information security. *Computers & Security* 26, 4 (2007), 276–289.

[6] Saad Alharbi and Dimitrios Rigas. 2008. Graphical browsing of email data: An empirical investigation. In *Information Technology: New Generations, 2008. ITNG 2008. Fifth International Conference on.* IEEE, 495–499.

[7] Leonelo Dell Anhol Almeida and Maria Cecília Calani Baranauskas. 2010. *Merging Technical Guidelines for Accessible Web Content with Universal Design Principles.* Technical Report IC-10-020.

[8] Bonnie Anderson, Tony Vance, Brock Kirwan, David Eargle, and Seth Howard. 2014. Users aren't (necessarily) lazy: using neuroIS to explain habituation to security warnings. In *Thirty Fifth International Conference on Information Systems.* Auckland.

[9] Bonnie Brinton Anderson, C Brock Kirwan, Jeffrey L Jenkins, David Eargle, Seth Howard, and Anthony Vance. 2015. 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.* ACM, 2883–2892.

[10] Bonnie Brinton Anderson, Anthony Vance, C Brock Kirwan, Jeffrey L Jenkins, and David Eargle. 2016. From warning to wallpaper: Why the brain habituates to security warnings and what can be done about it. *Journal of Management Information Systems* 33, 3 (2016), 713–743.

[11] Rebecca Balebako, Jaeyeon Jung, Wei Lu, Lorrie Faith Cranor, and Carolyn Nguyen. 2013. Little brothers watching you: Raising awareness of data leaks on smartphones. In *Proceedings of the Ninth Symposium on Usable Privacy and Security.* ACM, 12.

[12] Derek E Bambauer. 2013. Privacy versus security. *J. Crim. L. & Criminology* 103 (2013), 667.

[13] Lujo Bauer, Cristian Bravo-Lillo, Lorrie Cranor, and Elli Fragkaki. 2013. *Warning Design Guidelines.* Technical Report CMU-CyLab-13-002. Carnegie Mellon University.

[14] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie S. Downs, Saranga Komanduri, and Manya Sleeper. 2011. Improving Computer Security Dialogs. In *Human-Computer Interaction - INTERACT 2011 -13th IFIP TC 13 International Conference, Lisbon, Portugal, September 5-9, 2011, Proceedings, Part IV.* 18–35.

[15] Yan Chen, Fatemeh Zahedi, and Ahmed Abbasi. 2011. Interface Design Elements for Anti-phishing Systems. In *Proceedings of the 6th International Conference on Service-oriented Perspectives in Design Science Research (DESRIST'11).* Springer-Verlag, Berlin, Heidelberg, 253–265.

[16] Jessica Helena Colnago. 2016. *Privacy agents in the IoT: considerations on how to balance agent autonomy and user control in privacy decisions.* Ph.D. Dissertation. Universidade Federal de São Carlos.

[17] Lorrie Faith Cranor. 2008. A framework for reasoning about the human in the loop. *UPSEC* 8, 2008 (2008), 1–15.

[18] Xun Dong, John A. Clark, and Jeremy L. Jacob. 2010. Defending the weakest link: phishing websites detection by analysing user behaviours. *Telecommunication Systems* 45, 2-3 (2010), 215–226.

[19] David W. Eargle. 2017. Security Messages: Or, How I Learned to Stop Disregarding and Heed the Warning. (August 2017). http://d-scholarship.pitt.edu/31614/

[20] Richard Kissel (Ed.). 2013. *Glossary of Key Information Security Terms.* Technical Report NISTIR 7298 Revision 2.

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

[22] Adrienne Porter Felt, Alex Ainslie, Robert W. Reeder, Sunny Consolvo, Somas Thyagaraja, Alan Bettes, Helen Harris, and Jeff Grimes. 2015. Improving SSL Warnings: Comprehension and Adherence. In *Proceedings of the Conference on Human Factors and Computing Systems.*

[23] Jeffrey S Goldberg. 2009. *State of Texas Municipal Web Sites: A Description of Website Attributes and Features of Municipalities with Populations Between 50,000-125,000.* Master's thesis. Public Administration.

[24] Stefanos Gritzalis and Costas Lambrinoudakis. 2008. Privacy in the digital world. In *Encyclopedia of Internet Technologies and Applications.* IGI Global, 411–417.

[25] Greg Guest, Namey MacQueen, and EE Namey. 2012. Introduction to thematic analysis. *Applied Thematic Analysis* 12 (2012).

[26] Marian Harbach, Sascha Fahl, Thomas Muders, and Matthew Smith. 2012. Towards Measuring Warning Readability. In *Proceedings of the 2012 ACM Conference on Computer and Communications Security (CCS '12).* ACM, New York, NY, USA, 989–991. https://doi.org/10.1145/2382196.2382301

[27] Marian Harbach, Sascha Fahl, Polina Yakovleva, and Matthew Smith. 2013. Sorry, I Don't Get It: An Analysis of Warning Message Texts. In *Proceedings of the 2013 International Conference on Financial Cryptography and Data Security (FC13), Workshop on Usable Security (Lecture Notes in Computer Science).*

[28] Almut Herzog and Nahid Shahmehri. 2007. User Help Techniques for Usable Security. In *Proceedings of the 2007 Symposium on Computer Human Interaction for the Management of Information Technology (CHIMIT '07).* ACM, New York, NY, USA, Article 11.

[29] John Sören Petterson (Ed.). 2008. HCI Guidelines. (2008). PRIME (Privacy and Identity Management for Europe) EU Project Report.

[30] Michaela Kauer, Thomas Pfeiffer, Melanie Volkamer, Heike Theuerling, and Ralph Bruder. 2012. It is not about the design - it is about the content! Making warnings more efficient by communicating risks appropriately. (2012).

[31] Aharon Kellerman. 2010. Mobile broadband services and the availability of instant access to cyberspace. *Environment and Planning A* 42, 12 (2010), 2990–3005.

[32] Patrick Gage Kelley. 2009. Designing a privacy label: assisting consumer understanding of online privacy practices. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. ACM, 3347–3352.

[33] Soyun Kim and Michael S Wogalter. 2009. Habituation, dishabituation, and recovery effects in visual warnings. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 53. Sage Publications Sage CA: Los Angeles, CA, 1612–1616.

[34] Bart Piet Knijnenburg. 2015. *A User-Tailored Approach to Privacy Decision Support*. Ph.D. Dissertation. UC Irvine Information and Computer Sciences.

[35] Robert LaRose and Nora J Rifon. 2007. Promoting i-safety: effects of privacy warnings and privacy seals on risk assessment and online privacy behavior. *Journal of Consumer Affairs* 41, 1 (2007), 127–149.

[36] Harold D Lasswell. 1948. The Structure and Function of Communication in Society. *The Communication of Ideas* 37 (1948), 215–228.

[37] Jialiu Lin. 2013. *Understanding and capturing people's mobile app privacy preferences*. Ph.D. Dissertation. Carnegie Mellon University.

[38] Ewa Luger and Tom Rodden. 2014. The value of consent: Discussions with designers of ubiquitous computing systems. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*. IEEE, 388–393.

[39] Llúcia Masip, Célia Martinie, Marco Winckler, Philippe A. Palanque, Toni Granollers, and Marta Oliva. 2012. A Design Process for Exhibiting Design Choices and Trade-Offs in (Potentially) Conflicting User Interface Guidelines. In *Human-Centered Software Engineering - 4th International Conference, HCSE 2012, Toulouse, France, October 29-31, 2012. Proceedings*. 53–71.

[40] Ricardo Mendoza-González and Sandra Jimenez-González. 2015. Guidelines to Design Usable Security Feedback for Identity Management Applications. In *Mathematical Methods and Systems in Science and Engineering*. 256–264.

[41] Michael S Mikowski and Josh C Powell. 2013. *Single Page Web Applications*. Manning Publications.

[42] Emerson Murphy-Hill and Gail C Murphy. 2014. Recommendation delivery. In *Recommendation Systems in Software Engineering*. Springer, 223–242.

[43] Stefan Nafra. 2014. *Aligning Privacy and Usability: Designing a Privacy-Aware Mobile Application that People Can Use*. Master's thesis. Vienna University of Economics and Business.

[44] Annie W. Y. Ng and Alan H. S. Chan. 2017. Mental Models of Construction Workers for Safety-Sign Representation. *Journal of Construction Engineering Management* 143, 2 (2017).

[45] Jason R.C. Nurse. 2013. Effective Communication of Cyber Security Risks. In *7th International Scientific Conference on Security and Protection of Information (SPI 2013)*.

[46] Jason R.C. Nurse, Sadie Creese, Michael Goldsmith, and Koen Lamberts. 2011. Guidelines for Usable Cybersecurity: Past and Present. In *The 3rd International Workshop on Cyberspace Safety and Security (CSS 2011) at The 5th International Conference on Network and System Security (NSS 2011)*. IEEE.

[47] A Ant Ozok, Quyin Fan, and Anthony F Norcio. 2010. Design guidelines for effective recommender system interfaces based on a usability criteria conceptual model: results from a college student population. *Behaviour & Information Technology* 29, 1 (2010), 57–83.

[48] Massimiliano Pala and Yifei Wang. 2010. On the Usability of User Interfaces for Secure Website Authentication in Browsers. In *Proceedings of the 6th European Conference on Public Key Infrastructures, Services and Applications (EuroPKI'09)*. Springer-Verlag, Berlin, Heidelberg, 239–254.

[49] Kara Pernice. 2017. F-Shaped Pattern of Reading on the Web: Misunderstood, But Still Relevant (Even on Mobile). (Nov 2017). https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content/

[50] Catherine Pickering, Julien Grignon, Rochelle Steven, Daniela Guitart, and Jason Byrne. 2015. Publishing not perishing: How research students transition from novice to knowledgeable using systematic quantitative literature reviews. *Studies in Higher Education* 40, 10 (2015), 1756–1769.

[51] Irene Pollach. 2007. What's wrong with online privacy policies? *Commun. ACM* 50, 9 (2007), 103–108.

[52] Elissa M Redmiles, Everest Liu, and Michelle L Mazurek. 2017. You Want Me To Do What? A Design Study of Two-Factor Authentication Messages. In *Thirteenth Symposium on Usable Privacy and Security (SOUPS 2017)*. USENIX Association, Santa Clara, CA.

[53] K Renaud and J van Biljon. 2017. Demarcating Mobile Phone Interface Design Guidelines to Expedite Selection. *South African Computing Journal* 29, 3 (2017).

[54] Florian Schaub, Rebecca Balebako, Adam L Durity, and Lorrie Faith Cranor. 2015. A design space for effective privacy notices. In *Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)*. USENIX Association, 1–17.

[55] Ronak Shah and Kailas Patil. 2016. Evaluating Effectiveness of Mobile Browser Security Warnings. *ICTACT Journal on Communication Technology* 7, 3 (2016), 1373–1378.

[56] Umesh Shankar and Chris Karlof. 2006. Doppelganger: Better browser privacy without the bother. In *Proceedings of the 13th ACM Conference on Computer and Communications Security*. ACM, 154–167.

[57] Claude E Shannon. 2001. A mathematical theory of communication. *ACM SIGMOBILE Mobile Computing and Communications Review* 5, 1 (2001), 3–55.

[58] Mario Silic, Jordan Barlow, and Dustin Ormond. 2015. Warning! A comprehensive model of the effects of digital information security warning messages. In *Conference Proceedings The 2015 Dewald Roode Workshop on Information Systems Security Research, October, IFIP*. 1–32.

[59] Geoff Skinner, Song Han, and Elizabeth Chang. 2005. A framework of privacy shield in organizational information systems. In *Mobile Business, 2005. ICMB 2005. International Conference on*. IEEE, 647–650.

[60] Asimina Vasalou, Anne-Marie Oostveen, Chris Bowers, and Russell Beale. 2015. Understanding engagement with the privacy domain through design research. *Journal of the Association for Information Science and Technology* 66, 6 (2015), 1263–1273.

[61] Melanie Volkamer, Karen Renaud, Gamze Canova, Benjamin Reinheimer, and Kristoffer Braun. 2015. Design and Field Evaluation of PassSec: Raising and Sustaining Web Surfer Risk Awareness. In *Trust and Trustworthy Computing - 8th International Conference, TRUST 2015, Heraklion, Greece, August 24-26, 2015, Proceedings*. 104–122.

[62] Tilo Westermann. 2017. *User Acceptance of Mobile Notifications*. Ph.D. Dissertation. Institute of Software Engineering and Theoretical Computer Science, Berlin Institute of Technology Berlin, Germany.

[63] Michael Wogalter and Christopher Mayhorn. 2017. Warning Design. In *Information Design: Research and Practice*, Alison Black, Paul Luna, Ole Lund, and Sue Walker (Eds.). Chapter 20.

[64] Michael S Wogalter. 1999. Factors Influencing the Effectiveness of Warnings. *Visual Information for Everyday Use: Design and Research Perspectives* (1999), 93–110.

[65] Michael S Wogalter, David M DeJoy, and Kenneth R Laughery. 1999. Organizing theoretical framework: a consolidated communication-human information processing (C-HIP) model. *Warnings and Risk Communication* (1999), 15–23.

[66] Heng Xu, Robert E Crossler, and France BéLanger. 2012. A value sensitive design investigation of privacy enhancing tools in web browsers. *Decision Support Systems* 54, 1 (2012), 424–433.

The Development of Intervention E-Learning Materials and Implementation Techniques For Cyber-Security Behaviour Change

Tiffany Skinner
LiMETOOLS Ltd
Bournemouth, United Kingdom
tiff@limetools.biz

Prof. Jacqui Taylor
Bournemouth University
United Kingdom
jtaylor@bournemouth.ac.uk

John Dale
LiMETOOLS Ltd
Bournemouth, United Kingdom
john@limetools.biz

Dr John McAlaney
Bournemouth University
United Kingdom
jMcAlaney@bournemouth.ac.uk

ABSTRACT

Many organisations show compliance in running security awareness programmes, but this does not necessarily mean end users will change their behavior. This highlights one of the main challenges in cyber security. Providing awareness in a tool is a useful first step but it doesn't necessarily lead to changing behaviour [3]. In contrast, completing compliance or achieving competence can actually lead people to being more averse to change than before or even partaking in risky behaviour. This paper describes the collaboration between a specialist computer business (LiMETOOLS) and psychology academics to draw on psychology theory (e.g. Social Cognitive Theory, [4]) and pedagogy (e.g. self-directed learning) to create innovative techniques using interactive learning tools resulting in behaviour change. The aim of this article is to show how we have moved beyond developing materials that change awareness, to those that effectively change digital behaviour. We examine methodologies that can be integrated within online learning tools to embed text, video clips, gamification, and quizzes to encourage measurable cyber security behaviour change. A challenge within behaviour change is the maintenance of these behaviours and we are exploring the potential impact of using 'drip-feed learning' in the form of a short video magazine with embedded quizzes and 'nudges' of behaviour changes that have previously learnt, delivered over a long period of time in very short stimulus packages.

CCS CONCEPTS

• Interactive learning environments → Learning management systems ; • Systems security → Network security ; • Network Security → Human and societal aspects of security and privacy ; •Interaction paradigms → Interactive systems and tools;

KEYWORDS

e-learning, cyber security, behavior change, intermittent learning

1.BACKGROUND TO EXISTING LEARNING TOOLS

LiMETOOLS creates cyber security interactive learning tools by using a blend of text, games, videos, and quizzes. Tools are developed within this scope to encourage and enable behaviour change. The business has started to draw on psychology research within the cyber security operational field to inform the product development. This process utilises techniques used by broadcast digital storytellers who make complex dramas and documentaries that recognise the human factor element of cyber security. LiMETOOLS has developed an integrated storytelling, scoring and authoring tool platform, for the creation of the e-learning content. The e-learning tool is uploaded onto a Learning Management System (LMS) allowing for learner data capture that can be aggregated to report on specific individuals and teams, to assess knowledge retention, and provide valuable feedback to businesses about their employees, across departments, site locations and countries. The tools aim to provide behavior change in cyber security through perceived susceptibility which could increase with relevant and consistent messages about behavioural cyber security changes that need to take place that are personal to the individual [24]. This can lead to the necessary change in individuals' attitudes and intentions. The Elaboration Likelihood Model (Petty and Cacioppo 1986) [19] describes how attitudes are formed and persist and how this information can be used to persuade people to change be-haviour. It is based on the notion that there is a central route and a peripheral route to attitude change. The central route is a conscious process with thoughtful, logical decisions, where the peripheral route is automatic and unconscious. Therefore, decision making when using the central route can only be processed through motivation with intention to change whilst paying attention to information, leading to a more permanent attitude change. Although attitudes are often different to enacted behaviour, providing a message that is personally relevant to the individual should motivate them to take the central route and a change in attitude and intention (Bada and Sasse 2014) [3]. Blythe (2013) [5] supports this notion in the workplace, to improve employee's virus prevention behaviour. Specific and personalised messages were designed as an intervention tailored from anti-virus software and other security systems that enforce cyber secure behaviour. By providing each employee with the consequences of visiting particular websites and attachment downloads lead to an increase in more secure behaviours. LiMETOOLS aim to deliver their tools in a personally relevant way using knowledge awareness, documentaries, dramas and quizzes.

2. OVERVIEW OF INNOVATIVE METHODS AND TECHNIQUES TO ACHIEVE AND MEASURE BEHAVIOUR CHANGE IN CYBER SECURITY WITHIN THE WORKPLACE

There are three projects currently taking place within our collaborative research programme and in this article we will outline the work-in-progress for these and future plans.

2.1 Methods and Techniques for Motivating Learning and Improving Risk Perception and Risk-Related Behaviour

2.1.1 Knowledge based awareness via text.

Increasing an individual's knowledge in cyber security, increases their awareness of the risks involved online. This further increases individual motivation which could lead to a change in their risky online behaviour. Heuristic decision making underpins this theory. Heuristics are a set of simplifying rules for processing information selectively in memory [11]. Kahneman (2011) [15] suggests there is System 1 processing which is automatic, fast, and unconscious decision making, whereas System 2 processing is deliberate, slow, and effortful. Typical online behaviours are likely to involve System 1 processing, using our existing knowledge, where individuals may not have time to deliberate their online behavior. For example, Guadagno and Cialdini (2005) [13] found that online users evoke cognitive heuristics to evaluate the sources information credibility decreasing cognitive effort and time pressures through confirmation bias to disregard information that is not consistent with ones' own beliefs. By increasing the knowledge in cyber security within the tools, individuals may be more likely to consider cues to make better estimations, which could alert us to risky online behaviour. Davinson and Sillence (2010) [9] support this notion where participants were provided with phishing threat information, methods, and consequences. After reading and acknowledging this information, intention for users to behave more securely increased. However, a problem with many security awareness programmes based on solely knowledge, is that users are expected to identify the argument rationality from the information given, that cyber security is important and motivations to act accordingly will pursue that. The Cognitive Moral Development (CMD) theory (Kohlberg 1981) [16] demonstrates most rational people, in moral cases, would like an explanation for orders that they are given. Siponen (2000) [26] supports this theory where information security awareness achievement or failure correlates with behavioural readjustment in a positive way of acceptance, internalization and co-operation or alternatively, if the information is received without rationale, resistance or hate may be felt by some individuals. Further, internalizing security guidelines cannot be assumed to be achieved instantly, where if employees take a learning tool it cannot be presumed they will follow the guidelines at once. Therefore, it may be a long process to get staff to comply with guidelines. This means it is important to understand individual differences in learning where some individuals may learn better with the use of video content, or gamification to understand the content and achieve the intended learning objectives. Therefore, implementation of other delivery methods into e-learning programmes may increase cyber security behavior change [2].

2.1.2 Video Drama.

Delivery methods and techniques are crucial in changing cyber security behaviour [17]. Interactive training though video drama that identifies with a recognisable peer could increase motivation to change online behaviour. The Social Cognitive Theory [4] underpins this notion and proposes that people learn by watching what others do in the social context of experiences, outside media influences and social interactions. Video-based delivery methods can be adapted to a particular audience to consider a recognisable peer. Recognisable peers could be shown through similar gender, age, ethnicity, or the personfis situation. Pfleeger and Caputo (2012) [20] suggest that a recognisable peer could have the user gain a greater sense of self-efficacy and to further influence an imitation of their behaviour. The central route of heuristic decision making supports this notion whereby if the message is personally relevant and from a recognisable source, individuals will be motivated to make a more effortful decision and want to imitate the behaviour of their peer [3]. If people believe that they can act to solve a problem they become more inclined and committed to do so. This is linked to emotion, which is a fundamental part of rational decision making and individuals thought processes and linked to past experiences. The past experiences lead to emotional learning when people are confronted with a set of choices and this is what guides their decision by highlighting potential decisions and eliminating others (Goleman, 1995) [12]. Consequently, security measures should aim at provoking emotions, therefore appealing to them in order to affect attitudes and motivation in a positive manner. This means that by providing video drama content that relates specifically to the user and the correct cyber secure behaviours can motivate the user to change their own behaviour.

2.1.3 Gamification.

Game-based delivery methods are used to challenge, engage, and motivate individuals to offer effective learning compared to more traditional modes of awareness. Interactive serious games combine graphics, play and training concepts to enhance behavioural change. Anti-Phishing Phil [25] was one of the initial interactive games which taught users how to identify phishing URLs. The authors found that the participants who played the serious game were better able to identify fraudulent web sites. However, serious games can be suggested to be over simplified and experiences in the game do not will not reflect long-term habit change. This means, although Anti-Phishing Phil resulted in learning, there is a need to have a shared meaning between the individual and the environment of why they are playing the game [17]. In light of this, Boopathi, Sreejith and Bithin (2015) [6] provide a game which not only provides a capture the flag gamification of cyber security attacks but also puts the attacks into context by embedding tutorials in the first learning round which is then tested in the next stage of the game. Now that users have developed an understanding of the context of the attacks, the interactive capture the flag gamification of cyber security attacks enhances real-world scenario application by using their new knowledge to cyber-attack other teams to capture their flag. Educational gamification therefore shows it can increase understanding of cyber security threats in a more engaging manner, to better help implement behavior change.

2.1.4 Blended Learning.

To maximise full cyber security awareness, motivate, and ultimately change behaviour, implementing a blend of learning is important. Abawajy (2014) [1] evaluated the various channels of text-based, video-based and game-based security awareness delivery methods for phishing attacks. Within these delivery methods, the sessions consisted of informing participants of tactics and behaviours of exploitation attacks to encourage the learning to enable avoidance of phishing attacks. Furthermore, information of the over-arching aims of the attacks and the dangerous consequences involved if individuals submit to this type of attack is also given. The authors found the text-based training materials when read properly are helpful in identifying phishing websites but more importantly, game-based, and video-based delivery models are more suitable security delivery methods. Although this has been carried out solely on phishing scams, it is a promising implication for further research to examine the effects of delivery methods in alternate aspects of cyber security; for example, smart home working vulnerabilities. This could mean that creating eLearning tools that deliver a blend of learning will be better able to appeal to a wider range of individuals to enable change in security behaviour.

2.2 Measuring Capacity Growth in Organisations

Virtual learning environments and e-learning systems are fast becoming an important part of the organizational education and learning process (Pituch & Lee, 2006) [21]. Some e-learning systems only provide a measure of compliance in an awareness programme; however this type of analytics does not necessarily provide evidence that the programme has created the learners to have the capabilities to act in the desired manner. This means, cyber security awareness programmes ideally have to deliver measurable benefits to influence behaviour changes [3]. The use of a Learning Management System (LMS) enables businesses to do this where reporting can show departmental or site-specific vulnerabilities and not just the ability to show a pass and fail compliance. By collecting regular metrics or 'business analytics' it can measure the effectiveness of the leaners and then allows adjustment to the learning tool for the visualization and investigation of company employee data sets allowing strategic advantages and improvement of education and learning of a workforce (Ferguson, 2012) [11]. A robust LMS can capture data about a course results and answers and data about the user activity. This data can be interpreted to assess questions and responses that a lot of people are struggling with through reporting features and therefore can adapt the course accordingly to work on delivering the material in a more understandable way (Dawson, McWilliam, & Tan, 2008). Schläfke, Silvi and Möller (2012) [22] support this notion where performance management analytics has been found to increase performance. This could be explained as business competition increases, slight advantages of data analytics can make all the difference to help best support management decision-making and employee performance. LiMETOOLS incorporates the LMS, Litmos, to capture appropriate metrics for example reporting on specific individuals and teams to access knowledge retention of eLearning module. The analytics provide valuable feedback to businesses about their employees to allow for the determination of which cyber security areas are better understood and where vulnerabilities lie. By determining what vulnerabilities there are in a business as a whole, a team or individual, enables future e-learning programmes to be tailored to improve these specific behaviours.

2.3 LONG-TERM HABIT CHANGE AND RETENTION OF LEARNING

In this section we consider the timing of persuasive messages/when to interrupt the user and effective presentation of communications.

2.3.1 Proposal of a Habit Retention Product and Intermittent Learning.

We are looking to explore the most powerful way to increase eLearning retention in the workplace by using smaller, more incremental interventions. Ebbinghaus (1964) [10] proposes distributed practice or spacing effect increases memory compared to massed learning. This suggests, rather than providing information through an e-learning module on a one-time basis, delivering information using different formats could develop learning and increase habit retention. This may be through a video magazine with embedded learning information and quizzes, which could lead to the retention of a behaviour change. Schwarz and Clore (1983) [23] suggest messages are perceived more persuasive if it is consistent with userfis mental representations. This may mean that by taking the full eLearning tool, the users are more likely to have built up representations of the material and therefore presenting 'bite-sized' reminders of the same tool, will aim to persuade them to maintain cyber secure behaviour and potentially cause long term habit change. The challenge that arises is for Bournemouth University and LiMETOOLS to ensure that these behavioural cyber security changes are retained permanently, as it could be suggested the longer retention, the greater return on investment for the business.

2.3.2 Intermittent Learning.

Research in academic settings has shown that in order for learning to occur, self-paced learning needs to arise outside of the formal teaching activities such as assessment, seminars and lectures. Effective spacing requires students to understand the benefits and also to possess a certain degree of self-regulated learning. Although numerous studies demonstrate the benefits of spacing learning activities, many students seem unaware of this strategy. Instead, Taraban, Maki, & Rynearson (1999) [27] found that students crammed revision before exams, thinking that this is effective learning. Similarly, across three studies, Kornell (2009) [12] showed that even when students experienced the benefits of spaced learning they still retained the false belief that learning a large quantity of material at one time was more effective than spacing. Therefore, when designing behaviour change materials it is important to enhance awareness of the need for spacing. Although there is a growing number of studies investigating spacing in educational environments demonstrating convincing evidence that spacing is an effective approach for enhancing learning, there is less research showing how this can be applied in non-educational contexts and in online settings. One such study by Pereira, Taylor and Jones (2009) [18] found that spacing was an effective technique in improving retention and test performance for adults working in project management teams within industry and using an online training system in both company and social time.

2.3.3 Timing and Presentation of the Habit Change Tool.

It is essential to develop a positive habit change programme to better protect individuals and organisations to help prevent cyber security incidents [1]. The timing of the habit change tool will directly affect whether individuals are willing to take on further learning and continue with cyber secure behaviour change. The timing and presentation of persuasive technology of real-time reminders to maintain behaviour has been researched in the health sector. IJsselsteijn et al. (2006) [14] explored strategies that interrupt users to perform healthy behaviours and whether there is a long-term health change. Users were presented with interrupting commands at fixed intervals within their working day. The results found that when the command was polite, it positively correlated with compliance and predicted a long-term habit change. Moreover, when there was an annoying interruption, compliance dramatically decreased. This therefore expresses that timing of messages could interrupt the working day but providing the presentation of the command is articulated in a polite, respectful manner long-term habit change could occur. However, Cutrell et al., (2001) [7] expressed that any interruptions in the working day will affect performance negatively even if they are ignored. Alternatively, notification-based alerts to remind individuals that there is a new behaviour change command ready to be viewed may work better in changing behaviour. Czerwinski and Schumacher (1991) [8] found the pre-warning notification did not affect day to day performance because a user is put in control of when they view the full content of the notification. LiMETOOLS could implement a habit change tool whereby notifications are used to alert that a new intermittent learning tool is available, and individuals are more likely to participate and in turn could maintain cyber secure behaviours.

3. DISCUSSION AND CONCLUSION

We suggest in this article that the collaboration between technology companies and psychology researchers improves the quality of cybersecurity education and behavior change amongst end users. Also, we have highlighted that to enable behaviour change in cyber security practices in end users a blend of delivery methods is essential through knowledge-based awareness, video dramas and gamification methods. Applying analytics to a workforce's responses to eLearning can enable better understanding of where parts of the businesses are vulnerable in areas of cyber security. Tailored eLearning programmes can then be adapted to teams or individuals to help prevent cyber-attacks. To continue and maintain behaviour change, collaboration between academics and companies will conduct further research into the timing and delivery of intermittent learning strategies. By enabling better understanding of the underlying processes, better learning retention and long term cyber security habit changes can occur. Future research for this area would be using existing LiMETOOLS cyber learning tools and developing some smaller incremental learning product as oppose to massed learning. Comparative research can then be undertaken to see whether recall or habit change is better incrementally or in a massed fashion for both versions of the same tool. Moreover, future research could be the timing of the incremental tool within a corporate environment for when the best time to interrupt users in their working day is in order to watch the tool and how long the incremental learning tool should be.

REFERENCES

[1] J. Abawajy. 2014. User preference of cyber security awareness delivery methods. *Behaviour & Information Technology* 33, 3 (2014), 237–248.

[2] D. Ashenden and D. Lawrence. 2013. Can we sell security like soap?: a new approach to behaviour change.. In *Proceedings of the 2013 Workshop on New Security Paradigms*. ACM, 87–94.

[3] M. Bada and A. Sasse. 2014. Cyber Security Awareness Campaigns: Why do they fail to change behaviour? (2014).

[4] A. Bandura. 1986. *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice-Hall.

[5] J. Blythe. 2013. Cyber security in the workplace: Understanding and promoting behaviour change. In *Proceedings of CHI'2013 Doctoral Consortium, 1065.* 92–101.

[6] K. Boopathi, S. Sreejith, and A. Bithin. 2015. Learning cyber security through gamification. *Indian Journal of Science and Technology* 8, 7 (2015), 642–649.

[7] E. Cutrell, M. Czerwinski, and E. Horvitz. 2001. Notification, Disruption, and Memory: Effects of Messaging Interruptions on Memory and Performance. In *INTERACT'01.* 263–269.

[8] M. Czerwinski, S. Chrisman, and B. Schumacher. 1991. The effects of warnings and display similarities on interruption in multitasking environments. *SIGCHI Bulletin* 4 (1991), 38–39.

[9] N. Davinson and E. Sillence. 2010. It won't happen to me: Promoting secure behaviour among internet users. *Computers in Human Behavior* 26, 6 (2010), 1739–1747. DOI: http://dx.doi.org/10.1016/j.chb.2010.06.023

[10] H. Ebbinghaus. 1964. *Memory: A contribution to experimental psychology.* New York: Dover Publications. Ruger H.A. and Bussenius, C.E. and Hilgard, E.R., translators.

[11] R. Ferguson. 2012. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning* 4, 5–6 (2012), 304–317.

[12] D. Goleman. 1995. *Emotional Intelligence.* Bantam Books, New York, NY.

[13] R.E. Guadagno and R.B. Cialdini. 2005. Online persuasion and compliance: Social influence on the Internet and beyond. In *The social net: The social psychology of the Internet.* 91–113.

[14] W. IJsselsteijn, Y. de Kort, C. Midden, B. Eggen, and E. van den Hoven. 2006. Persuasive technology for human well-being: setting the scene. In *Persuasive technology.* 1–5.

[15] D. Kahneman. 2011. *Thinking, fast and slow.* London: Allen Lane.

[16] L. Kohlberg. 1981. *The Philosophy of Moral Development.* San Francisco, CA.

[17] G Light, R. Cox, and S. Calkins. 2010. *Learning and Teaching in Higher Education: The Reflective Professional* (2nd ed.). Sage, London.

[18] C. Pereira, J. Taylor, and M. Jones. 2009. Less learning, more often: the impact of the spacing effect in an adult e-learning environment. *Journal of Adult and Continuing Education* 15, 1 (2009).

[19] R.E. Petty and J.T. Cacioppo. 1986. The elaboration likelihood model of persuasion. *Advances in experimental social psychology* 19 (1986), 123–205.

[20] S.L. Pfleeger and D.D. Caputo. 2012. Leveraging behavioral science to mitigate cyber security risk. *Computers & security* 31, 4 (2012), 597–611.

[21] K. A. Pituch and Y.-K. Lee. 2006. The influence of system characteristics on e-learning use. *Computers & Education* 47 (2006), 222–244.

[22] M. Schläfke, R. Silvi, and K. Möller. 2012. A framework for business analytics in performance management. *International Journal of Productivity and Performance Management* 62, 1 (2012), 110–122.

[23] N. Schwarz and G.L Clore. 1983. Mood, misattribution, and judgments of well-being: Informative and directive functions of affective states. *Journal of Personality and Social Psychology* 45 (1983), 513–523.

[24] R.S Shaw. 2009. The impact of information richness on information security awareness training effectiveness. *Computers & Education* 52, 1 (2009), 92–100. DOI: http://dx.doi.org/doi:10.1016/j.compedu.2008.06.011

How can persuasion reduce user cyber security vulnerabilities?

John Paul Vargheese
University of Aberdeen
Aberdeen, United Kingdom
jpvargheese@acm.org

Matthew Collinson
University of Aberdeen
Aberdeen, United Kingdom
matthew.collinson@abdn.ac.uk

Judith Masthoff
University of Aberdeen
Aberdeen, United Kingdom
j.masthoff@abdn.ac.uk

ABSTRACT

End users' behaviour often leads to cyber security vulnerabilities. Recent studies investigating user error as a critical vulnerability within an organisation's security have emphasised the need for closer consideration of human factors when designing and implementing security solutions. This has led to an increasing focus on usable security design principles that seek to inform, enable and assist users to perform tasks securely and appropriately respond to potential threats. These objectives are often reinforced through provision of user security training and information awareness campaigns. However, despite these efforts, end users remain the target for increasingly sophisticated cyber attacks. In this paper, we discuss recent studies that stress the need for greater attention to human factors and motivate the potential for behavioural change interventions to support cyber security. We review behavioural insights into possible causes of user non-compliance with cyber security best practice, cognitive biases that are related to these insecure behaviours and cyber security risk communication challenges and recommendations. We discuss how persuasion incorporated within cyber security risk communications could reduce user cyber security vulnerabilities and outline a proposed methodology for designing persuasive strategies for this purpose.

CCS CONCEPTS

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

KEYWORDS

Human factors, Cyber Security, Persuasive technology, Behaviour change

1 INTRODUCTION

There is widely held consensus that end user behaviours are critical to an organisation's cyber security [36, 46]. This is due to the potential for end users' actions and decisions to result in cyber security vulnerabilities. Users have been described as the "weakest link" within an organisation's security chain and many cyber attacks, including the "professional", are designed to exploit this weakness [36, 37].

Organisations seek to address the threat of cyber attacks by designing policies for users to follow, and implementing various technical and non technical processes for maintaining security. However, there is increasing evidence that these provisions alone are not sufficient [19, 34]. Users' behaviour continues to be regularly reported as a primary cause of successful cyber security attacks [7] and this is often related to user non-compliance with policies [38, 40, 41].

Persuasive technologies and behaviour change interventions may offer a means to reduce user vulnerabilities by encouraging users to change their behaviour and act more securely [2, 4, 8, 9, 21, 24, 32]. Interventions have been shown effective across many domains, including health, sustainable behaviour, education, commerce, security, safety [17, 44, 45].

The adversarial context for intervention is noteworthy. Persuasion, in the form of deceit and manipulation, is a standard tool of attackers, and is sometimes personalised.

This paper discusses how persuasion could help to reduce user cyber security vulnerabilities. We discuss causes of insecure user behaviours, and related cognitive factors that offer possible explanations for insecure user actions. We discuss challenges in risk communication with regards to informing users about potential cyber security threats, with a view to how to incorporate persuasion to increase user risk perception, thus encouraging users to follow best practice.

2 CAUSES OF INSECURE USER BEHAVIOURS

Insecure user behaviours are often, though not always, linked to non-compliance with best practice as prescribed through an organisation's cyber security policy [4]. By examining the causes of this behaviour, it is possible to identify relevant behavioural determinants that can be used to develop persuasive strategies for encouraging users to act more securely, including complying with cyber security policy and best practice [25, 26]. Blythe, Coventry et al. [4, 8] summarize commonly reported causes of these behaviours, which we now discuss briefly, noting that this may be incomplete.

Determinants of insecure behaviours often overlap, particularly those associated with habits and expectations, desirability, convenience and incentives. For example, users often

AISB 2018, April 2018, Liverpool, UK

demand constant connectivity that may form both a habit and expectation and result in users selecting insecure connections or performing tasks involving sensitive data within an insecure environment such as a public area. The convenience, desirability and incentives that motivate perceived benefits for acquiring access to a resource or service often surpass concerns over security. From the user perspective, the perceived benefits gained from performing these actions are easily realised, whereas perceived benefits for performing recommended security behaviours are less so. Habits and expectations that drive insecure user behaviours may be reinforced if users suffer no immediate ill consequences. This can have a negative impact on users' motivation to follow security best practice and comply with policy. Security considerations are typically not the primary focus of users, and the required effort and additional cognitive load required to account for the potential consequences of users' actions may also negatively impact users' motivation to follow best practice [1, 32, 35, 39]. In addition, users may perceive behaviours aligned with security policy and best practice as unnecessary obstacles and interruptions [1, 32].

Insecure behaviours are often attributed to users' lack of knowledge and skills. Studies have emphasised the importance of information security training and awareness campaigns to improve cyber security [5, 12, 16, 18, 33]. However, the knowledge and skills required to perform tasks securely changes frequently as the diversity and volume of cyber attacks increases, requiring additional user effort.

Users may not consider cyber security as their responsibility but that of IT departments, others, or the system itself [1, 32]. They may be unable or unwilling to dedicate the required cognitive resources to consider all possible outcomes of their actions and decisions. This, combined with the need for maintaining up-to-date security knowledge and skills, can lead to uncertainty and ambiguity amongst users, who consequently may perceive additional security related actions as obstructive and having a potential negative impact on their productivity in contrast with less secure behaviours [1, 32]. This may also negatively impact their motivation.

Additional causes of insecure behaviours include failure to recall procedures related to best practice, conflicts on trusted sources of information about this and error prone task and mental models. Users may continue to perform tasks without understanding the relationship between their actions and potential threats, risk, susceptibility and vulnerability to cyber attacks. There is typically little to no positive feedback provided for following best practice for cyber security in a manner that encourages and reinforces these behaviours. Users may also over-estimate their ability to identify and respond to security risks or based on previous experience without negative outcomes, have a low risk perception and be unaware of what actions increase their vulnerability.

The diverse and complex range of causes that may lead to insecure behaviour highlights the challenges involved with addressing human aspects of cyber security. Behavioural determinants associated with non-compliance with policy and best practice often overlap and point towards additional cognitive factors that may cause such behaviours. The next section discusses cognitive biases and how these offer further explanation towards possible causes of insecure behaviours.

3 COGNITIVE BIASES AND INSECURE USER BEHAVIOURS

The term *cognitive biases* refers to individuals' irrational judgement and decision making that is commonly a result of subjective, and often incorrect or inaccurate, beliefs and perceptions. Together with behavioural determinants of insecure user behaviours, cognitive biases offer insights into why users may perform such behaviours. We discuss security-relevant factors related to cognitive bias, without claiming this list to be exhaustive or mutually exclusive.

Firstly, *anchoring and adjustment* refers to decision making based on available information at the time which is adjusted as required to assist with finalising the decision or judgement [14, 43]. Users may inadvertently compromise security by making decisions regarding their behaviours based on incomplete, inaccurate or out-of-date information [1].

Secondly, the *availability heuristic* describes how individuals assess the likelihood or frequency of an event occuring, based on the ease at which information about the event is cognitively available [14, 43]. The information available for considering the likelihood of a positive outcome from a decision or action (such as gaining access to a resource or service) may create bias in the users' judgement that may lead to insecure behaviours and cyber security vulnerabilities.

Thirdly, *framing* refers to how decision making can be affected by presenting decisions in either a positive or negative manner [22]. This can be used to exploit users. and encourage insecure behaviour. For example, users may be requested to grant permission to software related to a task to be performed. Framing that emphasises gains from granting permission, rather than associated risks, may encourage users to grant permission. Users may also be manipulated into performing insecure behaviours by exploiting *loss aversion*, whereby losses are valued more highly than gains.

Fourthly, both framing and loss aversion have been linked to the *endowment effect* [32]. This describes how individuals place a higher value on an object or item when in their possession compared to when not [14]. For example, if a user has a choice of continuing to use a service with a newly known risk that should be unacceptable, or using an alternative, the user may choose to continue using the original.

Fifthly, *(hyperbolic) discounting* explains how individuals will prefer an immediate gain compared to a higher gain later [1]. For example, users may install malicious software to increase download speeds in contrast to using existing secure tools already available.

Sixthly, users may overestimate their capability to identify and respond to potential security threats [4, 8]. *Overconfidence* and *optimum bias* lead individuals to incorrectly estimate the potential for negative outcomes to occur. They may account for users' low risk perception, lack of awareness of their susceptibility and vulnerability to potential threats, and reduce motiviation for following best practice [32]. *Control bias*, where individuals falsely believe they are fully capable of controlling and determining the outcome of an event or decision despite evidence to the contrary [23], may lead to similar outcomes. For example, users may fail to take preventative measures such as conducting a virus scan [32].

Seventhly, further causes of insecure user behaviours may be explained by *post completion errors*. This may occur when users are engaged with tasks that involve a sequence of actions some of which may by omitted despite achieving the main task objective [1], e.g. not logging off a shared PC [11].

Finally, *status quo bias* refers to how individuals' decision making often involves preferences towards a default actions [1]. This may reinforce user habits associated with insecure behaviours that may lead to cyber security vulnerabilities.

One possible approach to addressing the negative impact of cognitive biases that may lead to insecure user behaviours involves providing users with relevant and pertinent information concerning their actions. Acquisti et al. argue that providing effective information that increases users' risk perception and awareness of cyber security threats may also provide a means to encourage users to act more securely [1].

4 CYBER SECURITY RISK COMMUNICATION

Many studies have emphasised the role of *cyber security risk communications (CSRCs)* for addressing vulnerabilities caused by user behaviours. [3, 10, 15, 29]. CSRCs may help reduce vulnerabilities by increasing user risk perception and lowering susceptibility to potential threats. CSRCs may also provide a means for applying persuasive strategies and behavioural change techniques for encouraging users to act more securely, while increasing awareness and risk perception to vulnerabilities caused by insecure actions [1, 27].

Designing an effective CSRC involves considering what information is presented and how, in addition to who are the receiver and source [3]. Recommendations for increasing the effectiveness of CSRCs include reducing cognitive load required by users to understand and acknowledge potential risks [20, 30] and ensuring information presented is consistent, clarified and provides a means for users to visualise potential outcomes of risks involved [3].

The original idea behind CSRC (and the underlying ELM and HSM models [31]) is that behaviour depends on intentions, which in turn depend on attitudes. So, informing people about risks can change the way they think about their behaviour, and thereby lead to behaviour change. However, much research suggests a large attitude-behaviour gap, and also, that the effectiveness of interventions targetting attitudes is not always high. Therefore, we want to amend CSRC by incorporating persuasive strategies into communications that target behaviour directly, in line with work in persuasive technology and by psychologists such as Cialdini [6].

To develop persuasive strategies that may be incorporated within CSRCs for encouraging behaviour change, we propose adopting techniques from design guidelines and recommendations for developing persuasive interactive systems and behavioural change interventions. Our proposed methodological approach towards developing persuasive strategies for CSRCs is as follows:

(1) Identify which events should trigger appropriate CSRCs
(2) Identify relevant behavioural determinants related to the event that may lead to insecure behaviours
(3) Specify what risks are associated with these behaviours following recommendations from [3]
(4) Define goals of persuasive strategies based on behaviours associated with best practice for cyber security that are suitable responses to the event
(5) Define persuasive strategies based on the relationship between behavioural determinants of the event and goals of persuasion, as recommended by [25, 26, 28]
(6) Refine persuasive strategies by type, indicating how behavioural determinants of the event, related to the goals of persuasion, interact with either or both reflective and impulsive cognitive systems of the user[1]
(7) Refine persuasive strategies using generic persuasive design principles techniques, described by, e.g., [6, 13]
(8) Evaluate effectiveness of persuasive strategies.

5 DISCUSSION

We propose integrating persuasive strategies within CSRCs as a potential solution for reducing user security vulnerabilities. We present a proposed methodology for developing persuasive strategies for this purpose drawing from guidelines for designing persuasive interactive systems and behavioural change interventions. In future work, we plan to conduct user evaluation studies of CSRCs incorporating persuasive strategies developed using our proposed methodology.

REFERENCES

[1] A. Acquisti, I. Adjerid, R. Balebako, L. Brandimarte, L.F. Cranor, S. Komanduri, P.G. Leon, N. Sadeh, F. Schaub, M. Sleeper, et al. 2017.

[1] For more details, please refer to the Reflective Impulsive Model [42].

Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Comput. Surv.* 50, 3, Article 44 (2017).

[2] J. Blythe. 2013. Cyber security in the workplace: Understanding and promoting behaviour change. *CHItaly Doctoral Consortium* (2013).

[3] J. Blythe, J. Camp, and V. Garg. 2011. Targeted risk communication for computer security. In *Proceedings of IUI*. ACM, 295–298.

[4] J. Blythe, L. Coventry, and L. Little. 2015. Unpacking security policy compliance: The motivators and barriers of employees' security behaviors. In *Symposium on Usable Privacy and Security 2015*. 103–122.

[5] H. Cavusoglu, H. Cavusoglu, J.-Y. Son, and I. Benbasat. 2009. Information security control resources in organizations: A multidimensional view and their key drivers. *UBC Working Paper* (2009).

[6] R. Cialdini. 2016. *Pre-Suasion: A Revolutionary Way to Influence and Persuade*. Simon and Schuster.

[7] J.W. Coffey. 2017. Ameliorating Sources of Human Error in CyberSecurity: Technological and Human-Centered Approaches. In *8th Internat. Multi-Conference on Complexity, Informatics and Cybernetics*. 85–88.

[8] L. Coventry, P. Briggs, J. Blythe, and M. Tran. 2014. Using behavioural insights to improve the public's use of cyber security best practices. (2014). UK Government Office for Science, Ref: GS/14/835.

[9] L. Coventry, P. Briggs, D. Jeske, and A. van Moorsel. 2014. SCENE: A structured means for creating and evaluating behavioral nudges in a cyber security environment. In *International Conference of Design, User Experience, and Usability*. Springer, 229–239.

[10] S. Creese and K. Lamberts. 2009. Can cognitive science help us make online risk more tangible? *IEEE Intelligent Systems* 24, 6 (2009).

[11] P. Curzon and A. Blandford. 2004. Formally justifying user-centred design rules: a case study on post-completion errors. In *IFM*, Vol. 2999. Springer, 461–480.

[12] J. D'Arcy, A. Hovav, and D. Galletta. 2009. User awareness of security countermeasures and its impact on information systems misuse: a deterrence approach. *Inf. Systems Research* 20, 1 (2009), 79–98.

[13] P. Dolan, M. Hallsworth, D. Halpern, D. King, R. Metcalfe, and I. Vlaev. 2012. Influencing behaviour: The mindspace way. *Journal of Economic Psychology* 33, 1 (2012), 264–277.

[14] J. Ehrlinger, W.O. Readinger, and B. Kim. 2016. Decision-Making and Cognitive Biases. In *Encyclopedia of Mental Health (Second Edition)*, H.S. Friedman (Ed.). Academic Press, 5–12.

[15] F. Farahmand, M. Dark, S. Liles, and B. Sorge. 2009. Risk perceptions of information security: A measurement study. In *Computational Science and Engineering, 2009*, Vol. 3. IEEE, 462–469.

[16] S.M. Furnell, M Gennatou, and P.S. Dowland. 2002. A prototype tool for information security awareness and training. *Logistics Information Management* 15, 5/6 (2002), 352–357.

[17] J. Hamari, J. Koivisto, and T. Pakkanen. 2014. Do Persuasive Technologies Persuade?-A Review of Empirical Studies. In *International Conference on Persuasive Technology*. Springer, 118–136.

[18] M. Hentea. 2005. A Perspective on Achieving Information Security Awareness. *Issues in Informing Science & Information Tech.* 2 (2005).

[19] T. Herath and H.R. Rao. 2009. Protection motivation and deterrence: a framework for security policy compliance in organisations. *European Journal of Information Systems* 18, 2 (2009), 106–125.

[20] J.H. Hibbard and E. Peters. 2003. Supporting informed consumer health care decisions: data presentation approaches that facilitate the use of information in choice. *Annu. Rev. Public Health* 24, 1 (2003), 413–433.

[21] Q. Hu, T. Dinev, P. Hart, and D. Cooke. 2012. Managing employee compliance with information security policies: The critical role of top management and organizational culture. *Dec. Sci* 43, 4 (2012), 615–660.

[22] D. Kahneman and A. Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the Fundamentals of Financial Decision Making: Part I*. World Scientific, 99–127.

[23] E.J. Langer. 1975. The illusion of control. *Journal of Personality and Social Psychology* 32, 2 (1975), 311.

[24] N. Malkin, A. Mathur, M. Harbach, and S. Egelman. 2017. Personalized security messaging: Nudges for compliance with browser warnings. In *2nd European Workshop on Usable Security. Internet Society.*

[25] S. Michie, L. Atkins, and R. West. 2014. *The Behaviour Change Wheel: A Guide to Designing Interventions*. Silverback Publishing.

[26] S. Michie, M. Johnston, J. Francis, W. Hardeman, and M. Eccles. 2008. From theory to intervention: mapping theoretically derived behavioural determinants to behaviour change techniques. *Applied Psychology* 57, 4 (2008), 660–680.

[27] J.R.C. Nurse, S. Creese, M. Goldsmith, and K. Lamberts. 2011. Trustworthy and effective communication of cybersecurity risks: A review. In *Socio-Technical Aspects in Security and Trust (STAST)*. IEEE, 60–68.

[28] H. Oinas-Kukkonen and M. Harjumaa. 2009. Persuasive systems design: Key issues, process model, and system features. *Communications of the Association for Information Systems* 24, 1 (2009), 28.

[29] M.R. Pattinson and G. Anderson. 2007. How well are information risks being communicated to your computer end-users? *Information Management & Computer Security* 15, 5 (2007), 362–371.

[30] E. Peters, N. Dieckmann, A. Dixon, J.H. Hibbard, and C.K. Mertz. 2007. Less is more in presenting quality information to consumers. *Medical Care Research and Review* 64, 2 (2007), 169–190.

[31] R.E. Petty. 2013. Two Routes to Persuasion: State of the Art. *Int. Perspectives On Psych. Science, II: The State of the Art* (2013), 229.

[32] S. Lawrence Pfleeger and D.D. Caputo. 2012. Leveraging behavioral science to mitigate cyber security risk. *Comp. & Security* 31, 4 (2012).

[33] P. Puhakainen. 2006. *Design theory for information security awareness*. Ph.D. Dissertation. University of Oulu.

[34] P. Puhakainen and M. Siponen. 2010. Improving employees' compliance through information systems security training: an action research study. *MIS Quarterly* (2010), 757–778.

[35] M.A. Sasse, S. Brostoff, and D. Weirich. 2001. Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. *BT Technology Journal* 19, 3 (2001), 122–131.

[36] B. Schneier. 2000. *Secrets & Lies: Digital Security in a Networked World* (1st ed.). John Wiley & Sons, Inc., New York, NY, USA.

[37] B. Schneier. 2000. Semantic attacks: The third wave of network attacks. *Crypto-Gram Newsletter* 14 (2000).

[38] M. Siponen and R. Willison. 2009. Information security management standards: Problems and solutions. *Inf. & Manag.* 46, 5 (2009), 267–270.

[39] W. Smith, B. Hill, J. Long, and A. Whitefield. 1997. A design-oriented framework for modelling the planning and control of multiple task work in secretarial office administration. *Behaviour & Information Technology* 16, 3 (1997), 161–183.

[40] J.-Y. Son. 2011. Out of fear or desire? Toward a better understanding of employees' motivation to follow IS security policies. *Information & Management* 48, 7 (2011), 296–302.

[41] J.M. Stanton, K.R. Stam, P. Mastrangelo, and J. Jolton. 2005. Analysis of end user security behaviors. *Comp. & Security* 24, 2 (2005), 124–133.

[42] F. Strack and R. Deutsch. 2004. Reflective and impulsive determinants of social behavior. *Pers. Soc Psychol. Rev.* 8, 3 (2004), 220–247.

[43] A. Tversky and D. Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science* 185, 4157 (1974), 1124–1131.

[44] A. Xu, T. Chomutare, and S. Iyengar. 2014. Persuasive attributes of medication adherence interventions for older adults: A systematic review. *Technology and Health Care* 22, 2 (2014), 189–198.

[45] A. Xu, T. Chomutare, and S. Iyengar. 2014. Systematic review of behavioral obesity interventions and their persuasive qualities. In *International Conference on Persuasive Technology*. Springer, 291–301.

[46] M.E. Zurko and R.T. Simon. 1996. User-centered security. In *Proceedings of the 1996 Workshop on New Security Paradigms*. ACM, 27–33.