Eyes on URLs: Relating Visual Behavior to Safety Decisions

Niveta Ramkumar
nr1063@wildcats.unh.edu
Univ. of New Hampshire
Durham, NH, USA

Vijay Kothari
vijayk@cs.dartmouth.edu
Dartmouth College
Hanover, NH, USA

Caitlin Mills
caitlin.mills@unh.edu
Univ. of New Hampshire
Durham, NH, USA

Ross Koppel
rkoppel@sas.upenn.edu
Univ. of Pennsylvania
Philadelphia, PA, USA

Jim Blythe
blythe@isi.edu
Information Sciences Inst.
Marina del Rey, CA, USA

Sean Smith
sws@cs.dartmouth.edu
Dartmouth College
Hanover, NH, USA

Andrew L. Kun
andrew.kun@unh.edu
Univ. of New Hampshire
Durham, NH, USA

ABSTRACT

Individual and organizational computer security rests on how people interpret and use the security information they are presented. One challenge is determining whether a given URL is safe or not. This paper explores the visual behaviors that users employ to gauge URL safety. We conducted a user study on 20 participants wherein participants classified URLs as safe or unsafe while wearing an eye tracker that recorded eye gaze (where they look) and pupil dilation (a proxy for cognitive effort). Among other things, our findings suggest that: users have a cap on the amount of cognitive resources they are willing to expend on vetting a URL; they tend to believe that the presence of *www* in the domain name indicates that the URL is safe; and they do not carefully parse the URL beyond what they *perceive* as the domain name.

CCS CONCEPTS

• **Security and privacy** → **Social engineering attacks**; **Social aspects of security and privacy**; **Usability in security and privacy**; • **Human-centered computing** → *User studies*; • **Social and professional topics** → Spoofing attacks.

KEYWORDS

usable security, phishing, user study, eye tracking, cognitive psychology, reading

ACM Reference Format:
Niveta Ramkumar, Vijay Kothari, Caitlin Mills, Ross Koppel, Jim Blythe, Sean Smith, and Andrew L. Kun. 2020. Eyes on URLs: Relating Visual Behavior to Safety Decisions. In *Symposium on Eye Tracking Research and Applications (ETRA '20 Full Papers), June 2–5, 2020, Stuttgart, Germany.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3379155.3391328

1 INTRODUCTION

As people surf the web, check their email, and do other computer-related tasks, they interact with web addresses or Uniform Resource Locators (URLs) [Wikipedia contributors 2019c]. Unfortunately, URLs do not only serve legitimate content; bad actors may use URLs under their control to conduct attacks, e.g., to serve malware or steal credentials by masquerading as a legitimate service. Thus, users must be vigilant. Trusting an unsafe URL could present a security threat to the individual or their organization. Yet users don't want to ignore safe URLs either. This problem is compounded by user misperceptions of URL syntax, the sheer time required to vet URLs, and some practices of legitimate services (e.g., use of URL redirectors). These factors make it very difficult for users to vet URLs. Consequently, many attacks rely on the victim unwittingly clicking on a malicious URL.

From a security standpoint, it is critical to safeguard users from malicious websites. And so, numerous solutions have been developed. Some companies specialize in security training for users (e.g., [KnowBe4 2019; Proofpoint 2019b]). Others focus on limiting user exposure to unsafe URLs: Products and services like Microsoft Office 365 APT Safelinks [Microsoft 2019] and Proofpoint URLDefense [Proofpoint 2019a] check for malicious content served by URLs before allowing users to visit them. Some browsers similarly warn the user when they detect unsafe URLs (e.g., [Mozilla 2019]). There is also abundant research on why users fall for URL-based phishing attacks (e.g., [Dhamija et al. 2006; Hong et al. 2013]), on training techniques (e.g., [Kumaraguru et al. 2009; Miyamoto et al. 2014; Wen et al. 2019]), and on defenses (e.g., [Fette et al. 2007; Maurer et al. 2011]), as well as other foci. However, to the best of our knowledge, this is the first study that solely focuses on understanding users' visual attention as they process URLs. Studying users' visual attention while processing URLs allows us to determine why certain attacks succeed, to measure the influence of URL characteristics on visual processing and cognition, and to determine the efficacy of countermeasures.

The work presented here serves as a first step toward developing a descriptive model of the relationship between URL characteristics and user visual behavior. We conducted a user study where users were asked to classify URLs as safe or unsafe while wearing an eye tracker. One key finding is that participants spent more time on processing URLs as URL length increased but only up to a point. Another is that participants relied more upon the `authority` component of URLs than any other component.

ETRA '20 Full Papers, June 2–5, 2020, Stuttgart, Germany

ACM ISBN 978-1-4503-7133-9/20/06…$15.00
https://doi.org/10.1145/3379155.3391328

2 RELATED WORK

2.1 Eye Tracking and Reading

Eye tracking is considered to be a window into users' cognitive states [König et al. 2016; Reichle et al. 2012]. It has been employed to assess cognitive load [Palinko et al. 2010; Pappusetty et al. 2017; Pomplun and Sunkara 2003; Zagermann et al. 2016], reading strategies [Beymer et al. 2008; Hyönä et al. 2002, 2003; Rayner et al. 2006], and design implications [Bergstrom and Schall 2014; Goldberg et al. 2002]. We study users' eyes as they process URLs.

Users assess the safety of a URL by reading. The amount of visual attention given while reading reflects moment-to-moment cognitive processing [Rayner 1998; Zagermann et al. 2016]. Researchers have sought to examine the relationships between reading and eye movements by using measures like fixations, saccades, regressions, and backtracks [Beymer and Russell 2005; Sibert et al. 2000]. Fixations are pauses in eye movements during which new information is acquired. Research has shown that users fixate longer while reading when "the processing load is greater" [Just and Carpenter 1980].

Reading and scanning text differs with respect to fixations and word skipping [Rayner and Fischer 1996]. When and where someone looks next while reading is influenced by the reader's ongoing mental processing [Rayner and Fischer 1996]. Six commonly used eye-tracking measures are: fixation count, fixation count on various areas of interest (AOIs), proportion of time spent on each AOI, average fixation duration, fixation rate (fixation count/second), and gaze duration mean on each AOI [Lai et al. 2013]. We used all these measures, as well as pupil dilation and backtrack fixation count.

2.2 Pupil Dilation and Cognitive Load

As users read and evaluate URLs, they use cognitive resources. A common measure of cognitive load is pupil dilation [Kun et al. 2013; Palinko et al. 2010; Poole and Ball 2006]. When users face challenging tasks, their pupils dilate on the order of 0.1 to 0.5 mm [Beatty 1982; Pfleging et al. 2016]. This task-evoked pupillary response (TEPR) indicates the cognitive load of the task. However, pupil dilation is also influenced by other factors like the amount of light entering the pupil (pupillary light reflex) [Palinko and Kun 2012; Pfleging et al. 2016] and one's emotional state [Bradley et al. 2008; Stanners et al. 1979; Xu et al. 2011]. To reduce these effects, we conducted the experiment in a windowless light-controlled room.

2.3 Neutral Mood Induction

Mood can affect a person's ability to comprehend text and their judgment [Bohn-Gettler and Rapp 2011; Forgas 1989]. Mood induction is used to understand and reduce the effect of mood [Mills et al. 2019]. Watching a film or a story is one of the most effective mood induction techniques [Westermann et al. 1996]. To reduce the effect of mood and improve replicability, we had participants watch a video chosen to induce a neutral mood.

2.4 URL Security and Phishing

Phishing is the act of masquerading as a legitimate entity to gather sensitive user information [Wikipedia contributors 2019b]. Adversaries often use URL obfuscation to carry out phishing attacks. In fact, URL security is primarily studied in relation to phishing.

Researchers have studied the efficacy of different phishing techniques and demographic factors affecting phishing susceptibility [Dhamija et al. 2006; Downs et al. 2007; Hong et al. 2013; Sheng et al. 2010; Wu et al. 2006]; the impact of psychological manipulation on phishing susceptibility [Goel et al. 2017]; and the effect of communication medium on phishing susceptibility [Benenson et al. 2017; Benenson et al. 2014]. Phishing and URL obfuscation techniques have been categorized, e.g., [Althobaiti et al. 2019; Drake et al. 2004; Ollmann 2004]. However, there are also (ostensibly) legitimate reasons to obfuscate or otherwise break user expectations of where URLs go, e.g., URL redirection [Wikipedia contributors 2019d], tracking links [Cyphers et al. 2018]. Researchers have developed and compared phishing training approaches and educational materials [Arachchilage et al. 2016; Kumaraguru et al. 2007; Sheng et al. 2010, 2007; Stockhardt et al. 2016; Wen et al. 2019]. Companies even provide security training [KnowBe4 2019; PhishingBox 2019; PhishLabs 2019; Proofpoint 2019b; SANS 2019a,b].

Many defenses have also been pursued. Researchers have: compared browser indicators and warnings [Dhamija et al. 2006; Egelman et al. 2008]; developed ways to effectively convey security information [Maurer et al. 2011; Schechter et al. 2007]; and studied ML approaches for email filtering and URL classification [Almomani et al. 2013; Bergholz et al. 2010; Blum et al. 2010; Fette et al. 2007]. Browsers [Mozilla 2019] and search engines [Whittaker et al. 2010] use blacklists and other techniques to protect users. Some products vet URLs in emails before allowing user access, e.g., [Microsoft 2019; Proofpoint 2019a]. However, these defenses are not always foolproof, e.g., [Nathaniel 2017].

Recently, there has been growing interest in using eye trackers for usable security. An eye-tracking based system was developed to train users to look at the status bar [Miyamoto et al. 2014]. Another study involved participants classifying websites, not just URLs, while wearing an eye tracker to examine how users gauge website legitimacy and evaluate security indicators [Alsharnouby et al. 2015]. Our study is similar in spirit. However, we exclusively focus on how users visually process URLs. This narrow focus lets us dissect URLs into smaller components and examine how people process them. We seek to understand which parts people pay attention to, when people give up, and how their eyes process different URLs, amongst other things.

2.5 A Brief Introduction to URL Structure

A uniform resource locator (URL) is a string of characters that specifies the location of a web resource and how to access it [Wikipedia contributors 2019c]. The original URL specification details URL structure [Berners-Lee et al. 1994]. Here, we present the bare essentials of URL structure at an appropriate level of granularity to understand our work.[1]

Each URL in our corpus has the form:

```
<scheme>://<authority><rest>
```

The `scheme` component [Berners-Lee et al. 1998, 1994; WHATWG 2019] corresponds to the scheme name, which specifies how to interpret the text following the colon. Common schemes are *http*, *ftp*, and *file*. Every URL in our corpus uses the *https* scheme.

[1] A more thorough treatment of URLs can be found in URL and URI specifications and standards [Berners-Lee et al. 1998, 1994; WHATWG 2019].

Table 1: Disaggregation of a URL into its three components.

scheme	*delims.*	authority	rest
https	://	www.google.com	/forms/about/

The authority component specifies a subset of the host, port, username, and password [Berners-Lee et al. 1998; WHATWG 2019]. For URLs in our corpus,, the authority component has either the form host or user@host where host represents the host and user represents the username. In this study, the host is always a fully qualified domain name (e.g., *www.wikipedia.org*) - "a sequence of domain labels separated by '.' " [Berners-Lee et al. 1994]. The last domain label is the top-level domain. For URLs in our corpus, the authority component comprises everything following the leading *https://* until either the next /, if present, or the end of the line.

We call the last component rest, a catch-all term that is *not* borrowed from any specification or standard. It captures everything following the authority component. The rest component includes the path [Berners-Lee et al. 1998, 1994; WHATWG 2019], which may be empty; it may also include queries, fragments, and accompanying delimiters [Berners-Lee et al. 1998, 1994; WHATWG 2019]. For every URL in our corpus, if the rest component is non-empty, it includes a path that "[identifies] the resource within the scope of [the] scheme and authority" [Berners-Lee et al. 1998], it begins at the first / character following the authority component, and it is the last part of the URL. Table 1 provides an example of a URL disaggregation into these three components. Please note the formatting style used for these components. Later, we define areas of interest of the same names but different formatting styles.

3 STUDY OUTLINE

Our long-term goal is to understand users' visual behaviors (and the underlying cognitive processes they manifest) as they process, interpret, and operationalize security information (including information embedded in URLs) when making security decisions. Identifying which factors affect visual behavior and how they affect it is vital in informing security solutions. Such information can be used to improve security awareness training or to better design user interfaces that aid in decision-making.

The work presented in this paper is one step towards this long-term goal. We aim to capture how some URL properties affect visual behaviors. We attempt to control for other factors, but we do not explore them in this initial study. We propose hypotheses pertaining to how various aspects of a URL affect visual processing of the URL, test these hypotheses, and observe trends in users' visual behaviors.

3.1 Hypotheses

We created hypotheses to examine how users visually process URLs and how URL features affect this processing:

H1: Total time spent on processing a URL is longer for complex URLs than it is for simple URLs.

H2: Total time spent on processing a URL, normalized by the URL length, is shorter for complex URLs than it is for simple URLs.

H3: There exists a URL length threshold over which increasing URL length does not result in more time being spent on processing URLs.

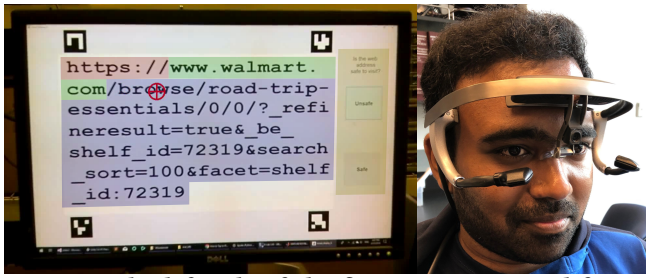

Figure 1: The left side of the figure is a processed frame from the eye tracker video (This is not the same as what the participant sees). The red cursor indicates gaze position and the four colored boxes represent four AOIs: the *scheme AOI* (red), the *authority AOI* (green), the *rest AOI* (blue), and the *response AOI* (yellow). The right side is an image of a participant performing the task wearing the eye tracker.

H4: Total time spent on the scheme per character is less than that of the authority and rest components.

H5: For URLs that have an authority component of form user@host where user ends with ".com", participants spend significantly more time per character looking at the user component than the host component.

4 METHOD

4.1 URL Corpus and Classification

We created a URL corpus comprising 64 URLs partitioned into 8 categories.[2] Categories are defined by features corresponding to (1) safety, (2) complexity, (3) a leading *www* in the authority component, and (4) the attack type for unsafe URLs. The corpus contains 8 URLs for each of the 8 categories. To reduce variability and maintain uniformity between categories, every URL uses *https* as the scheme component and *com* as the top-level domain.

The categories are defined by the following 4 features:

4.1.1 Safety: URLs that are *safe* use domain names associated with popular services within the USA, such as Facebook. We selected the fully qualified domain names used in these URLs primarily from the top 1,000 US websites in the Quantcast Top One Million list[3], although we consulted other lists as well. For the subset that were complex and included rest components, we chose the rest components by searching for legitimate content served by these domain names.

URLs that are *unsafe* have fully qualified domain names that, at the time of corpus construction, were eligible for purchase, did not have a domain name server record, or were spoofed websites. While many URLs with the *unsafe* feature were not actually unsafe to visit, it is exceedingly unlikely that participants would be knowledgeable about the status of the URLs tagged as *unsafe*, and, if an adversary wished to acquire the corresponding domains, they could do so. This decision allowed for greater control over the corpus.

[2]Materials used in this study can be found at https://drive.google.com/drive/folders/1ZNMLoXBxOU4R2nela-6d7MxsaQGrdyg4

[3]https://www.quantcast.com/top-sites

Table 2: Mean values and standard deviations of measurements for the eight URL categories (not normalized by length).

Category	Safety	Complexity	www	Attack Type	URL Length	Time Spent	Score	Fix. Ct.	Backtracking Fix. Ct.
C1	*safe*	*simple*	*www*	N/A	25.0 (4.8)	4.1 (2.3)	7.2 (1.1)	7.9 (4.9)	1.9 (1.8)
C2	*safe*	*simple*	*non-www*	N/A	19.8 (2.0)	4.0 (1.9)	3.8 (2.3)	7.1 (4.3)	1.6 (1.5)
C3	*safe*	*complex*	*www*	N/A	124.0 (13.2)	7.5 (3.8)	5.8 (1.5)	15.3 (8.2)	3.7 (3.1)
C4	*safe*	*complex*	*non-www*	N/A	105.3 (13.5)	7.9 (4.2)	4.4 (1.5)	15.9 (9.0)	4.1 (3.8)
C5	*unsafe*	*simple*	*www*	*positive*	28.5 (2.4)	5.4 (2.1)	4.4 (2.8)	9.5 (4.7)	2.4 (1.9)
C6	*unsafe*	*simple*	*www*	*negative*	29.3 (3.6)	4.8 (1.9)	5.9 (2.2)	9.2 (4.9)	2.4 (2.0)
C7	*unsafe*	*complex*	*non-www*	substring	96.0 (20.4)	7.4 (4.0)	5.5 (2.1)	14.5 (8.3)	3.7 (3.2)
C8	*unsafe*	*complex*	*www*	*user@host*	95.0 (17.4)	6.3 (3.4)	3.4 (2.4)	12.6 (7.4)	3.2 (3.2)

4.1.2 Complexity: URLs were grouped into two complexity classes: *simple* and *complex*. We define complexity in terms of (a) URL length and (b) URL features. A URL is *simple* if it is at most 36 characters long and does not contain a `rest` component. A URL is *complex* if it is at least 48 characters long and contains a non-empty path; it may also contain queries and fragments.

4.1.3 Presence of www: URLs with the *www* attribute begin with *https://www*. URLs with the *non-www* attribute do not.

4.1.4 Attack Type: We chose to explore four conditions for unsafe URLs. They are neither exhaustive nor fully representative of real-world attacks. Rather, our aim was to explore a variety of conditions that may affect visual behaviors and/or classification:

- ***positive*:** The fully qualified domain name contains positive or feel-good words or phrases, e.g., "happy", "bliss".
- ***negative*:** The fully qualified domain name contains words or phrases with a negative, technical, or a security connotation, e.g., "malware", "antivirus", "techsupport".
- ***substring*:** The fully qualified domain name has the form *https://X.Y.com* where *https://X.com* is a safe URL.
- ***user@host*:** The `authority` component has form *www.X.com@Y* where *https://www.X.com* is a legitimate URL. Moreover, some of the last four characters of *Y* are obfuscated using a hexadecimal representation, e.g., representing "*.com*" as "`.%63o%6D`".

The eight URL categories are presented in Table 2. In Section 4.5, we will discuss the measures in this table.

4.2 Experimental Design and Task

We conducted a within-subject experiment that was approved by the Institutional Review Board (IRB). Each of the 20 participants were shown the 64 URLs from the corpus over two sessions. The task was to classify each URL as safe or unsafe. Participants completed this task by viewing one URL at a time and clicking a button on the GUI to indicate whether they believed the URL was safe.

The URL corpus was split into two equal-sized sets presented over two sessions, such that four URLs from each category were represented in each set. For each session, the order in which URLs were presented was randomly determined but held fixed for all participants. However, session order alternated between participants.

4.3 Data Collection, Processing, & Analysis

We discuss the participant selection, the GUI, data collection, data processing, and data analysis:

4.3.1 Participants: We collected data from 20 participants (3 female, mean age = 22.68, SD = 2.65). All participants were students who participated in the user study as part of their coursework. We discarded data from 4 participants due to technical issues with the data extraction from the eye tracker. Hence, we report on the data from 16 participants (2 female).

4.3.2 User interface: The application was created using GUIs in MATLAB. It was presented to participants on a 24" monitor with a resolution of 1920x1200. Each URL image was created using bold monospace font [Wikipedia contributors 2019a] of size 64. The screen was made up of two panes. The first included the URL image, which was scaled and displayed on screen over 2-7 lines with a full line having approximate height of 20mm and width of 280mm. The second pane included the question "Is the web address safe to visit?", accompanied by two response buttons that read "Safe" and "Unsafe" (see Figure 1). Four markers were embedded in the application to identify the surface plane to mark various AOIs during post-processing of the eye-tracking data. Times of clicks and corresponding classifications/responses captured via button clicks were also recorded.

4.3.3 Eye Tracking: We used the head-mounted Dikablis eye tracker to collect gaze positions. It contains three cameras: two eye cameras sampling the eye at 60 Hz and a scene camera sampling at 30 Hz. Gaze positions are computed from the pupil movements and mapped onto the video from the scene camera. Establishing a mathematical mapping between the features of eye and the target being looked at is referred to as calibration. We used the four-point operator-controlled calibration method [Nyström et al. 2013].

4.3.4 Post-task questionnaire: Following the URL classification task, the participant filled in a questionnaire comprising: demographics questions; questions pertaining to security knowledge and behaviors, especially regarding URLs and phishing; and questions to help assess experimental validity.

4.3.5 Data Analysis: We used MATLAB for post-processing the eye-tracking data. We used JMP Pro 14 and R for statistical analysis. The Shapiro-Wilk test indicated that all of our data were non-normally distributed, thus we used non-parametric tests (Kruskal-Wallis test and Wilcoxon test) for analysis.

4.4 Procedure

After signing the consent form, the participant was given a brief introduction to the study and the user interface. They then saw a short neutral mood induction video to control for the effects of

Table 3: Disaggregation of a URL in accordance with the first three AOIs. This differs from Table 1 in that the scheme AOI includes the "://" following the scheme.

scheme AOI	*authority AOI*	*rest AOI*
https://	www.google.com	/forms/about/

mood. They then filled in a pre-task questionnaire to assess their mood [Schaefer et al. 2010], wore the eye tracker, and completed a practice trial to familiarize themselves with the task and the GUI.

Before calibration, we adjusted a nose pin and head band to reduce the movement of the eye tracker during the study; we did not use a chin rest. Next, we focused the eye and scene cameras and calibrated the eye tracker using the four-point operator-controlled calibration method. The participant then classified URLs for the first session and took a break. The calibration procedure was then repeated and the participant classified URLs for the second session. Last, they filled in the post-task questionnaire. The distance between the screen and the participant was kept at about 0.6 meters.

4.5 Measures

4.5.1 Mood: Each participant's mood was assessed along six emotional states: awake, pleasant, angry, fearful, happy, and sad [Mills et al. 2019]. The assessment used a 10-point scale, where 1 indicated that the participant's mood was not associated with the given emotional state, and 10 indicated that it was highly associated.

4.5.2 Score: The score represents the number of correctly classified URLs within a set with no penalty for incorrect classification.

4.5.3 Total Time Spent: The total time spent on classifying a URL is the time (seconds) from the presentation of the URL to the time when the user clicks on a button to classify it. This is a proxy for the cumulative effort and engagement in classifying the URL.

4.5.4 Time Spent on Areas of Interest: Using the UTC timestamps of each data point recorded by the eye tracker, we computed the percentage dwell time on five AOIs (Areas of Interest). These measures express the distribution of users' visual attention and help us understand which URL components users use to gauge URL safety. We examined five AOIs. Figure 1 captures the first four AOIs and Table 3 gives a disaggregation of a URL in accordance with the AOIs that correspond to the URL. We now present the five AOIs.

- The ***scheme AOI*** captures the `scheme` component and the delimiters immediately following it. As every URL in our corpus uses the *https* as the `scheme`, this AOI always corresponds to the leading *https://* in the URL.
- The ***authority AOI*** captures the `authority` component. For classes C1 through C7, the `authority` component is a fully qualified domain name, e.g., *www.google.com* is the `authority` component of *https://www.google.com*. For class C8, the `authority` component has form `user@host`, e.g., as in *www.google.com@evil.com*. To test ***H5***, the ***authority AOI*** was further split into two smaller AOIs, the ***user AOI*** and the ***host AOI*** corresponding to the `user` and `host` components.
- The ***rest AOI*** captures the `rest` component.
- The ***response AOI*** captures the response portion of the screen containing the "Safe" and "Unsafe" buttons.

Table 4: Probabilities of correctly classifying *safe* URLs given the participant knew of the service.

Probabilities	P[correct|known]	P[correct|unknown]
C1 (simple, www)	0.92	0.63
C2 (simple, non-www)	0.83	0.19
C3 (complex, www)	0.76	0.5
C4 (complex, non-www)	0.58	0.46

- The last AOI captured visual targets other than the previous four areas of interest.

4.5.5 Fixations and Backtracking Fixations: Fixating is the act of maintaining one's gaze at a particular target for a certain duration of time. It represents the time where new information is gathered [Ramkumar et al. 2019]. We extracted fixations of 100ms or more following prior research guidelines [Irwin and Zelinsky 2002; Munn et al. 2008; Salvucci and Goldberg 2000].

Backtracking is the process of revisiting information that was previously processed or skipped [Bruneau et al. 2002]. It usually occurs to re-establish previously processed information or it signifies a cognitive interest in an area with respect to the given task [Burton and Daneman 2007]. We measured the backtrack fixation count, i.e., the number of fixations involving backtracking.

4.5.6 Normalized Pupil Area: : The eye tracker records raw pupil area of both eyes in pixels. We used the right eye pupil area. We used the Hampel identifier technique to remove outliers [Foroughi et al.2017; Pearson et al. 2016]. Due to the non-uniform sampling rate, we interpolated the data to obtain a uniform sampling frequency of 60 Hz [Pfleging et al. 2016]. Then, we normalized the data to compare it between participants.

4.5.7 Accounting for Length Differences in URLs: URLs may differ in the number of characters in their `scheme`, `authority`, and `rest` components. Thus, for the corresponding AOIs, we calculated the time spent per character (total time spent on AOI divided by number of characters in AOI) and the fixation count per character (total number of fixations occurring on AOI divided by total number of characters in AOI). For the overall comparison, we computed overall time spent per character (total time spent/total URL length), overall fixation count per character (total fixation count/total URL length), and backtrack fixation count as a function of URL length (total backtrack fixations/total URL length).

5 RESULTS

5.1 Mood Induction Measures

On average participants were awake (ranking of M=7.50, SD=1.59), felt relatively pleasant (M=7.69, SD=1.40), and were mildly happy (M=6.75, SD=1.44). They did not feel angry (M=1.81, SD=0.83), fearful (M=1.56, SD=1.09), or sad (M=1.50, SD=0.82).

5.2 Scores

The average score was 40.44 out of 64. From the post-task questionnaire, we were able to identify whether the participants knew of the services associated with the *safe* URLs. Table 4 indicates the probabilities of participants correctly classifying the URL given that

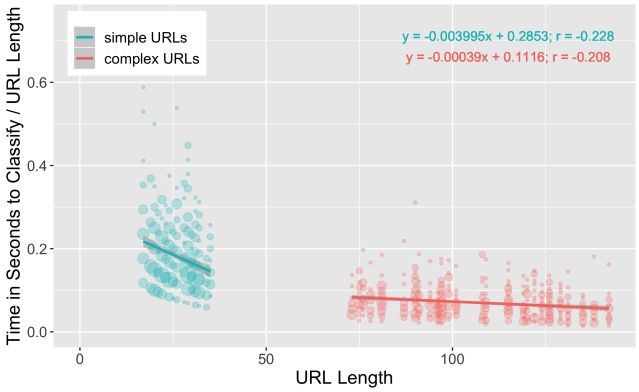

Figure 2: Time spent per character to classify URL vs. URL length with linear regression lines.

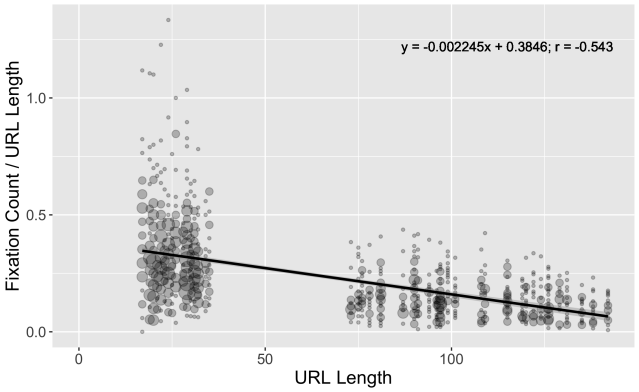

Figure 3: Fixation count per character vs. URL length with a linear regression line.

they knew the service. The Kruskal-Wallis test showed no significant difference between the four categories of *safe* URLs (C1-C4) in terms of the participant knowing the services associated with the domain names [$X^2(3)$=6.9674, p=0.0729].

5.3 Overview of Eye-Tracking Results

Table 2 presents some key results. The overall distribution of visual attention on the AOIs is shown in Figure 6. Using Kruskal-Wallis test, we found that the time spent per character was significantly different between the three AOIs corresponding to the URL [$X^2(2)$=30.4152, p<0.0001]. Post hoc analysis indicated time spent per character on the ***authority AOI*** was significantly higher than that of the ***scheme AOI*** and that of the ***rest AOI***. The fixation count per character was significantly different between the three AOIs [Kruskal-Wallis test: $X^2(2)$=23.9356, p<0.0001]. Post hoc analysis indicated that fixation count per character on the ***rest AOI*** was significantly lower than the other two. However, we found no evidence that fixation duration was significantly different between the three AOIs [Kruskal-Wallis test: $X^2(2)$=3.1692, p=0.0516].

The Kruskal-Wallis test indicated a significant difference in normalized pupil area [$X^2(2)$=8.7532, p=0.0126]. Post hoc analysis indicated a lower pupil area for the ***scheme AOI*** relative to other AOIs, suggesting less cognitive effort was expended on the ***scheme AOI***.

5.4 Complexity

We saw a significant difference in overall time spent (seconds) processing between *complex* and *simple* URLs [Wilcoxon test: Z=3.4865, p=0.0005]. More time was spent on *complex* URLs (M=7.26, SD=2.41) compared to *simple* URLs (M=4.58, SD=1.35). This can also be seen pictorially in Figure 4. Wilcoxon test indicated significant differences in overall time spent per character [Z=8.9998, p<0.0001], overall fixation count per character [Z=6.4883, p<0.0001], and backtrack fixation count as a function of URL length [Z=4.4399, p<0.0001].

People spent less time per character on *complex* URLs (M=0.06, SD=0.01) than *simple* URLs (M=0.13, SD=0.04). Figure 2 shows the time spent per character decreases as URL length increases. Also, the fixation count per character was smaller for *complex* URLs (M=0.12, SD=0.04) than for *simple* URLs (M=0.22, SD=0.10). Figure 3 shows a decrease in fixation count per character as URL length increases. But the backtrack fixation count was higher on *complex* URLs (M=3.68, SD=2.44) relative to *simple* ones (M=2.08, SD=1.18). We found no significant difference in the score between *complex* (M=4.76, SD=2.10) and *simple* URLs (M=5.34, SD=2.51). Examining *complex* URLs of different lengths tells a more nuanced story. Figure 5 suggests a peak in time spent per character that occurs near 100 characters. We observed similar trends with fixation count per character and backtrack fixation count as a function of URL length for *complex* URLs.

5.5 Existence of *www*

We compared *safe* URLs that have `authority` components that begin with *www* (C1&C3) to those that do not (C2&C4). Wilcoxon test results indicated a significant difference in time spent per character on the ***authority AOI*** between *www* URLs (M=0.16, SD=0.04) and *non-www* URLs (M=0.21, SD=0.04); [Z=4.2094, p<0.0001]. Also, there was a significant difference in the fixation count per character on the ***authority AOI*** between *www* URLs (M=0.24, SD=0.09) and *non-www* URLs (M=0.34, SD=0.12); [Wilcoxon test: Z=3.2292, p=0.0012]. The score obtained (maximum score: 8) was also significantly different between *www* URLs (M=6.50, SD=1.48) and *non-www* URLs (M=4.09, SD=1.90); [Wilcoxon test: Z=4.7020, p<0.001].

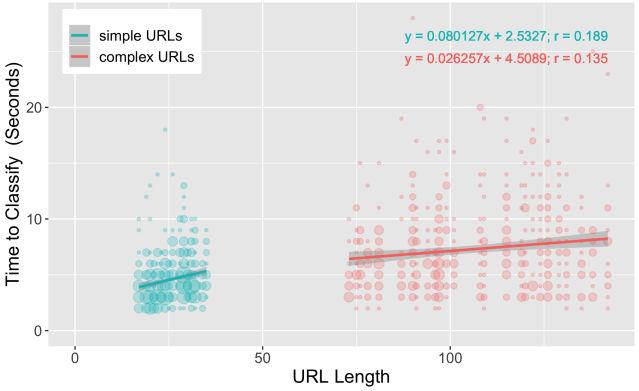

Figure 4: Time spent to classify URL vs. URL length with linear regression lines for simple and complex URLs.

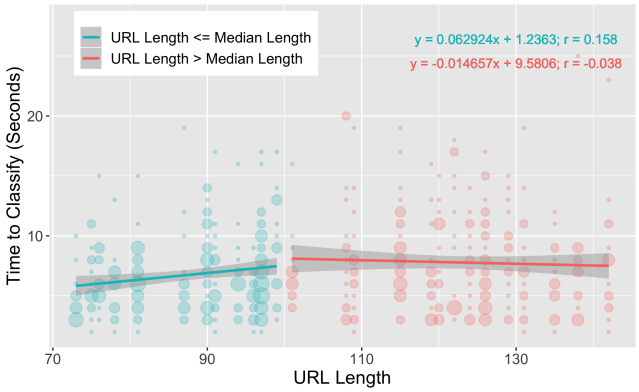

Figure 5: Time spent to classify URLs vs. URL length with two linear regression lines for data points separated by the median URL length (complex URLs).

5.6 User@Host Attack Type vs. Regular URLs

To examine user visual attention for the *user@host* URLs (C8), we considered two special AOIs at a finer granularity than the ***authority AOI***: the ***user AOI*** and ***host AOI***. We compared measurements on these two AOIs for the *user@host* URLs (C8) to those for the ***authority AOI*** for safe URLs of similar structure (C3). Using the Kruskal-Wallis test we found a significant difference on time spent per character between the ***authority AOI*** of C3, the ***user AOI*** of C8, and the ***host AOI*** of C8 [$X^2(2)$=32.1735, p<0.0001]. A significant difference was also observed with fixation count per character [Kruskal-Wallis test: $X^2(2)$=11.3323, p=0.0035]. Post hoc analysis indicated that both sets of measurements for the ***host AOI*** for C8 were lower than those of the ***user AOI*** for C8 and the ***authority AOI*** for C3; the measurements between the ***user AOI*** for C8 were comparable to those of the ***authority AOI*** for C3. These results suggests that users process the ***user AOI*** of C8 and the ***authority AOI*** of C3 similarly. Also, there was a significant difference in the score between the user@host attack type (M=3.37, SD=2.41) and safe URLs of similar structure (M=5.81, SD=1.51); [Wilcoxon test: Z=2.9176, p=0.0035].

6 DISCUSSION

First, participant responses to the pre-task questionnaire following the mood induction video [Schaefer et al. 2010] indicate they were awake and in a neutral mood. Responses to the post-task questionnaire reveal that participants did not fatigue, and, on average, correctly identified the safety of about 40 of the 64 URLs (63%).

Let us now turn to a detailed discussion of the results.

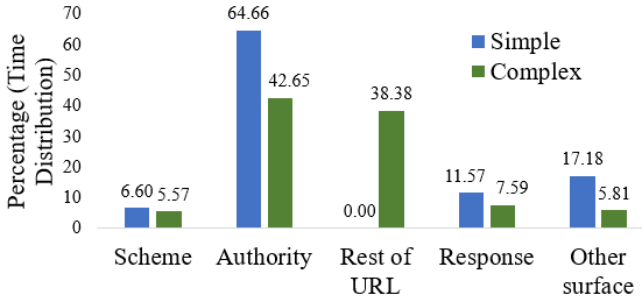

Figure 6: Percentage of time distribution on various AOIs.

6.1 URL Processing & Classification Factors

6.1.1 URL Length: The overall time spent on classifying *simple* (and shorter) URLs (C1, C2, C5, C6) was less than the total time spent on classifying *complex* (and longer) URLs (C3, C4, C7, C8). This weakly supports ***H1***, though follow-up work must be done to disentangle length from other complexity factors.

For *complex* URLs, we found URL length negatively correlated with time spent per character and fixation count per character. This supports ***H2***.

We did not observe a correlation between URL length and score. Also, while Figure 4 suggests participants spent more time parsing URLs as URL length increases, Figure 2 suggests time spent per character decreases as we increase URL length. Moreover, the positive correlation between URL length and time spent seems to cease at a point, which supports ***H3***. Specifically, Figure 5 suggests that at a threshold of approximately 100 characters, time spent stops increasing as we increase URL length. Similar trends were observed with fixation count per character and backtrack fixation count per character. We also observed no statistical difference between time spent on complex URLs under 100 characters and those above. One interpretation is captured by a notion similar to that of the compliance budget proposed by Beautement et al. [Beautement et al. 2008]: the user may only expend a finite budget of resources (here, time is a proxy for expended resources) to classify a URL, and, if the resources required to fully process a URL exceeds this budget, the user will not expend them. While the peculiarities of where that threshold is may depend on factors other than just URL length, we expect this notion of a finite budget applies more generally.

6.1.2 AOI:. We examine the influence of the AOIs:

- **Scheme AOI**: The decrease in the pupil area for the ***scheme AOI*** indicates reduced cognitive attention. Previous work found the frequency with which a user encounters a word affects the fixation duration and processing of that word [Rayner and Duffy 1986]. Users usually spend less time on frequently encountered words. Most legitimate websites use *https* nowadays, which is also used in each of the 64 URLs in our corpus. This explains the decrease in cognitive load for the ***scheme AOI***. We observed a statistically significant difference in time spent per character between the ***scheme AOI*** and the ***authority AOI*** (with the latter being higher); however, we did not observe such a difference for the ***scheme AOI*** and the ***rest AOI***. Therefore, we do not have evidence to support ***H4***.
- **Authority AOI:** The results indicate the time spent per character on the ***authority AOI*** is significantly higher than that of other AOIs. Time spent and fixation count per character on the ***authority AOI*** suggests users find *www* at the beginning of the domain name to be a strong indicator of URL safety.
- **Rest AOI:** Reduced fixation count while reading is characteristic of scanning text [Rayner and Fischer 1996]. The fixation count per character for the rest AOI is significantly

lower than it is for other AOIs, which suggests participants scanned the ***rest AOI***.

6.1.3 Attack Types: Participants classified *positive*, *unsafe* URLs (C5) correctly 55% of the time and they classified *negative*, *unsafe* URLs (C6) correctly 74% of the time. This suggests people are more inclined to trust URLs that use positive words or phrases, even if they have no familiarity with the domain name. Table 4 shows that participants, on average, correctly classified the URLs 77% of the time, given that they had heard of the associated services.

Results suggest users visually process the `user` component of URLs with the *user@host* attack type (C8) similar to how they process the `authority` of URLs without a `user` component. In general, the fixation count per character was low for the `rest` component relative to both the `scheme` and `authority` components. For C8, we observed a reduced fixation count per character and time spent per character on the `host` component, which suggests participants perceived the `host` component as part of the `rest` component. Visual evidence suggests participants misidentified the `user` component as the `host` for URLs in C8. Of the *unsafe* URL categories, participants scored worst on C8. Participants spent significantly more time per character on the `user` component than the `host` component for C8, in support of ***H5***.

We expect classification accuracies observed in this study are upper bounds on what users achieve in practice without additional safeguards in place. Sophisticated attacks that use URL features participants do not know about will likely be more effective. We also expect that attacks that use obfuscation in the `rest` component - or what users *perceive* as the `rest` component - are more likely to succeed given that participants spent less time on the `rest` component than the `authority` component in our study.

6.2 Improving Security in Practice

The study suggests a sort of ceiling effect: as URL length increases, participants spent more time vetting the URL until it capped out at around 100 characters. It also provides visual evidence of user misperceptions regarding URL structure. These insights into how users process and perceive URLs suggest concrete steps and best practices for services to improve the perceived security - and, we argue, the *actual* security - associated with the URLs they serve. For example, from a purely technical standpoint, there is no intrinsic security benefit to serving a URL that is short, has a domain name that begins with *www*, and has few special characters. But if those URLs match users' safety expectations, users would be better at classifying both safe URLs served by the service and unsafe, obfuscated URLs served by adversaries.

Some *unsafe* URLs from our corpus were classified as safe because they exploited uncommon URL features that users rarely encounter in practice with legitimate services. Ironically, this makes such URLs easy for a computer to classify as risky. Surprisingly, we found that some web browsers offer no user protection against such URLs, even though simple-to-write parsers could easily detect them. This provides an opportunity to improve security at minimal cost.

Last, our findings can improve the quality of security awareness training programs. Our study identifies various misperceptions held by users. It also provides concrete evidence of where users look as they process URLs. This study's methods and data may help in assessing, comparing, and improving training modules that aim to help users correctly identify URLs.

7 LIMITATIONS

Several considerations may have affected study generalizability: Participants were predominantly male college students pursuing electrical engineering degrees. To ensure the eye tracker accurately picked up on AOIs, we used a large font and displayed URLs over multiple lines. URLs were presented in isolation; contextual factors (e.g., the device on which a URL is displayed, the application on which a URL is viewed, or beliefs regarding who sent it) may affect visual behaviors and responses. Also, repeatedly asking participants whether URLs were safe likely sensitized them to phishing attacks.

However, we took precautions to minimize unintended effects. We conducted pilot runs to ensure the interface was clear and user fatigue was minimized. We used the post-experiment questionnaire to evaluate experimental validity. And we used a neutral-mood-inducing video to reduce variability in mood.

The available indicators provide some evidence of the study's validity. The average participant score of 63% is within the ballpark of similar studies, e.g., [Dhamija et al. 2006; Sheng et al. 2010]. Post-task survey responses indicate most participants took the task seriously, exercised equal or only slightly more caution than they would in practice, and were not fatigued. Though we did not observe significant bias, we believe any bias would be in the direction of more caution and would be unlikely to invalidate our security recommendations as problems during the classification task would also be at play in the real world. We also note that applications and interfaces in the wild may vary regarding font properties so there is no one-size-fits-all approach for conducting such studies.

Last, the URLs may have had features we could not identify that affected participants' visual behaviors and responses. We attempted to mitigate these concerns by including eight URLs per category, but further work is needed. Also, we only considered a few flavors of URL-based attacks. Notably, no attacks made use of the `rest` component, which may have affected participants' visual behaviors.

8 CONCLUSION AND FUTURE WORK

Eye tracking provides a window to examine security behavior. This paper is a first step toward developing a model that captures how users visually process, derive meaning from, and operationalize URL security information to gauge URL safety. We conducted a user study in which participants saw URLs and then classified them while wearing an eye tracker. The findings suggest that participants relied on poor security indicators such as presence of *www* to gauge URL legitimacy, that they spent more time and cognitive resources to vet longer URLs but only up to a point, and that, for the unsafe, *user@host* URLs, participants perceived the `user` component to be the `host` component. In future work, we plan to study other contextual factors such as mood, additional flavors of URL obfuscation, and the effectiveness of training the user.

ACKNOWLEDGMENTS

The work of Ramkumar and Kun was supported in part by NSF grant OISE 1658594.

REFERENCES

A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg, and E. Almomani. 2013. A Survey of Phishing Email Filtering Techniques. *IEEE Communications Surveys Tutorials* 15, 4 (Fourth 2013), 2070–2090. https://doi.org/10.1109/SURV.2013.030713.00020

Mohamed Alsharnouby, Furkan Alaca, and Sonia Chiasson. 2015. Why phishing still works: User strategies for combating phishing attacks. *International Journal of Human-Computer Studies* 82 (2015), 69 – 82. https://doi.org/10.1016/j.ijhcs.2015.05.005

K. Althobaiti, G. Rummani, and K. Vaniea. 2019. A Review of Human- and Computer-Facing URL Phishing Features. In *2019 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. 182–191. https://doi.org/10.1109/EuroSPW.2019.00027

Nalin Asanka Gamagedara Arachchilage, Steve Love, and Konstantin Beznosov. 2016. Phishing threat avoidance behaviour: An empirical investigation. *Computers in Human Behavior* 60 (2016), 185–197.

Jackson Beatty. 1982. Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin* 91, 2 (1982), 276.

Adam Beautement, M. Angela Sasse, and Mike Wonham. 2008. The Compliance Budget: Managing Security Behaviour in Organisations. In *Proceedings of the 2008 New Security Paradigms Workshop (NSPW '08)*. ACM, New York, NY, USA, 47–58. https://doi.org/10.1145/1595676.1595684

Zinaida Benenson, Freya Gassmann, and Robert Landwirth. 2017. Unpacking Spear Phishing Susceptibility. In *Financial Cryptography and Data Security*, Michael Brenner, Kurt Rohloff, Joseph Bonneau, Andrew Miller, Peter Y.A. Ryan, Vanessa Teague, Andrea Bracciali, Massimiliano Sala, Federico Pintore, and Markus Jakobsson (Eds.). Springer International Publishing, Cham, 610–627.

Z. Benenson, A. Girard, N. Hintz, and A. Luder. 2014. Susceptibility to URL-based Internet attacks: Facebook vs. email. In *2014 IEEE International Conference on Pervasive Computing and Communication Workshops (PERCOM WORKSHOPS)*. IEEE, 604–609. https://doi.org/10.1109/PerComW.2014.6815275

André Bergholz, Jan De Beer, Sebastian Glahn, Marie-Francine Moens, Gerhard Paaß, and Siehyun Strobel. 2010. New Filtering Approaches for Phishing Email. *Journal of Computer Security* 18, 1 (Jan. 2010), 7–35. http://dl.acm.org/citation.cfm?id=1734234.1734239

Jennifer Romano Bergstrom and Andrew Schall. 2014. *Eye Tracking in User Experience Design*. Elsevier.

Tim Berners-Lee, Roy Fielding, and Larry Masinter. 1998. *Uniform Resource Identifiers (URI): Generic Syntax*. RFC 3986. RFC Editor. 1–61 pages. https://tools.ietf.org/html/rfc3986

Tim Berners-Lee, Larry Masinter, Mark McCahill, et al. 1994. *Uniform Resource Locators (URL)*. RFC 1738. RFC Editor. 1–25 pages. https://tools.ietf.org/html/rfc1738

David Beymer, Daniel Russell, and Peter Orton. 2008. An Eye Tracking Study of How Font Size and Type Influence Online Reading. In *Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 2 (BCS-HCI '08)*. BCS Learning & Development Ltd., Swindon, UK, 15–18. http://dl.acm.org/citation.cfm?id=1531826.1531831

David Beymer and Daniel M. Russell. 2005. WebGazeAnalyzer: A System for Capturing and Analyzing Web Reading Behavior Using Eye Gaze. In *CHI '05 Extended Abstracts on Human Factors in Computing Systems (CHI EA '05)*. ACM, New York, NY, USA, 1913–1916. https://doi.org/10.1145/1056808.1057055

Aaron Blum, Brad Wardman, Thamar Solorio, and Gary Warner. 2010. Lexical Feature Based Phishing URL Detection Using Online Learning. In *Proceedings of the 3rd ACM Workshop on Artificial Intelligence and Security (AISec '10)*. ACM, New York, NY, USA, 54–60. https://doi.org/10.1145/1866423.1866434

Catherine M Bohn-Gettler and David N Rapp. 2011. Depending on My Mood: Mood-Driven Influences on Text Comprehension. *Journal of Educational Psychology* 103, 3 (2011), 562.

Margaret M. Bradley, Laura Miccoli, Miguel A. Escrig, and Peter J. Lang. 2008. The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology* 45, 4 (2008), 602–607. https://doi.org/10.1111/j.1469-8986.2008.00654.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-8986.2008.00654.x

Daniel Bruneau, M. Angela Sasse, and John Mccarthy. 2002. The Eyes Never Lie: The Use of Eye Tracking Data in HCI Research. In *In Proceedings of the CHI'02 Workshop on Physiological Computing*. ACM Press.

Christine Burton and Meredyth Daneman. 2007. Compensating for a Limited Working Memory Capacity During Reading: Evidence from Eye Movements. *Reading Psychology* 28, 2 (2007), 163–186. https://doi.org/10.1080/02702710601186407 arXiv:https://doi.org/10.1080/02702710601186407

Bennett Cyphers, Alexei Miagkov, and Andrés Arrieta. 2018. Privacy Badger Now Fights More Sneaky Google Tracking. https://www.eff.org/deeplinks/2018/10/privacy-badger-now-fights-more-sneaky-google-tracking.

Rachna Dhamija, J. D. Tygar, and Marti Hearst. 2006. Why Phishing Works. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 581–590. https://doi.org/10.1145/1124772.1124861

Julie S. Downs, Mandy Holbrook, and Lorrie Faith Cranor. 2007. Behavioral Response to Phishing Risk. In *Proceedings of the Anti-phishing Working Groups 2nd Annual eCrime Researchers Summit (eCrime '07)*. ACM, New York, NY, USA, 37–44. https://doi.org/10.1145/1299015.1299019

Christine E. Drake, Jonathan J. Oliver, and Eugene J. Koontz. 2004. Anatomy of a Phishing Email. In *CEAS 2004 - First Conference on Email and Anti-Spam, July 30-31, 2004, Mountain View, California, USA*.

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 (CHI '08)*. ACM, New York, NY, USA, 1065–1074. https://doi.org/10.1145/1357054.1357219

Ian Fette, Norman Sadeh, and Anthony Tomasic. 2007. Learning to Detect Phishing Emails. In *Proceedings of the 16th International Conference on World Wide Web (WWW '07)*. ACM, New York, NY, USA, 649–656. https://doi.org/10.1145/1242572.1242660

Joseph P Forgas. 1989. Mood effects on decision making strategies. *Australian Journal of Psychology* 41, 2 (1989), 197–214.

Sanjay Goel, Kevin Williams, and Ersin Dincelli. 2017. Got Phished? Internet Security and Human Vulnerability. *Journal of the Association for Information Systems* 18, 1 (2017), 2.

Joseph H. Goldberg, Mark J. Stimson, Marion Lewenstein, Neil Scott, and Anna M. Wichansky. 2002. Eye Tracking in Web Search Tasks: Design Implications. In *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications (ETRA '02)*. ACM, New York, NY, USA, 51–58. https://doi.org/10.1145/507072.507082

Kyung Wha Hong, Christopher M. Kelley, Rucha Tembe, Emerson Murphy-Hill, and Christopher B. Mayhorn. 2013. Keeping Up With The Joneses: Assessing Phishing Susceptibility in an Email Task. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 57, 1, 1012–1016. https://doi.org/10.1177/1541931213571226 arXiv:https://doi.org/10.1177/1541931213571226

Jukka Hyönä, Robert F Lorch Jr, and Johanna K Kaakinen. 2002. Individual Differences in Reading to Summarize Expository Text: Evidence From Eye Fixation Patterns. *Journal of Educational Psychology* 94, 1 (2002), 44. https://doi.org/10.1037//0022-0663.94.1.44

Jukka Hyönä, Robert F Lorch Jr, and Mike Rinck. 2003. Chapter 16 - Eye Movement Measures to Study Global Text Processing. In *The Mind's Eye*, Jukka Hyönä, Ralph Radach, and Heiner Deubel (Eds.). North-Holland, Amsterdam, 313–334. https://doi.org/10.1016/B978-044451020-4/50018-9

David E Irwin and Gregory J Zelinsky. 2002. Eye movements and scene perception: Memory for things observed. *Perception & Psychophysics* 64, 6 (2002), 882–895.

Marcel A Just and Patricia A Carpenter. 1980. A Theory of Reading: From Eye Fixations to Comprehension. *Psychological review* 87, 4 (1980), 329.

KnowBe4. 2019. Kevin Mitnick Security Awareness Training. https://www.knowbe4.com/products/kevin-mitnick-security-awareness-training/.

Peter König, Niklas Wilming, Tim C Kietzmann, Jose P Ossandón, Selim Onat, Benedikt V Ehinger, Ricardo R Gameiro, and Kai Kaspar. 2016. Eye movements as a window to cognitive processes. *Journal of Eye Movement Research* 9, 5 (2016), 1–16.

Ponnurangam Kumaraguru, Justin Cranshaw, Alessandro Acquisti, Lorrie Cranor, Jason Hong, Mary Ann Blair, and Theodore Pham. 2009. School of Phish: A Real-world Evaluation of Anti-phishing Training. In *Proceedings of the 5th Symposium on Usable Privacy and Security (SOUPS '09)*. ACM, New York, NY, USA, Article 3, 12 pages. https://doi.org/10.1145/1572532.1572536

Ponnurangam Kumaraguru, Yong Rhee, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. 2007. Protecting People from Phishing: The Design and Evaluation of an Embedded Training Email System. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '07)*. ACM, New York, NY, USA, 905–914. https://doi.org/10.1145/1240624.1240760

Andrew L. Kun, Oskar Palinko, Zeljko Medenica, and Peter Heeman. 2013. On the Feasibility of Using Pupil Diameter to Estimate Cognitive Load Changes for In-Vehicle Spoken Dialogues. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*. International Speech and Communication Association, 3766–3770.

Meng-Lung Lai, Meng-Jung Tsai, Fang-Ying Yang, Chung-Yuan Hsu, Tzu-Chien Liu, Silvia Wen-Yu Lee, Min-Hsien Lee, Guo-Li Chiou, Jyh-Chong Liang, and Chin-Chung Tsai. 2013. A review of using eye-tracking technology in exploring learning from 2000 to 2012. *Educational Research Review* 10 (2013), 90 – 115. https://doi.org/10.1016/j.edurev.2013.10.001

Max-Emanuel Maurer, Alexander De Luca, and Sylvia Kempe. 2011. Using Data Type Based Security Alert Dialogs to Raise Online Security Awareness. In *Proceedings of the Seventh Symposium on Usable Privacy and Security (SOUPS '11)*. ACM, New York, NY, USA, Article 2, 13 pages. https://doi.org/10.1145/2078827.2078830

Microsoft. 2019. Office 365 ATP Safe Links. https://docs.microsoft.com/en-us/office365/securitycompliance/atp-safe-links.

Caitlin Mills, Jennifer Wu, and Sidney D'Mello. 2019. Being Sad Is Not Always Bad: The Influence of Affect on Expository Text Comprehension. *Discourse Processes* 56, 2 (2019), 99–116. https://doi.org/10.1080/0163853X.2017.1381059 arXiv:https://doi.org/10.1080/0163853X.2017.1381059

D. Miyamoto, T. Iimura, G. Blanc, H. Tazaki, and Y. Kadobayashi. 2014. EyeBit: Eye-Tracking Approach for Enforcing Phishing Prevention Habits. In *2014 Third International Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*. 56–65. https://doi.org/10.1109/BADGERS.2014.14

Mozilla. 2019. How does built-in Phishing and Malware Protection work? https://support.mozilla.org/en-US/kb/how-does-phishing-and-malware-protection-work.

Susan M. Munn, Leanne Stefano, and Jeff B. Pelz. 2008. Fixation-identification in dynamic scenes: Comparing an automated algorithm to manual coding. In *Proceedings of the 5th Symposium on Applied Perception in Graphics and Visualization (APGV '08)*. ACM, New York, NY, USA, 33–42. https://doi.org/10.1145/1394281.1394287

Yoav Nathaniel. 2017. Attack Report: Office 365 Security Hacked Using Google Redirect. https://www.avanan.com/resources/open-redirect-vulnerability.

Marcus Nyström, Richard Andersson, Kenneth Holmqvist, and Joost van de Weijer. 2013. The influence of calibration method and eye physiology on eyetracking data quality. *Behavior Research Methods* 45, 1 (01 Mar 2013), 272–288. https://doi.org/10.3758/s13428-012-0247-4

Gunter Ollmann. 2004. The Phishing Guide–Understanding & Preventing Phishing Attacks. *NGS Software Insight Security Research* (2004).

Oskar Palinko and Andrew L. Kun. 2012. Exploring the Effects of Visual Cognitive Load and Illumination on Pupil Diameter in Driving Simulators. In *Proceedings of the Symposium on Eye Tracking Research & Applications (ETRA '12)*. ACM, New York, NY, USA, 413–416. https://doi.org/10.1145/2168556.2168650

Oskar Palinko, Andrew L. Kun, Alexander Shyrokov, and Peter Heeman. 2010. Estimating Cognitive Load Using Remote Eye Tracking in a Driving Simulator. In *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA '10)*. ACM, New York, NY, USA, 141–144. https://doi.org/10.1145/1743666.1743701

D. Pappusetty, V. V. R. Chinta, and H. Kalva. 2017. Using Pupillary Response to Assess Video Quality. In *2017 IEEE International Conference on Consumer Electronics (ICCE)*. 64–65. https://doi.org/10.1109/ICCE.2017.7889231

Bastian Pfleging, Drea K. Fekety, Albrecht Schmidt, and Andrew L. Kun. 2016. A Model Relating Pupil Diameter to Mental Workload and Lighting Conditions. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 5776–5788. https://doi.org/10.1145/2858036.2858117

PhishingBox. 2019. phishingbox. https://www.phishingbox.com

PhishLabs. 2019. Security Awareness Training. https://www.phishlabs.com/security-awareness-training/.

Marc Pomplun and Sindhura Sunkara. 2003. Pupil Dilation as an Indicator of Cognitive Workload in Human-Computer Interaction. In *Proceedings of the International Conference on HCI*.

A Poole and Linden Ball. 2006. *Eye Tracking in Human-Computer Interaction and Usability Research: Current Status and Future Prospects*. Idea Group, Inc, 211–219.

Proofpoint. 2019a. Proofpoint Essentials URL Defense: Advanced Protection with Proofpoint's Targeted Attack Protection. https://www.proofpoint.com/us/resources/data-sheets/essentials-url-defense.

Proofpoint. 2019b. Proofpoint Security Awareness Training. https://www.proofpoint.com/us/product-family/security-awareness-training.

Niveta Ramkumar, Nadia Fereydooni, Orit Shaer, and Andrew L. Kun. 2019. Visual Behavior During Engagement with Tangible and Virtual Representations of Archaeological Artifacts. In *Proceedings of the 8th ACM International Symposium on Pervasive Displays (PerDis '19)*. ACM, New York, NY, USA, Article 21, 7 pages. https://doi.org/10.1145/3321335.3324930

Keith Rayner. 1998. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychological bulletin* 124 3 (1998), 372–422.

Keith Rayner, Kathryn H Chace, Timothy J Slattery, and Jane Ashby. 2006. Eye Movements as Reflections of Comprehension Processes in Reading. *Scientific Studies of Reading* 10, 3 (2006), 241–255. https://doi.org/10.1207/s1532799xssr1003_3 arXiv:https://doi.org/10.1207/s1532799xssr1003_3

Keith Rayner and Susan A. Duffy. 1986. Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity. *Memory & Cognition* 14, 3 (01 May 1986), 191–201. https://doi.org/10.3758/BF03197692

Keith Rayner and Martin H. Fischer. 1996. Mindless reading revisited: Eye movements during reading and scanning are different. *Perception & Psychophysics* 58, 5 (01 Jul 1996), 734–747. https://doi.org/10.3758/BF03213106

Erik D Reichle, Alexander Pollatsek, and Keith Rayner. 2012. Using E-Z Reader to simulate eye movements in nonreading tasks: A unified framework for understanding the eye–mind link. *Psychological Review* 119, 1 (2012), 155–185.

Dario D. Salvucci and Joseph H. Goldberg. 2000. Identifying Fixations and Saccades in Eye-tracking Protocols. In *Proceedings of the 2000 Symposium on Eye Tracking Research & Applications (ETRA '00)*. ACM, New York, NY, USA, 71–78. https://doi.org/10.1145/355017.355028

SANS. 2019a. Robust Phishing Awareness Simulation Training that Changes Behavior. https://www.sans.org/security-awareness-training/products/phishing.

SANS. 2019b. The 2019 SANS EndUser Training Suite. https://www.sans.org/security-awareness-training/products/end-user.

Alexandre Schaefer, Frédéric Nils, Xavier Sanchez, and Pierre Philippot. 2010. Assessing the effectiveness of a large database of emotion-eliciting films: A new tool for emotion researchers. *Cognition and Emotion* 24, 7 (2010), 1153–1172. https://doi.org/10.1080/02699930903274322 arXiv:https://doi.org/10.1080/02699930903274322

S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. 2007. The Emperor's New Security Indicators. In *2007 IEEE Symposium on Security and Privacy (SP '07)*. 51–65. https://doi.org/10.1109/SP.2007.35

Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who Falls for Phish?: A Demographic Analysis of Phishing Susceptibility and Effectiveness of Interventions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '10)*. ACM, New York, NY, USA, 373–382. https://doi.org/10.1145/1753326.1753383

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 (SOUPS '07)*. ACM, New York, NY, USA, 88–99. https://doi.org/10.1145/1280680.1280692

John L. Sibert, Mehmet Gokturk, and Robert A. Lavine. 2000. The Reading Assistant: Eye Gaze Triggered Auditory Prompting for Reading Remediation. In *Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology (UIST '00)*. ACM, New York, NY, USA, 101–107. https://doi.org/10.1145/354401.354418

Robert F. Stanners, Michelle Coulter, Allen W. Sweet, and Philip Murphy. 1979. The Pupillary Response as an Indicator of Arousal and Cognition. *Motivation and Emotion* 3, 4 (01 Dec 1979), 319–340. https://doi.org/10.1007/BF00994048

Simon Stockhardt, Benjamin Reinheimer, Melanie Volkamer, Peter Mayer, Alexandra Kunz, Philipp Rack, and Daniel Lehmann. 2016. Teaching Phishing-Security: Which Way is Best?. In *ICT Systems Security and Privacy Protection*, Jaap-Henk Hoepman and Stefan Katzenbeisser (Eds.). Springer International Publishing, Cham, 135–149.

Zikai Alex Wen, Zhiqiu Lin, Rowena Chen, and Erik Andersen. 2019. What.Hack: Engaging Anti-Phishing Training Through a Role-playing Phishing Simulation Game. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 108, 12 pages. https://doi.org/10.1145/3290605.3300338

Rainer Westermann, Kordelia Spies, Günter Stahl, and Friedrich W Hesse. 1996. Relative effectiveness and validity of mood induction procedures: A meta-analysis. *European Journal of Social Psychology* 26, 4 (1996), 557–580.

WHATWG. 2019. URL Living Standard. https://url.spec.whatwg.org.

Colin Whittaker, Brian Ryner, and Marria Nazif. 2010. Large-Scale Automatic Classification of Phishing Pages. In *NDSS '10*. http://www.isoc.org/isoc/conferences/ndss/10/pdf/08.pdf

Wikipedia contributors. 2019a. Monospace (typeface) — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Monospace_(typeface)&oldid=884652181 [Online; accessed 20-February-2020].

Wikipedia contributors. 2019b. Phishing — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=Phishing&oldid=914413259. [Online; accessed 8-September-2019].

Wikipedia contributors. 2019c. URL — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=URL&oldid=909233629 [Online; accessed 11-September-2019].

Wikipedia contributors. 2019d. URL redirection — Wikipedia, The Free Encyclopedia. https://en.wikipedia.org/w/index.php?title=URL_redirection&oldid=916373985. [Online; accessed 19-September-2019].

Min Wu, Robert C. Miller, and Simson L. Garfinkel. 2006. Do Security Toolbars Actually Prevent Phishing Attacks?. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '06)*. ACM, New York, NY, USA, 601–610. https://doi.org/10.1145/1124772.1124863

Jie Xu, Yang Wang, Fang Chen, Ho Choi, Guanzhong Li, Siyuan Chen, and Sazzad Hussain. 2011. Pupillary Response Based Cognitive Workload Index Under Luminance and Emotional Changes. In *CHI '11 Extended Abstracts on Human Factors in Computing Systems (CHI EA '11)*. ACM, New York, NY, USA, 1627–1632. https://doi.org/10.1145/1979742.1979819

Johannes Zagermann, Ulrike Pfeil, and Harald Reiterer. 2016. Measuring Cognitive Load Using Eye Tracking Technology in Visual Computing. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization (BELIV '16)*. ACM, New York, NY, USA, 78–85. https://doi.org/10.1145/2993901.2993908