A Multi-Modal Neuro-Physiological Study of Phishing Detection and Malware Warnings

Ajaya Neupane
Department of Computer and Information Sciences
University of Alabama at Birmingham
aneupane@uab.edu

Md. Lutfor Rahman*
Aegis Foundry LLC
mdlutforrahman@csebuet.org

Nitesh Saxena
Department of Computer and Information Sciences
University of Alabama at Birmingham
saxena@uab.edu

Leanne Hirshfield
Newhouse School of Public Communications
Syracuse University
lmhirshf@syr.edu

ABSTRACT

Detecting phishing attacks (identifying fake vs. real websites) and heeding security warnings represent classical user-centered security tasks subjected to a series of prior investigations. However, our understanding of user behavior underlying these tasks is still not fully mature, motivating further work concentrating at the neurophysiological level governing the human processing of such tasks.

We pursue a comprehensive *three-dimensional* study of phishing detection and malware warnings, focusing not only on what users' task performance is but also on how users process these tasks based on: (1) *neural activity* captured using Electroencephalogram (EEG) cognitive metrics, and (2) *eye gaze patterns* captured using an eye-tracker. Our primary novelty lies in employing *multi-modal neuro-physiological* measures in a single study and providing a *near realistic* set-up (in contrast to a recent neuro-study conducted inside an fMRI scanner). Our work serves to advance, extend and support prior knowledge in several significant ways. Specifically, in the context of phishing detection, we show that users do not spend enough time analyzing key phishing indicators and often fail at detecting these attacks, although they may be mentally engaged in the task and *subconsciously processing real sites differently from fake sites*. In the malware warning tasks, in contrast, we show that users are frequently reading, possibly comprehending, and eventually heeding the message embedded in the warning.

Our study provides an initial foundation for building future mechanisms based on the studied real-time neural and eye gaze features, that can automatically infer a user's "alertness" state, and determine whether or not the user's response should be relied upon.

Categories and Subject Descriptors

K.4.1 [**Computer and Society**]: Public Policy Issues—*Privacy*; D.4.6 [**Operating System**]: Security and Protection—*Authentication*; H.1.2 [**Information Systems**]: User/Machine Systems —*Human Factors; Human Information Processing*

General Terms

Security and privacy, Human-centered computing

Keywords

Phishing Detection; Malware Warnings; EEG; Eye Tracking; Neuroscience

1. INTRODUCTION

Cyber-security is undoubtedly a topic of national importance. While some cyber-attacks exploit the flaws in the system design or implementation itself, others are successful due to the potential negligence or mistakes of end users. This latter aspect of computer systems security, commonly referred to as "user-centered security," forms the central focus of our work. There exists a number of attacks and vulnerabilities underlying user-centered security systems. For example, users are frequently subject to phishing attacks (i.e., presented with malicious websites which may look very similar to real websites), which they may fail to detect, eventually undermining the privacy of their sensitive information. Similarly, warnings are regularly shown to users in order to alert them against potential security risks (e.g., while connecting to a potentially malicious site), which they may not read or comprehend, or may simply ignore.

There exists a large body of recent literature focusing on user-centered security (e.g., [9, 17, 18, 20, 32, 34, 39].), However, our understanding of end user performance in user-controlled security tasks is still not fully mature at this point. In this light, there is a need for a detailed, root-level, *neuro-physiological* investigation of human behavior pertaining to user-centered security.

In this paper, we concentrate on two classical user-centered security tasks: (1) *phishing detection* – distinguishing fake sites from real sites, and (2) *malware warnings* – heeding malware warnings shown by modern browsers when connecting to potentially malicious sites. We pursue a comprehensive *three-dimensional* study of

*Work done while being a student at UAB

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.

CCS'15, October 12–16, 2015, Denver, Colorado, USA.
© 2015 ACM. ISBN 978-1-4503-3832-5/15/10 $15.00.
DOI: http://dx.doi.org/10.1145/2810103.2813660.

not only what users' performance is in these security tasks (*first dimension*: task performance) but also on how users actually process these tasks based on (1) *neural activity* (*second dimension*) captured using Electroencephalogram (EEG) cognitive metrics, and (2) *eye gaze patterns* (*third dimension*) captured using an eye-tracker. An additional dimension we incorporate in our study is a user's individual personality traits measured with simple questionnaire.

Our Contributions: We believe that we make measurable progress towards advancing the science of user-centered security. We are reporting on a first triangular study of users' neural response (EEG), eye focus and dynamics, and task performance, with respect to phishing detection and malware warnings. Our work makes several contributions.

1. We pursue a novel methodology that combines multi-modal neuro-physiological measures in a single study shedding light on multiple facets of human processing of phishing detection and malware warnings tasks. This methodology might be generally applicable to other user-centered security tasks.
2. We employ a neuroimaging technique (EEG) complementary to the one employed in a recent "neuro-only" study (fMRI) [26]. The most notable advantage of using EEG (and a wireless EEG headset) is that the participants can pursue the tasks in a more realistic web browsing scenario. In contrast, the study of [26] was conducted inside a scanner, under a supine posture, and with "constrained" interfaces.
3. Our work advances, extends and supports prior studies in several significant ways (our results summary is below). On many fronts, it also serves to independently re-affirm the findings of previous work and provides further support to the existing knowledge in user-centered security.[1]

Summary of Key Results: Our study provides several interesting insights and results. A detailed listing of our results, positioned with respect to prior results, is provided in Section 8. In the phishing detection task, we found that the users' task accuracy is low, which is mirrored by their gaze activity that concentrated more on the "login region" and/or "company logo region", and less on the "URL region", the key indicator of the authenticity of a website. At the same time, however, users' neural activity shows that they were exhibiting high workload and were highly engaged in making the real-fake decisions (and more engaged than distracted or sleep-prone). In addition, there were some differences, neurologically, in the way they processed the real sites and the fake sites. This three-way result suggests that users may not be fully aware of, and equipped to fully analyze, the main parameter indicative of the legitimacy of a site, but they were certainly making an active effort in this task (i.e., not ignoring it) and *subconsciously processing the real sites differently from the fake sites*. This clearly underscores the importance of continued training and education against phishing attacks, and also suggests the possibility of detecting phishing attacks programmatically based on users' neural patterns.

The way users respond to and process malware warnings seems to be good news all-around. The gaze patterns show that users are reading the warnings, the neural activity shows that users are undergoing high workload (more so when subject to casual news abstracts) and are highly engaged (more engaged than distracted or sleep-prone) when warnings were displayed, and the task accuracy shows that users heed warnings on a large majority of occasions. This may constitute a proof that users are reading, understanding and acting upon malware warnings as stipulated, and emphasizes the continued importance of warnings as an effective means of communicating potential security risks to users in real-time.

Finally, there exists a direct impact of users' "attention control" on their accuracy of phishing detection (the higher the attention control, the higher the accuracy). This suggests that users' susceptibility to phishing attacks is a function of their personality traits (besides their level of awareness).

Implications of Our Work: We believe that our study provides a concrete foundation for building future mechanisms based on real-time neural and eye gaze data, that can automatically detect whether users are in "attentive" or "inattentive" states, i.e., whether or not they are performing the security task as stipulated. Such mechanisms can be developed using machine learning techniques. "Fusing" neural and ocular features may provide a robust detection mechanism (resulting in low error rates).

Another important insight from our study, in the context of phishing detection, is that users' mental activity may be implicitly indicative of whether a given website is real or fake (although users' eventual decision may be incorrect), i.e., users process fake and real sites differently – this suggests that the system could automatically detect a phishing site based on a users' neural activity.

2. BACKGROUND & RELATED WORK

2.1 Overview: EEG and Eye-Tracking

Electroencephalography (EEG) is a non-invasive method of measuring postsynaptic brain activity from the surface of the scalp associated with task-related or internal stimulation. The temporal resolution of EEG is superior to many other methods of brain imaging. While other methods may experience a delay on the order of seconds or minutes (e.g., fMRI – functional magnetic resonance imaging), EEG is able to depict changes within milliseconds. Because of its higher temporal resolution, EEG is often used to evaluate the time course changes in brain activation across different brain regions. This neuroimaging modality is also a good choice as an investigative tool for assessing cognitive states (i.e., cognitive overload and lapses in focused attention) which are not visible to the observer's eye, and may be overlooked or forgotten by the participant in a self-report [11–13, 21]. Many commercial scale EEG monitoring devices exist today. In our study, we use a wireless and lightweight EEG headset (see Section 4).

Eye-tracking is the process of measuring the point of gaze and movement of the eye. The technology has been commonly deployed in many different domains including medical science, marketing research, and psychology to understand users' gaze trail different tasks. Many types of eye-tracking techniques are used today. A popular set of eye trackers uses video captured by a webcam capable of recording infra-red light and mounted on an external display, without the need for any physical contact with the user. In our study, we employ such an eye-tracker (see Section 4).

2.2 Related Work

Task Performance Studies: Closely relevant to the phishing component of our study is study by Dhamija et al. [17]. Their results indicated that users do not perform well at phishing detection and make incorrect choices 40% of the time. Recently, Neupane et

[1] Reproducing the results of prior user-centered security studies in independent settings is believed to be science in itself, constituting an established line of research in premier user-centered security venues, such as SOUPS.

al. [26] obtained very similar results based on an fMRI experiment. Our task performance data also yielded similar results.

The malware warnings fMRI study by Neupane et al. [26], and a field study based on real-world browser telemetry data by Akhawe and Felt [9], both suggest that users heed malware warning messages with a high likelihood. The malware warnings task performance results in our study are consistent with these prior studies.

Many other studies, focusing on SSL warnings and security indicators (e.g., [9, 17, 18, 20, 32, 34]) and measuring the users' task performance, generally suggest that users do not perform well at these security tasks.

Neural Activity and Personality Studies: Neupane et al. [26] conducted the first study of users' neural activity, measured with fMRI, in phishing detection and malware warnings. They showed that users exhibit higher activation in brain regions governing decision-making, attention, and problem-solving (phishing and malware warnings) as well as language comprehension (malware warnings). Our neural results are in line with these findings albeit using a different neuroimaging technique (EEG), and in a much more realistic set-up (outside scanner).

Neupane et al. [26] also showed a negative relationship between brain activation and impulsive personality traits under both phishing and warnings, although such traits did not influence task performance. Our study, in contrast, reveals a direct (positive) impact of attention control on users' task performance in the phishing task.

The study by Vance et al. [36] employed EEG to measure risk-taking behavior in an independent psychological task (Iowa Gambling Task) and predicted users' task performance in the warnings task. It argued that such EEG-based measures could predict warnings task performance. Unlike the fMRI study [26] and our current study, the work of [36] does not directly measure users' neural response in the security tasks themselves. The most recent study by Andersen et al. [14] used fMRI and mouse tracking to argue that polymorphic warnings can reduce the effect of warning habituation.

Eye Gaze Studies: There are also previous studies employing eye-trackers to study whether users look at security indicators [10, 37]. Whalen et al. [37] argued that users do not look at these indicators in general, but did not provide any quantitative results. Arianezhad et al. [10] provided a similar insight in the specific context of "single-sign-on" applications, based on gaze patterns (fixations and durations in areas of interest). Our study, in contrast, focuses on users' gaze patterns (fixations and durations, and movement dynamics) when subject to phishing detection and malware warnings tasks.

3. DESIGN OF EXPERIMENTS

The designs of our phishing detection and malware warnings experiments are in line with the ones previously employed in [17, 26] (phishing) and [26] (warnings). The fMRI experiments [26] had certain limitations, however. Specifically, participants had to lie down inside the scanner in a supine posture, look at low-resolution website images shown on a small screen (640x480) inside the scanner, and provide responses using a primitive button response system (i.e., output-input interfaces were very constrained). Thus, the participants' neural activity and task performance in this set-up might not have reflected their neural activity and task performance in the real-world. In this light, we felt the need for a much more realistic, EEG-based, set-up to measure users' cognitive states and performance, simulating a near real-world browsing experience.

3.1 Real World Browsing Experience

We designed *in-house* software to execute the phishing detection and malware warnings tasks in the Firefox browser (the study was limited to Firefox given its popularity). The participants interacted with websites displayed in the browser very much like a real-world environment. A lightweight wireless EEG headset was used emulating a minimally invasive browsing experience. The eye-tracker was placed directly below computer screen, centered on the screen. Figure 3 provides a snapshot of our experimental set-up.

3.2 Phishing Detection Experiment

Phishing involves stealing a users' private credentials by showing them fake replica of real websites. Fully in line with the design of prior phishing detection studies [17,26], our experiment assumes that the users are explicitly asked to identify fake sites from real sites, and our focus is then to determine users' performance, neural activity and eye gaze activity in making the real-vs-fake decisions. In our experiment, we presented the participants with real and fake versions of popular websites, such as Amazon, eBay, PayPal, Facebook and Citibank. The participants' task was to distinguish between real and fake websites.

Experiment Design and Implementation: Fake websites (denoted "Fake") were created by modifying the URL, logo and layout of the corresponding real websites, or by borrowing the phishing websites from phishtank.com. In order to protect the privacy of participants, while being subjected to real-world phishing sites, we pre-downloaded these sites for offline use and hosted them on our local web-server. The fake websites, which differ from the real websites (denoted "Real") only in the URL, are called "difficult fake (DFake)", assuming they might be difficult to detect. The other fake websites, which differ from real websites in more than one factor, such as layout, logo, fonts and URL, were referred to as "easy fake (EFake)", assuming these might be easier to detect.

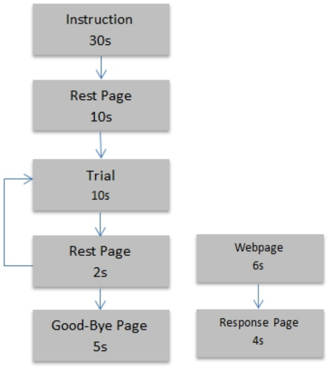


Figure 1: (left) flow chart of entire phishing experiment; (right) components of a trial page

There were 37 randomized trials in this experiment: 13 corresponding to real, and 12 each corresponding to easy fake and difficult fake websites. We set the number of trials consulting the thumb rule of EEG experiments design [24, 38], and previous relevant neuro-physiological studies [14, 26, 36]. Multiple trials are necessary in such experiments to achieve a high signal-to-noise ratio. The experiment started with the Firefox browser loading the instructions page (explaining the terms "real" and "fake", and specifying the tasks participants were to perform), which lasted for 30 seconds. This was followed by the trials pages, each displayed for 10s. Each trial consisted of a webpage (corresponding to a fake/real website) shown for 6s, followed by a 4s long response page. The response page had a dialog box with the question, "Do you think the shown website is real?" and the "Yes" and No" buttons. A rest page of 2s (+ sign shown at the center of a blank page), after each trial was added, during which participants were asked to relax. The experiment ended after 37 trials with the goodbye note, displayed for 5s. The process flow diagram of the experiment is shown in Figure 1.

3.3 Malware Warnings Experiment

Malware is malicious software aimed to obtain unauthorized access to computer resources and collect a users' private information. As a user visits a malicious website, such malware may infect the user's computer. However, modern browsers have devised warning mechanisms to alert the users in case they visit a potentially suspicious web site, relying upon users' input to proceed. Whether or not users read (measured via eye-tracker), understand (measured via EEG cognitive metrics) and heed (measured via task performance) these warnings, are the key questions we are exploring in this work. The EEG cognitive metrics were calculated using the data acquired from three baseline conditions (Section 5.3 provides details). In our experiment, participants were shown the real warnings employed by Firefox (sample shown in Appendix A).

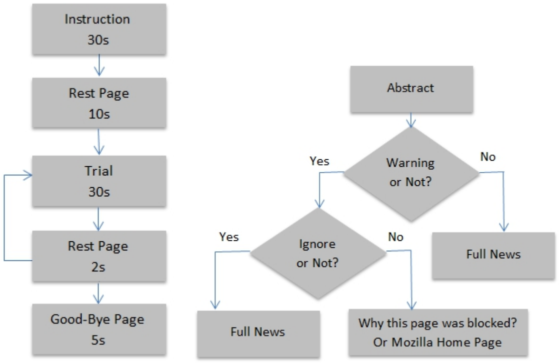


Figure 2: (left) flow chart of entire warnings experiment; (right) flow chart of components of a trial page

Experiment Design and Implementation: We extracted diverse interesting news samples from popular websites, including BBC, NYTimes, Daily Mirror, and CNN, and published them following our own news presentation template. The news samples were divided into two sections, Abstract and Full News. The abstract had a "read more" link which pointed to the corresponding full news. The primary task of the participants was to read the abstract of the news. Some of the news items were randomly intermixed with malware warnings. The warnings were unexpectedly displayed when participants were reading the abstract, or when they clicked on the read more link. Upon ignoring the warning, full news was displayed. The two buttons on the warnings mimicked the ones on real Firefox malware warnings. That is, the "*Get me out of here*" button linked to the home page of Firefox, and the "Why this page was blocked" button linked to the page providing the details as to why the page was blocked.

In this experiment, there were 20 randomized trials: 10 each for the warning and non-warning trials. Similar to the phishing detection experiment, we set the number of trials following the thumb rule of EEG experiments design [24, 38] and previous relevant neuro-physiological studies [14, 26, 36]. The non-warning trials are those in which full-news is shown immediately after the abstract. The experiment started with the instructions page which lasted for 30s, followed by trials, each 30s long. Each trial consisted of an abstract along with the read more link. Similar to the phishing detection experiment, the rest (+ sign) page of 2s, after each trial were added. The experiment ended with the goodbye note shown for 5s. The flow chart of the malware warning experiment is shown in Figure 2.

4. REPEATED MEASURES AND EXPERIMENTAL SET-UP

In our experiments, we recorded each participant's response, response time, neural (EEG) activity, and eye gaze activity while he/she performed the phishing detection and malware warnings tasks. The goal of the experiments was to measure the participants' task performance, cognitive states, and gaze patterns.

Task Performance: We created a custom software-hardware configuration that enabled us to log participants' responses and response times.

Neural Activity and Gaze Patterns: For measuring neural activity, we used a wireless EEG sensor B-Alert headset, X10-Standard, developed by Advanced Brain Monitoring (ABM) [1]. This EEG system, shown in Figure 3 (right), provides a lightweight (less than 3 oz.) means to acquire and analyze 10 channels of high-quality EEG data. The sensors of this EEG headset follow the 10-20 international system of placement. It uses the Fz, F3, F4, C3, Cz, C4, P3, POz, P4 sites to collect EEG data at 256 Hz (Figure 3 (left)). The portable unit worn on the back of the head (Figure 3 (right)) contains miniaturized electronics that amplify, digitize, and transmit the EEG data to the host computer over Bluetooth. The sensors require no scalp preparation; however, water soluble gel had to be applied at electrode sites for better conductance of signals between the skull and sensors. The headset provided a comfortable sensor-scalp interface for 8-12 hours of continuous use.

For measuring gaze patterns, we used EyeTech DS TM3 (remote desk mounted) eye tracker system with a frequency of 60 Hz. The TM3 uses infra-red lights to illuminate the eyes and provide reference points for the eye tracker. Data is captured for both eyes including X/Y gaze coordinates, timestamp, image pixel data, and location and size of pupils within the image.

System Set-Up: Our experimental set-up comprised of a collection of four computers: (1) "Survey Computer" to administer the surveys (described in Section 5.2), (2) "Stimuli Computer" to present experiment tasks (and collect eye-gaze data), (3) "Data Collection Computer" to collect neural data and (4) "Data Visualization Computer" to remotely monitor neural data to ensure its quality. The Stimuli Computer was a laptop with a 15.6 inch screen at a resolution of 1600 x 900.

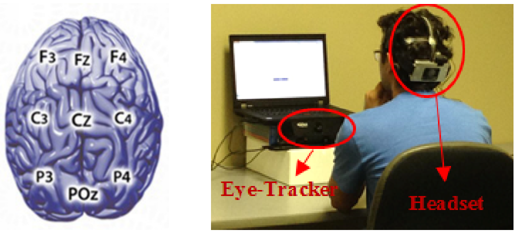


Figure 3: (left) B-Alert electrode arrangement; (right) experimental set-up

5. STUDY PROCEDURES

Our study followed a *within-study design,* i.e., all participants performed the same set of (randomized) trials.

5.1 Ethical and Safety Considerations

The study was approved by our University's IRB. The participation in the study was strictly voluntary. The participants were given the option to withdraw from the study at any point of time. The standard best practices were followed to protect the confidentiality and privacy of participants' data (survey responses, task responses, EEG and eye tracker data) acquired during the study.

5.2 Recruitment and Preparation Phase

The participants were recruited by distributing the study advertisements across our University's campus and on online-media (Facebook & Twitter). Twenty-five healthy participants were recruited for the study. Due to the EEG component of our study, the participants were excluded from the study if they had a history of neurological disorders, anxiety disorder, schizophrenia, and if they were on any psychotropic drugs. Each participant took about a total of 2 hours to complete the study, and was compensated with $40 cash for their time.

Appendix C provides participants' demographics. The majority of our participants were young, students and males. However, our sample was fairly diversified. In particular, *none* of the participants were computer scientists, but rather had diverse backgrounds, such as engineering, education, medical science, physics, and physical health. There were 28% working professionals and non-working people. 35% were above the age of 27, and 36% were females. Future studies might be needed to further validate our results with broader participant samples.

During the preparation phase of the study, informed consent was obtained from each participant. In this phase, we also administered two surveys to our participants to measure their personality traits: (1) *impulsivity* using the Barrat's Impulsivity Scale (BIS) [27], and (2) *attention control* using the Attention Control Scale (ATTC) [16]. BIS is a 30-question-set questionnaire. The higher the BIS score, the higher the impulsivity. ATTC is a 20-question-set questionnaire used to assess executive control of individuals over their attention. The higher the ATTC score, the higher the attention control in an individual. For each of BIS and ATTC, we calculated aggregated scores derived from all of the questions as stipulated in [16, 27].

5.3 Testing (Data Collection) Phase

A measurement of each participant's head was first taken to determine the best size of the B-Alert headset that would fit that participant (our headset came in three sizes: small, medium and large). The EEG headset was then placed on the participant's head, and the participants were moved to Data Collection Computer for an impedance check to ensure the quality of the EEG data.

The participants next completed a 15-minute baseline EEG session that included three 5-minute baseline conditions, namely, standard Eyes-Closed, Eyes-Open, and proprietary 3-Choice Vigilance Task (3C-VT) (developed by ABM [23]). In 3C-VT, participants had to discriminate between one primary and two secondary geometric shapes with stimulus presentation interval of 1.5 to 3s. In the Eyes-Open task, the participants had to respond to visual probe every 2 seconds. In the Eyes-Closed task, they had to respond to audio probe every 2 seconds. These tasks defined the classes of distraction/relaxed wakefulness (DIS), low engagement (LENG), and high engagement (HENG), respectively [23]. The class of sleep onset (SO) is derived using stepwise linear regression using data from these three tasks. The baseline session data is used to create individualized EEG profiles required for the calculation of cognitive state metrics (i.e., SO, DIS, LENG, HENG, referred to as the cognitive states, and Workload) [23].

In our Stimuli Computer, calibration of the eye-tracker was done. Once the participant was ready to perform the experiment, the BAS data acquisition button in the Data Collection Computer, eye tracker gaze points capture module and in house software to execute the tasks were triggered. The phishing detection and malware warnings experiments were executed in random order for different participants. This was done to ensure none of the experiments yield biased results based on the order of their execution.

Impedance, noises and EEG signals were continuously monitored on the Data Visualization Computer to confirm the quality of the data collected. Eye-tracker calibration check was done after each experiment to ensure optimal functioning. A 5-minute gap between the two tasks was provided so participants could rest.

5.4 Post-Test Phase

After completing the security tasks, each participant was asked to fill out a post-test questionnaire (presented on Survey Computer). This questionnaire was designed to determine participants' knowledge of computer security, and to learn how they performed the security tasks they participated in. For example, the participants were asked if they had heard about phishing attacks and malware warnings, and whether they read the warnings and what the warning said. After the post-test questionnaire, the participants were provided with their cash reward.

6. ANALYSIS PROCEDURES & METRICS

6.1 Neural Data

The BAS software included real-time artifact removal for fast and slow eye blinks, muscle movement, and environmental/electrical interference such as spikes and saturations [29]. The data from two of our participants was excluded due to the presence of excessive noise, leaving us with the good quality data from 23 participants. We then used the B-Alert Lab (BAL) Software provided by ABM to conduct the offline data analysis.

We synchronized the EEG data collected during the experiments with the trial presentation time and order. The BAL software then took the synchronized data and the baseline model as its input, and classified each 1-second of EEG data, referred to as an epoch, into one of four cognitive states: *high engagement* (HENG), *low engagement* (LENG), *distraction* (DIS), and *sleep onset* (SO) [29] (see Appendix B for details). For example, for a 6-second time period when a participant was viewing a webpage during the phishing experiment, the BAL software produced 6 mental state values (HENG, LENG, DIS, or SO) for each of the 6 seconds that the participant viewed the webpage. ENG, either HENG or LENG, denotes the state in which users are paying attention to the information they are provided [12,23]. It reflects information-gathering, visual scanning and sustained attention of participants during a given task. DIS is the state when participants shift their attention from the primary task to focus on another activity [12, 23]. SO reflects the state in which people may be able to respond to stimuli but still not able to integrate all information and features [12, 23].

Mental workload (WL) [11, 12] was also calculated for each epoch using the BAL software (Appendix B). WL reflects the amount of neural effort and resources required for a given task. WL increases with increasing working memory load, and under problem-solving, integration of information and analytical reasoning, reflecting brain's executive functioning.

Based on these measures, we computed the average cognitive workload (*WL*) and average percentage of frequency (*pfr*) for which the participants were engaged (*pfrENG*), distracted (*pfrDIS*), and under sleep onset (*pfrSO*), corresponding to different types of trial. WL is calculated on a scale of 0-1; higher values denote higher workloads. Percentage frequency in a trial represents the fraction of the duration for which the participant was in a given mental state (ENG, DIS or SO) in that trial. For example, if someone was highly engaged for 2 epochs, lowly engaged for 2 epochs, distracted for 1 epoch and under sleep onset for 1 epoch, during the 6 second trial, the percentage frequency of engagement will be 4/6 (.67), percentage frequency of distraction will be 1/6 (.17), and percentage frequency of sleep onset will be 1/6 (.17).

6.2 Eye Tracking Data

The gaze data collected during the experiments was used to compute the mean *number of fixations* and mean *gaze duration* of participants in specific areas of the websites, referred as *Areas of Interest* (AOI). *Fixation* is defined as a pause made by a user looking at a specific area to extract meaningful information. We used a dispersion-based technique, *dispersion-threshold algorithm* [30], to compute fixations. This algorithm identified fixations as a group of consecutive points within a particular dispersion and duration threshold [30].

6.3 Statistical Testing

All statistical results in this paper are reported at the significance level (α) of .05. The Friedman test was used to test for the existence of differences within the groups, and, if it succeeded, the Wilcoxon Singed-Rank Test (WSRT) was used to examine in which pairs the differences occurred. The effect size of WSRT was calculated using the formula $r = Z/\sqrt{N}$, where Z is the value of the z-statistic and N is the number of observations on which Z is based. The statistically significant pairwise comparisons are reported with Holm-Bonferroni corrections for multiple testing. Correlations between different conditions were measured using the Spearman's rank correlation coefficient, with Holm-Bonferroni corrections.

7. RESULTS AND ANALYSIS

7.1 Phishing Detection Experiment

To recall, in the phishing detection task, the participants were asked to identify whether a given website is real or fake. We analyzed the neural data, gaze data, and the task performance data collected during the experiment, and studied their interrelationships with one another and with participants' personality scores.

7.1.1 Neural Activity Results

As described in Section 6.1, we computed the average cognitive workload (*WL*) and average percentage of frequency for which the participants were engaged (*pfrENG*), distracted (*pfrDIS*), and under sleep onset (*pfrSO*), for our different trials (Real, Fake, EFake and DFake). The results are shown in Table 1.

From Table 1 (column 1), we see that the average workload exhibited by our participants in identifying the websites as real or fake is high (more than 0.5) for all types of trials. Upon using the Friedman test to test for differences in workload among different types of trials, we did not find a statistically significant difference.

Considering the cognitive state results (columns 2-4), we see that the participants frequency of being engaged was high (at least 50% in all trials except DFake), and their frequency of being distracted or under sleep onset was low (at most 30%). This suggests that the participants were actively engaged, and lowly distracted or sleep-prone, during the experiment.

Metric →		**WL**	**pfrENG**	**pfrDIS**	**pfrSO**
Trials ↓		μ (σ)	μ (σ)	μ (σ)	μ (σ)
Real		.65 (.08)	.61 (.18)	.13 (.12)	.25 (.17)
Fake	Overall	.64 (.08)	.50 (.03)	.20 (.06)	.28 (.06)
	EFake	.64 (.08)	.62 (.17)	.11 (.09)	.25 (.17)
	DFake	.64 (.08)	.39 (.18)	.29 (.14)	.30 (.14)
Overall		.64 (.08)	.54 (.05)	.18 (.06)	.27 (.08)

Table 1: Neural Results for Phishing Detection: Average cognitive workload and average percentage of frequency of engagement, distraction and sleep onset.

We noticed a statistically significant difference in the means of overall pfrENG, overall pfrDIS and overall pfrSO upon testing with the Friedman test (*χ^2(2) = 34.6, p < .0005*). Upon performing pairwise comparisons using WSRT between the means of the three metrics, overall pfrENG, overall pfrDIS and overall pfrSO, statistically significant differences were found between pfrENG and pfrDIS (*p < .0005*) and pfrENG and pfrSO (*p < .0005*), both with a large effect size (*r = .61*). This pattern was visible upon performing pairwise comparisons using WSRT among the means of pfrENG and pfrDIS, and the means of pfrENG and pfrSO, across the Real trials and the EFake trials (all with *p<.0005),* all with a large effect size*(r > .5)*; we did not see a statistically significant difference in pairwise comparison of these metrics across the DFake trials, however. All these pairwise differences remained statistically significant even when applying Holm-Bonferroni correction.

This analysis shows that the participants' frequency of being in an engaged state was higher than their frequency of being in distracted state or sleep-onset state in the phishing task for all types of trials (except DFake). This means that the participants were actively involved in making fake vs. real decisions (not getting distracted by it or ignoring it).

Finally, we contrasted different categories of trials (rows of the Table 1.) with respect to our metrics. We found statistically significant differences in the means of pfrENG, pfrDIS and pfrSO among the different types of trials, with Friedman test (*χ^2(11) = 141.5, p < .001*). Further, upon comparing pfrENG in Real trials with pfrENG in Fake trials with WSRT, we saw a statistically significant difference (*p = .026*) with a medium effect size (*r = .32*). In addition, we found a statistically significant difference in pfrENG for Real trials and pfrENG for DFake trials (*p = .013*), with a medium effect size (*r = .36*). We also found a statistically significant difference in pfrDIS in Fake trials and pfrDIS in Real trials (*p = .031*) with a medium effect size (*r = .31*), and between pfrDIS in DFake trials and pfrDIS in Real trials (*p = .005*) with a medium effect size (*r = .41*). However, we did not see any statistically significant difference in pfrENG and pfrDIS between Real and EFake trials.

This analysis suggests that participants may have been more engaged and less distracted when processing real websites as opposed to fake, or difficult fake, websites. Except of difficult fake pfrDIS and real pfrDIS, these differences do not remain statistically significant upon applying Holm-Bonferroni correction. No statistically significant differences were found in pfrSO across different types of trials.

The last set of results shows that there might be differences in the processing of real and fake websites in human brain. Our participants may have been more engaged and less distracted when dealing with real websites when compared to fake or difficult fake websites. Real websites, because of the fact they were real (although participants did not know about it), might have triggered more engagement or less distraction at a subconscious level. The prior fMRI phishing detection study by Neupane et al. [26] also showed significant differences in activation of specific brain regions while participants were viewing real vs. fake (and difficult fake) websites. As a classical analogy, Huang et al. [22] found an increased activity in certain brain areas while subjects identified real-fake Rembrandt paintings.

7.1.2 Eye Gaze Results

Through the eye-tracking component of our experiment, we wanted to see whether participants look at the key areas of a website and how much time they spent on those areas. The prior studies [17,34, 39] suggest that users may not pay attention to security indicators, based on their low task performance in phishing detection. Our goal was to evaluate this hypothesis based on users' gaze patterns.

We marked the URL, logo and login form of a website as our areas of interest (AOI), since these are some specific locations or artifacts which people may examine while making real-fake decisions. Logo denotes the logo of the website/company, and login form denotes the small form where the user is to input the username and password to login to the site. Both logo and login form are not the true indicators of the legitimacy of a website, given phishers can spoof them relatively easily. For each of our AOIs, we computed mean *number of fixations* (*#fix*) and mean *gaze durations* (*dur*), as described in Section 6.2. The results are shown in Table 2.

AOIs →		**URL**		**Logo**		**Login Form**	
		# fix	**dur (ms)**	**# fix**	**dur (ms)**	**# fix**	**dur (ms)**
Trials ↓		μ (σ)	μ (σ)	μ (σ)	μ (σ)	μ (σ)	μ (σ)
Real		1.02 (1.09)	376 (449)	.56 (.34)	427 (198)	2.8 (1.32)	1370 (514)
Fake	Overall	1.03 (1.01)	345 (349)	1.28 (.66)	705 (310)	3.70 (1.79)	1527 (568)
	EFake	0.98 (1.03)	373 (389)	.99 (.53)	613 (310)	4.00 (.98)	1585 (596)
	DFake	1.09 (1.09)	317 (325)	1.57 (.92)	797 (341)	3.4 (1.83)	1469 (604)
Overall		1.03 (1.01)	355 (378)	1.04 (.52)	612 (258)	3.42 (1.59)	1475 (536)

Table 2: Gaze Results for Phishing Detection: Average number of fixations and average gaze durations in Areas Of Interests

From Table 2, we generally see a lower average number of fixations and average gaze duration at URL compared to logo and login form. The gaze duration seems the highest for the login form. And, this pattern seems to be similar across different types of trials.

Friedman test showed the presence of a statistically significant difference in mean gaze duration among URL, Logo and Login Form ($\chi^2(3) = 32.7,\ p < .0005$) in Real trials and Fake (EFake and DFake) trials. Further, comparing mean gaze durations across different AOIs using WSRT, we saw a statistically significant difference between the overall Logo and overall URL ($p = .004$) with a medium effect size ($r = .43$), overall Login and overall URL ($p < .0005$) with a large effect size ($r = .59$), and overall Login and overall Logo ($p < .0005$) with a large effect size ($r = .61$). Statistically significant differences were also seen between mean gaze durations of Login and URL, and Login and Logo, with respect to all types of trials ($p < .0005$ for all comparisons), with a large effect size ($r > .5$ for all comparisons). We also found a statistically significant difference between mean gaze durations of Logo and URL corresponding to Fake trials (p=*.001*) with a medium effect size ($r = .49$), EFake trials (p=*0.015*) with a medium effect size ($r = .35$), and DFake trials (*p<.0005)* with a large effect size ($r = .57$). All these differences remained statistically significant when applying Holm-Bonferroni correction. No significant difference was found between mean gaze duration in Logo and URL for the Efake trials.

Next, we analyzed the mean number of fixations. This metric also generally follows the same pattern as gaze durations, i.e., it seems participants were fixating the most on the login region, followed by the logo and URL regions. We found a statistically significant difference among the mean number of fixations in URL, Logo and Login Form across Real, and Fake (EFake and DFake) with Friedman test ($\chi^2(14) = 182.7,\ p<.0005$). Further, using WSRT, we found a statistically significant difference between overall Login and overall URL ($p < .0005$) with a large effect size ($r = .57$) and overall Login and overall Logo ($p < .0005$) with a large effect size ($r = .61$), whereas overall Logo and overall URL difference was not statistically significant. The exact same pattern was observed with respect to different types of trials. Statistically significant differences were also observed in the number of fixations in: Login and URL ($p = .001$) with a large effect size ($r = .50$) and Login and Logo ($p = .001$) with a large effect size ($r = .60$) in Real trials, and Login and URL (p<.*00*05) with a large effect size ($r = .57$) and Login and Logo ($p<.0005$) in Fake trials, EFake trials and DFake trials, with a large effect size ($r > .6$ for all comparisons); whereas Logo-URL difference was not statistically significant. All these differences remained statistically significant after Holm-Bonferroni corrections.

Based on the above analysis, we can conclude that participants were fixating more, and spending more time, at the Login and/or Logo regions compared to the URL region, for all categories of trials. This confirms our hypothesis that users may not be spending enough time analyzing the key indicators of phishing attacks. The users were actually looking more at the Login region than the Logo region, which means they may have regarded the login form as a better indicator of the legitimacy of the site than its logo. This insight helps to explain why their real-fake decisions were not accurate, as our task performance results show below.

When testing for differences between different categories of trials (rows of Table 2) with respect to number of fixations and gaze duration using WSRT, no statistically significant differences emerged.

Finally, we performed correlation analysis, using Spearman's correlation method, to elicit relationships in the mean gaze durations across different AOIs. We found a statistically significant positive correlation between mean duration in Login (overall trial) and mean duration in Logo (overall trial) (r_{cor} =*.606, p =.002*), mean duration Login(Fake) and mean duration Logo (Fake) (r_{cor} =*.591, p=.003*), mean duration Login and mean duration Logo (EFake) (r_{cor} =*.569, p=.005*), mean duration Login (Real) and mean duration Logo (r_{cor} = *.551, p=.006*) and between mean duration in Login (DFake trial) and mean duration in Logo (DFake trial) (r_{cor}= *.567, p=.005*). These differences remained statistically significant upon correcting with Holm-Bonferroni correction, and suggest that participants who spent more time at Login also spent more time at Logo overall (and in DFake trials). The other pairs did not show any significant relationship.

7.1.3 Task Performance Results

We calculated the *response times* and the percentage of correctly identified websites out of the total responses given by the participants (referred to as *accuracy*), for different types of trials. The response was counted as correct/incorrect only if the response was provided (6.15% of trials were not responded to and are excluded from our calculations). Table 3 summarizes our results.

The overall accuracy of correctly identifying a website is around 70%. It seems the highest for the real websites and the lowest for the difficult fake websites. Our average accuracy results are in line

Metric → Trial ↓		Accuracy (%) μ (σ)	Response Time (ms) μ (σ)
Real		83.24 (17.28)	1594 (339)
Fake	Overall	62.31 (20.62)	1663 (231)
	EFake	68.35 (21.68)	1667 (268)
	DFake	55.94 (25.30)	1655 (294)
Overall		69.69 (16.64)	1641 (257)

Table 3: Task Performance in Phishing Detection: Average accuracy and response time

with, but slightly better than, the results of [17,26]. They are further supported by our gaze pattern analysis which showed participants were spending more time looking at the login field and/or logo than analyzing the URL.

The Friedman test showed a statistically significant difference in mean accuracies across Real trials, Fake trials, EFake trials and DFake trials ($\chi^2(3) = 32.7, p<.0005$). On further contrasting the accuracy rates across different types of trials with WSRT, we found that participants identified real websites with a statistically significantly higher accuracy than fake websites ($p <.0005$), with a large effect size ($r = .53$). This seems to conform to our neural data analysis, which showed participants were seemingly more engaged, and less distracted, in real trials than they were in fake trials. We also found that the participants identified Real trials with statistically higher accuracy than EFake trials ($p=.003$), with a medium effect size ($r = .44$) and DFake trials ($p < .0005$), with a large effect size ($r = .54$). Further, we found the accuracy for EFake trials to be statistically significantly higher than the accuracy of DFake websites ($p =.017$) with a medium effect size ($r = .34$).

Difficult fake websites had a different URL, disguised to look like the original one, with the layout of the original (real) website. Each easy fake website, in contrast, had a URL and logo or layout different from the corresponding real website. Therefore, it is natural that people were less accurate with difficult fake websites. This difference, however, did not remain statistically significant when using Holm-Bonferroni correction; all others were still statistically significant.

Post-Test Survey Analysis: 52% of our participants reported that they had not heard about phishing attacks. The other 48% defined these attacks as, "*Attacks from unsecured websites and they cause viruses to occur*"; " *someone trying to get your information without you knowing*; *information can be stolen*"; "*Tracks cookies, privacy is reduced*"; "*steal your private information, lose money, ID stolen*". This suggests that participants had some, but not very precise, understanding of phishing, which may help explain the overall low accuracy.

7.1.4 Correlations

Upon using Spearman's correlation, we found a large, statistically significant decrease in the *overall* accuracy of phishing detection with the increase in the *overall* mean gaze duration in the login area of websites (r_{cor}=-.592, $p = .003$). This correlation remained statistically significant even when applying Holm-Bonferroni correction. It suggests that the participants who spent more time looking at the login form had lower accuracy rates. We did not find a statistically significant correlation of accuracy with gaze duration in URL or logo regions.

Spearman's correlation did not reveal significant correlations between task performance & neural metrics, and between neural metrics & gaze metrics.

We next explored correlation between task performance & personality traits. Neupane et al. [26] showed that users' individual personality traits might affect how they process security tasks. They specifically showed that impulsive persons had lower activation in certain decision-making regions of their brains. However, they did not report any direct significant effect of users' personality traits on their task performance [26]. In our experiment, Spearman's correlation revealed a medium, statistically significant, positive relationship between participants' ATTC personality scores and their task accuracy (r_{cor} = .477, p = .021). This correlation remained statistically significant even after applying Holm-Bonferroni correction. BIS did not yield any statistically significant relationship, however. This means that attention control has a positive effect on the performance of the users in the phishing detection task. Training to improve users' attention level, along with education, might therefore help them identify phishing attacks better (we will discuss this aspect in Section 9).

7.2 Malware Warnings Experiment

To recall, in the malware warnings task, participants were randomly exposed to malware warnings while reading abstracts of news items. As in the phishing detection experiment, we analyzed the neural data, gaze data, and task performance data collected during the malware warnings experiment.

7.2.1 Neural Activity Results

We computed average WL, and average pfrENG, pfrDIS and pfrSO, for the three trials – abstract, malware warning and full news. Our results are summarized in Table 4.

Metric → Trial ↓	WL μ (σ)	pfrENG μ (σ)	pfrDIS μ (σ)	pfrSO μ (σ)
Abstract	.65 (.08)	.63 (.17)	.13 (.13)	.22 (.14)
Warning	.69 (.09)	.60 (.21)	.13 (.60)	.25 (.17)
Full News	.67 (.11)	.65 (.20)	.12 (.16)	.22 (.16)

Table 4: Neural Results for Malware Warnings: Average cognitive workload and average percentage frequency of engagement, distraction and sleep onset.

From Table 4 (column 1), we observe that the average workload values across abstract, warnings and full news trials are all high (greater than .65). We also see a higher average workload in processing warnings when compared to abstract and full news. The Friedman test revealed a statistically significant difference among the mean workloads of the abstract, warning and full news trials (χ^2 $(2) = 6.0, p = .048$). Further, upon using WSRT for pair-wise comparisons, we found a statistically significant difference between the warnings and abstracts trials ($p=.005$), with a medium effect size ($r = .41$); the other two pairs of trials did not show up any statistically significant difference. This pairwise difference remained statistically significant even when applying Holm-Bonferroni correction.

This demonstrates that our participants were possibly exerting more effort on their memory and neural resources when subject to warnings in contrast to reading casual abstracts.

Considering the cognitive state metrics (columns 2-4), we see that the participants frequency of being engaged was high (at least 60% in all trials), and their frequency of being distracted or under sleep onset was low (at most 25%). This suggests that the participants were actively engaged, and less distracted or sleep-prone, during the experiment, including warnings. The Friedman test indicated the difference among the means of pfrENG, pfrDIS and pfrSO as significant ($\chi^2(2) = 26.9, p<.0005$). Upon performing pairwise comparisons between the means of the three metrics across the warnings trials using WSRT, statistically significant differences were found between pfrENG and pfrDIS ($p < .0005$) with a large effect size (r = *.59*) , pfrENG with pfrSO (p < *.0005*) with a large effect size (r =*.51*) and pfrDIS with pfrSO (p = *.027*) with a medium effect size (r = *.32*). This pairwise difference, apart from pfrDIS with pfrSO, remained statistically significant even when applying Holm-Bonferroni correction.

This analysis demonstrates that, when processing warnings, participants' frequency of being in an engaged state was higher than their frequency of being in the distracted state or sleep onset state. The high task performance results (discussed later in this section) conform to this high engagement level and high workload.

Last, we contrasted the different categories of trials (rows) with respect to our cognitive state metrics (pfrENG, pfrDIS and pfrSO) using the Friedman test. However, no significant differences emerged.

7.2.2 *Eye Gaze Results*

Our primary goal of employing gaze tracking in the warnings experiment was to determine if the participants actually read the message embedded within the warning, or just ignore it.

To this end, we considered the "red dialog box" of the warning page, called the warning area (see Appendix A), as our AOI, and calculated the average number of fixations (#fix) and average gaze duration (dur) inside it (just like the phishing detection task).

The participants spent almost 2.5s inside the warning area on average (Table 5). This means that participants' primary focus was inside the warning dialog.

Metric → Trial ↓	**#fix** μ (σ)	**dur (ms)** μ (σ)
Warning	7.4 (2.67)	.63 (.17)

Table 5: Gaze Results for Malware Warnings: Average number of fixations and average total gaze duration

To further understand how users processed the warnings while focusing on the warning window, we plotted the fixations in the first warning trial of all participants, in a scatterplot, overlaid on top of the warning (shown in Figure 4). The scatterplot has more fixation points in the middle of the plot, representing dense gaze points inside the warning area. These dense gaze points show that participants spent maximum percentage of their time inside the warning area, spread consistently across the sentences/tabs in the warning message. Furthermore, the fixation points in the scatterplot are flowing along the sentences in the warning area (as shown by sample snapshots in Appendix D), representing the movement of gaze points and demonstrating the "*warning reading flow*". This gaze pattern analysis shows that participants were not just fixating inside the warning window but actually reading the warning message. A similar pattern was observed across other warning trials. We also calculated the correlation between the trial number and the number of fixations (time spent) within the warning area. We only found a statistically significant negative correlation corresponding to two of our participants (*r = -.661, p = .038,; and r = -.72, p = .019*). This suggests that these participants spent lesser time in processing the warnings as the experiment proceeded. On average across all participants, however, no statistically significant correlation was found.

To our knowledge, no prior study has looked at eye gaze analysis in the context of warnings (malware or otherwise). Akhawe and Felt [9] showed that people are likely to heed warnings based on browser telemetry data. Neupane et al. [26], in their fMRI-based malware experiment, showed activity in language comprehension areas of the brain when subjects were exposed to warnings. We presented an analysis of the flow of fixation points over time, and validated that users are in fact reading warnings, which may serve to explain their high heeding rates and comprehension-relevant neural activity [26].

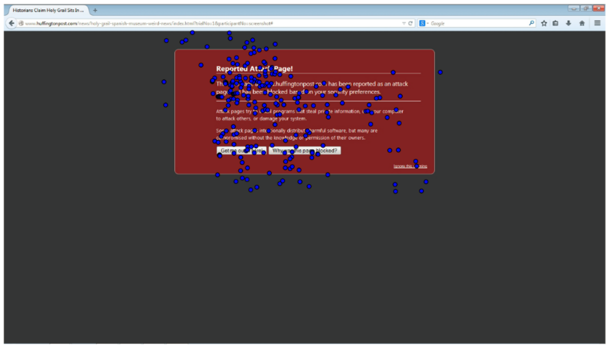

Historians Claim Holy Grail Sits In ...
www.huffingtonpost.com/news/holy-grail-spanish-museum-weird-news/index.html?trialNo=1&participantNo=screenshot#
Google
Reported Attack Page!
Get me out of here!
Why was this page blocked?
Ignore this warning

Figure 4: The flow of fixation points over the 1st Warning trial of all participants (others trials had a similar effect)

7.2.3 *Task Performance Results*

To measure the task performance in the warnings experiment, we recorded the participants' responses (and response time) when they were subject to warnings. We were mainly interested in determining the rate at which the participants may ignore the warning - the fraction of the time they hit the "Ignore this warning" button.

Metric → Trial ↓	**Ignoring Rate (%)** μ (σ)	**Response Time (ms)** μ (σ)
Warning	14.10 (27.79)	2580 (655)

Table 6: Task Performance in Malware Warnings: Average rate of ignoring warnings, and response time

Table 6 summarizes these results. We observe that almost 15% of the time, participants ignored the warnings (i.e., they heeded the warning almost 85% of the time). This result is well-aligned with prior studies [17, 26]. Both these studies suggest that users are highly likely to heed malware warning messages. The high level of workload and engagement reflected in our neural analysis, and the "reading effect" highlighted in our gaze analysis justify participants' heeding behavior.

Post-Test Questionnaire Analysis: Our post-experiment survey results further confirm that our participants read the warnings. 84% of them said they read the warnings. Following are a few excerpts of what information they read in the warnings: "*That the website was a potential threat, if I wanted to continue*"; "*Possible danger on the website*"; "*They asked if I wanted to "Get Me Out of Here!" or figure out why the page I was visiting had been blocked.*" 72% of our participants reported that they had heard about malware attacks.

7.2.4 *Correlations*

Spearman's correlation did not reveal statistically significant correlations between neural metrics, gaze metrics and task performance (heeding rate). It did not reveal statistically significant relationships between personality scores and heeding warning rates.

8. SUMMARY AND KEY INSIGHTS

The primary findings and insights from our study, with respect to our different dimensions, are itemized below. Whenever applicable, our results are positioned with respect the prior results.

Task Performance

- The users fail to identify phishing websites more than 37% of the time. This result is well-aligned with the results of several prior studies (e.g., [17,26]).
- The users are likely to heed warnings about 85% of the time. This serves to further validate the results of two recent studies [9,26].

Neural Activity

- The users' exhibit a high cognitive load in the two security tasks. The cognitive load in processing warnings is more than the cognitive load in processing casual abstracts of news articles. Moreover, users' frequency of being in an "engaged" state is more than their frequency of being in "distracted" and "sleep-prone" states for both tasks. This means that users are paying attention and making an active effort while performing these tasks (and not ignoring them). Although this level of involvement translated into high heeding rates in the warnings task, the phishing detection task accuracy is still quite low (as listed above). This result is in line with the findings presented in [26], but is based on a *complementary neuroimaging technique* having high temporal resolution (EEG vs. fMRI) and accomplished under a near ecologically valid setting (e.g., out-scanner vs. in-scanner, sitting vs. supine).
- At a *subconscious* level, there might be hidden differences in how users detect real and fake websites (in line with identifying real and fake paintings [22]). Real websites, which possibly simulate a more trustworthy environment, may have a higher frequency of engagement, and a lower frequency of distraction, compared to fake websites. This means that the computer system could use these subtle *implicit cues* to determine whether the site is fake (even though users may eventually fail to detect it).

Eye Gaze Patterns

- Eye gaze analysis in the phishing detection task shows that users do not spend enough time looking at the key areas of websites (less time on URL; more time on "login field" or "website logo") for identifying its trustworthiness. A prior work [37] made a similar conclusion regarding security indicators in general, but did not provide any quantitative results. The work of [10] provided a similar insight but in the context of "single-sign-on" applications, not phishing detection.
- The correlation of gaze "fixations" with phishing detection accuracy shows that users who look longer at the login field are likely to have lower accuracy. Also, users who look longer at the login field are more likely to look longer at the website logo (not an authentic indicator of the real website).
- Gaze pattern analysis of malware warnings shows that users are fixating inside the warning dialog and actually *reading the warnings* (also reflected in their high task performance). This is the first work that shows the warning "reading effect". Prior work [17,26] showed that users heed warnings (based on task performance data) and may trigger "language comprehension" activity in their brains. Overall, our work corroborates the previous findings demonstrating that users (1) read (based on eye gaze analysis), (2) understand (based on neural activity) and (3) heed (based on task performance) warnings on a large majority of occasions.

Personality Traits

- The difference in users' personal characteristics can have an effect on how well they perform in a security task. A user with high attention control (measured via a simple questionnaire [16]) is more likely to identify the real and fake websites correctly. Our study demonstrates a direct impact of personality traits on security task performance. The work of [26] showed a correlation between personality traits with neural activity, not task performance. Beyond raising people's awareness to phishing attacks, interventional training programs that can improve people's attention control [4, 5, 15,35] may therefore help reduce the impact of these attacks.

9. DISCUSSION

Implications of Our Work: A broader implication of our work is in leveraging real-time brain monitoring and eye tracking techniques to inform the design of user-centered security systems. The current user-centered security practices unconditionally rely upon users' input whether or not users pay attention. The use of real-time "brain-eye" measures, we investigated in this paper, could be used to build an automated mechanism where the system can determine whether user's response is reliable or not. For example, if the gaze patterns show that the user did not sufficiently look at the URL when connecting to a website, or did not read the message provided by a warning, the user's response would most likely not be reliable. Similarly, if neural features show that the user was not engaged, or was under a distracted state, when subject to a security task, the user's response may not be valid. In contrast, if eye gaze dynamics show that the user reads the warning and neural activity reveals that the user was highly engaged, a user's response can be deemed legitimate.

To formalize a bit, we are suggesting a mechanism based on real-time neural and eye gaze data, that can detect whether users are in an "attentive" or "inattentive" state, i.e., whether or not they are performing the security task as stipulated. Such mechanisms can be developed using machine learning techniques. "Fusing" neural and ocular features may provide a robust detection mechanism (resulting in low error rates).

While traditional security approaches either rely on machines alone or humans alone, what we are proposing is a hybrid approach where machines and humans work in conjunction with each other, possibly complementing each other's strengths and weaknesses in meaningful ways. This approach could be generally applicable to many security applications including other warnings (e.g., SSL warnings [34] or app permissions warnings [19]), user-aided device pairing [31], security and privacy indicators (e.g., webcam lights [28]), and more.

Although the design and evaluation of such a mechanism requires a comprehensive future investigation, we believe that our work lays out the necessary foundation at least in the realm of phishing detection and malware warnings. Given the rise of eye-tracking and neuroimaging devices in the commercial sectors, such as the adoption of eye-trackers in smart-glasses [7, 8] and gaming BCI headsets [2, 6] it seems feasible that such a mechanism could be used in practice once shown effective, especially in high-security settings (such as defense applications).

A malicious application having access to brain-eye measures could similarly be used for offensive purposes. For example, a user could be attacked at an opportune moment, i.e., when he/she is in the inattentive state as inferred by eye-brain features (e.g., when the user is sleep-prone or otherwise distracted). Commercial BCI devices have already been shown vulnerable to privacy attacks where a malicious app can infer sensitive user information (e.g., their PIN digits) based on recorded brain signals [25]. A similar attack model seems applicable to eye trackers. The attack vector we are envisioning aims to infer a user's neuro-physiological state so as to optimize the timing of the occurrence of the attack.

Strengths and Limitations: We believe that our study has several strengths. The neuro-physiological sensors chosen for our study – a lightweight and wireless EEG headset, and a remote desk mounted eye-tracker – allowed us to collect data almost transparently just like in day-to-day computer use. Also, we simulated a near real-world web browsing experience where participants interacted with a popular browser and actual websites.

Similar to any other study involving human subjects, our study also had certain limitations. The study was conducted in a lab, which might have impacted the performance of participants since they might not have felt the real security risks, similar to other prior lab studies. Due to the neuro-physiological focus of the study, it does not currently seem feasible to conduct such a study online or in field conditions. Although our EEG headset was very lightweight, performing the tasks with the headset on might have affected the experience of some participants.

Our participant sample comprised of a majority of young students. This represents a common constraint underlying many University lab studies, especially those involving neuro-physiological scanning due to logistical challenges (e.g., costly equipment, rigorous exclusion criteria and lengthy protocols), such as the recent fMRI study [26] (*N=25*; mostly students), the eye-tracking study [37] (*N=16*; students, faculty and staff) and the eye-tracking study [10] (*N=19*; mostly youth). However, our sample exhibited some diversity with respect to educational backgrounds (especially, no participant had a computer science background). Moreover, our sample, especially in terms of age, was closer to the group of users who use Internet frequently [3] and who are supposedly more vulnerable to phishing attacks [33]. Also, our result demonstrating subconscious differences in brain activation while processing real and fake websites may persist despite age differences. Indeed, we saw subconscious differences in the participants belonging to each of the 19-22 age group and 30 plus age group. Future studies might be needed to further validate our results with broader participant samples.

One limitation of our study pertains to the number of trials presented to the participants. Although multiple trials is a norm in EEG (and neuro-imaging) experimental design [24,38] to achieve a good quality signal-to-noise ratio, the participants may hardly face many security-related trials in a short span of time in real life. Our behavioral results, in the malware warnings experiment, are still well-aligned with a previous large-scale real life study [9].

Another limitation relates to the participants' motivation to disregard the warning. The reward for ignoring the warning might not have been high enough for our study participants, since they could only read full news on disregarding the warning. On the other hand, since the participants were performing the experiments on a lab computer, they could have ignored the warning more often when compared to using their own laptop in real-world. This suggests that the warning heeding rates may be higher than 85% in a real-life scenario, which is also reflected in the field study of [9].

Finally, in the phishing detection task, our participants were explicitly asked to identify a website as real or fake. However, in a real-world attack, the victims are driven to a phishing website from some primary task (e.g., reading email) and the decision about the legitimacy of the site needs to be made implicitly. Nevertheless, in any case, the users ultimately have to make the decision about the legitimacy of the site. Our results show that, despite being asked explicitly, users are not able to detect the legitimacy of the websites accurately, and therefore the result may be even worse in a real world attack where the decisions are to be made implicitly. The subconscious differences in real-fake processing, due to their implicit nature, may still persist.

10. CONCLUSIONS AND FUTURE WORK

We pursued a triangular study of phishing detection and malware warnings, measuring users' neural activity, eye gaze patterns, task performance, and inter-relationships thereof. In the realm of phishing detection, our results showed that users do not spend enough time looking at key phishing indicators and often fail at detecting these attacks, although they may be highly engaged in the task and subconsciously processing real sites differently than fake sites. In the malware warning tasks, on the other hand, our results demonstrated that users frequently read and eventually heed the message embedded within the warning. We also found that a user's personality traits (specifically, attention control) directly impact his/her phishing detection accuracy. This suggests that users may detect phishing attacks better if they could be trained to exercise attention control (beyond phishing awareness training). Further work is necessary to understand the effect of such interventional training on users' performance in the phishing detection task.

Based on our work, we suggested the possibility of building future automated mechanisms applying a fusion of real-time neural and eye gaze features that can infer users' "alertness" state, and determine whether or not users' responses should be relied upon. The proposed mechanism may be used to "sanitize" a user's response and enhance the credibility of human decisions in a user-centered security system, serving as a closed-loop between humans and machines. Future research is needed to design and validate such mechanisms in different security domains.

Acknowledgments

The authors thank: Cali Fidopiastis and Pankaj Mahajan for their help with data collection, Rajesh Kana for his feedback in the early stages of this work, John Sloan for feedback on a previous draft, and the ABM team, especially Stephanie Korszen, for valuable guidance on EEG data analysis. We also thank Konstantin (Kosta) Beznosov (our shepherd) and CCS'15 anonymous reviewers for their constructive comments and guidance.

11. REFERENCES

[1] B-Alert X-10 Set-Up Manual. http://www.biopac.com/Manuals/b-alert%20x10%20setup.pdf.

[2] Emotiv EEG Headset. http://emotiv.com/.

[3] Internet Users Demographics. http://www.pewinternet.org/data-trend/internet-use/latest-stats/. [Online; accessed 30-July-2015].

[4] Lumosity. www.lumosity.com.

[5] MindAscend . www.mindascend.com.

[6] Neurosky . http://neurosky.com/.

[7] SMI Eye-Tracking Glasses . http://eyetracking-glasses.com/.

[8] Tobi Gaze Glass. http://www.tobii.com/en/eye-tracking-research/global/landingpages/tobii-glasses-2/.

[9] Devdatta Akhawe and Adrienne Porter Felt. Alice in warningland: A large-scale field study of browser security warning effectiveness. In *Presented as part of the 22nd*

USENIX Security Symposium (USENIX Security 13), pages 257–272, Washington, D.C., 2013. USENIX.

[10] M. Arianezhad, L. J. Camp, T. Kelley, and D. Stebila. Comparative eye tracking of experts and novices in web single sign-on. In *Proceedings of the Third ACM Conference on Data and Application Security and Privacy*, CODASPY '13, pages 105–116. ACM, 2013.

[11] C. Berka, D. J. Levendowski, M. M. Cvetinovic, M. M. Petrovic, G. Davis, M. N. Lumicao, V. T. Zivkovic, M. V. Popovic, and R. Olmstead. Real-time analysis of eeg indexes of alertness, cognition, and memory acquired with a wireless eeg headset. *International Journal of Human-Computer Interaction*, 17(2):151–170, 2004.

[12] C. Berka, D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. T. Zivkovic, R. E. Olmstead, P. D. Tremoulet, and P. L. Craven. Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks. *Aviation, space, and environmental medicine*, 78(Supplement 1):B231–B244, 2007.

[13] C. Berka, D. J. Levendowski, C. K. Ramsey, G. Davis, M. N. Lumicao, K. Stanney, L. Reeves, S. H. Regli, P. D. Tremoulet, and K. Stibler. Evaluation of an eeg workload model in an aegis simulation environment. In *Defense and security*, pages 90–99. International Society for Optics and Photonics, 2005.

[14] Bonnie Brinton Anderson and C. Brock Kirwan and Jeffrey L. Jenkins and David Eargle and Seth Howard and Anthony Vance. How polymorphic warnings reduce habituation in the brain: Insights from an fMRI study. In *ACM Conference on Human Factors in Computing Systems, CHI*, pages 2883–2892, 2015.

[15] R. Chambers, B. C. Y. Lo, and N. B. Allen. The impact of intensive mindfulness training on attentional control, cognitive style, and affect. *Cognitive Therapy and Research*, 32(3):303–322, 2008.

[16] D. Derryberry and M. A. Reed. Anxiety-related attentional biases and their regulation by attentional control. *Journal of abnormal psychology*, 111(2):225, 2002.

[17] R. Dhamija, J. D. Tygar, and M. Hearst. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 581–590. ACM, 2006.

[18] S. Egelman, L. F. Cranor, and J. Hong. You've been warned: an empirical study of the effectiveness of web browser phishing warnings. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1065–1074. ACM, 2008.

[19] A. P. Felt, E. Ha, S. Egelman, A. Haney, E. Chin, and D. Wagner. Android permissions: User attention, comprehension, and behavior. In *Proceedings of the Eighth Symposium on Usable Privacy and Security*, page 3. ACM, 2012.

[20] B. Friedman, D. Hurley, D. C. Howe, E. Felten, and H. Nissenbaum. Users' conceptions of web security: A comparative study. In *CHI'02 extended abstracts on Human factors in computing systems*, pages 746–747. ACM, 2002.

[21] F. C. Galán and C. R. Beal. Eeg estimates of engagement and cognitive workload predict math problem solving outcomes. In *User Modeling, Adaptation, and Personalization*, pages 51–62. Springer, 2012.

[22] M. Huang, H. Bridge, M. J. Kemp, and A. J. Parker. Human cortical activity evoked by the assignment of authenticity when viewing works of art. *Frontiers in human neuroscience*, 5, 2011.

[23] R. R. Johnson, D. P. Popovic, R. E. Olmstead, M. Stikic, D. J. Levendowski, and C. Berka. Drowsiness/alertness algorithm development and validation using synchronized eeg and cognitive performance to individualize a generalized model. *Biological psychology*, 87(2):241–250, 2011.

[24] S. J. Luck. Ten simple rules for designing erp experiments. *Event-related potentials: A methods handbook*, 262083337, 2005.

[25] I. Martinovic, D. Davies, M. Frank, D. Perito, T. Ros, and D. Song. On the feasibility of side-channel attacks with brain-computer interfaces. In *USENIX Security Symposium*, pages 143–158, 2012.

[26] A. Neupane, N. Saxena, K. Kuruvilla, M. Georgescu, and R. Kana. Neural signatures of user-centered security: An fMRI study of phishing, and malware warnings. In *Proceedings of the Network and Distributed System Security Symposium (NDSS)*, pages 1–16, 2014.

[27] J. H. Patton, M. S. Stanford, and E. S. Barratt. Factor structure of the Barratt impulsiveness scale. *Journal of clinical psychology*, (51):768–74, 1995.

[28] R. S. Portnoff, L. N. Lee, S. Egelman, P. Mishra, D. Leung, and D. Wagner. Somebody's Watching Me? In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, 2015.

[29] M. Poythress, C. Russell, S. Siegel, P. Tremoulet, P. Craven, C. Berka, D. Levendowski, D. Chang, A. Baskin, R. Champney, et al. Correlation between expected workload and eeg indices of cognitive workload and task engagement. 2006.

[30] D. D. Salvucci and J. H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*, pages 71–78. ACM, 2000.

[31] N. Saxena, J.-E. Ekberg, K. Kostiainen, and N. Asokan. Secure device pairing based on a visual channel. In *Security and Privacy, 2006 IEEE Symposium on*, pages 6–pp. IEEE, 2006.

[32] S. E. Schechter, R. Dhamija, A. Ozment, and I. Fischer. The emperor's new security indicators. In *Security and Privacy, 2007. SP'07. IEEE Symposium on*, pages 51–65. IEEE, 2007.

[33] S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. 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*, pages 373–382. ACM, 2010.

[34] J. Sunshine, S. Egelman, H. Almuhimedi, N. Atri, and L. F. Cranor. Crying wolf: An empirical study of ssl warning effectiveness. In *USENIX Security Symposium*, pages 399–416, 2009.

[35] Y.-Y. Tang, Y. Ma, J. Wang, Y. Fan, S. Feng, Q. Lu, Q. Yu, D. Sui, M. K. Rothbart, M. Fan, et al. Short-term meditation training improves attention and self-regulation. *Proceedings of the National Academy of Sciences*, 104(43):17152–17156, 2007.

[36] A. Vance, B. B. Anderson, C. B. Kirwan, and D. Eargle. Using measures of risk perception to predict information security behavior: Insights from electroencephalography (eeg). *Journal of the Association for Information Systems*, 15(10):679–722, 2014.

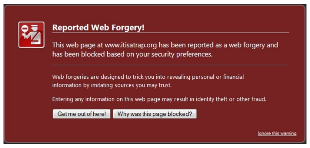
[37] T. Whalen and K. M. Inkpen. Gathering evidence: use of visual security cues in web browsers. In *Proceedings of Graphics Interface 2005*, pages 137–144. Canadian Human-Computer Communications Society, 2005.

[38] G. F. Woodman. A brief introduction to the use of event-related potentials in studies of perception and attention. *Attention, Perception, & Psychophysics*, 72(8):2031–2046, 2010.

[39] M. Wu, R. C. Miller, and S. L. Garfinkel. Do security toolbars actually prevent phishing attacks? In *Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 601–610. ACM, 2006.

APPENDIX

A. SAMPLE WARNING



Reported Web Forgery!

This web page at www.itisatrap.org has been reported as a web forgery and has been blocked based on your security preferences.

Web forgeries are designed to trick you into revealing personal or financial information by imitating sources you may trust.

Entering any information on this web page may result in identity theft or other fraud.

B. NEURAL METRICS

The B-Alert headset measures and records the electrical activity in the brain with sensors placed on the scalp. These signals are first decontaminated from any noises, e.g., the presence of eye-blinks, spikes, and muscle movement. The B-Alert cognitive metrics were obtained from ABM's four-class B-Alert quadratic discriminant function classification algorithm (see Berka et al. [29] and B-Alert User Manual for further details) for each second of data, referred to as epoch. It gives the mean probability of classifications for the four classes: high engagement, low engagement, distraction and sleep in a given epoch. The class with the greatest mean probability is the winning class. This is the class assigned to the epoch. For example, if an epoch is classified as high engagement with probability .45, low engagement as .30, distraction as .20 and sleep onset as 0.05, then the final class of the epoch will be high engagement. The workload was derived from a two-class Linear discriminant function algorithm (range 0.0 to 1.0) [11, 12, 29].

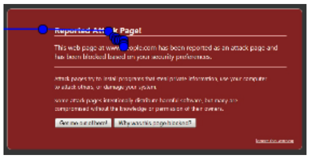
C. PARTICIPANT DEMOGRAPHICS

Participant Size (N = 25)	
Gender (%)	
Male	64
Female	36
Age (%)	
19-22 years	44
23-26 years	20
27-30 years	16
31-34 years	8
>35 years	12
Background (%)	
Students (undergrads and grads from different fields)	72
Working Professionals	16
Others	12

Table 7: Participant Demographics Distribution Summary

D. WARNING READING EFFECT

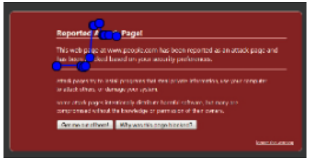
The sample frames from one of the participants, shown in Figures D.1–D.8, demonstrate the warning message reading effect (most other participants had a similar effect)



Reported Attack Page!

This web page at www.people.com has been reported as an attack page and has been blocked based on your security preferences.

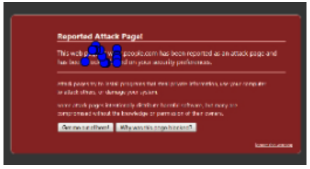
Figure D.1: Gaze plot Frame 1



Reported Attack Page!

This web page at www.people.com has been reported as an attack page and has been blocked based on your security preferences.

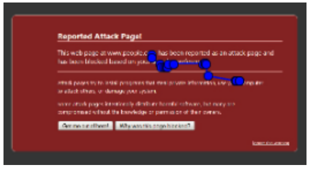
Figure D.2: Gaze plot Frame 2



Reported Attack Page!

This web page at www.people.com has been reported as an attack page and has been blocked based on your security preferences.

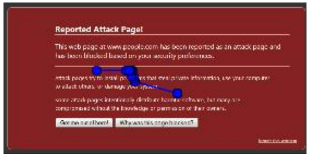
Figure D.3: Gaze plot Frame 3.



Reported Attack Page!

This web page at www.people.com has been reported as an attack page and has been blocked based on your security preferences.

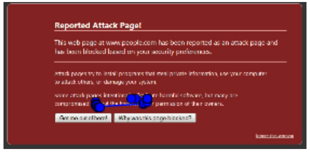
Figure D.4: Gaze plot Frame 4



Reported Attack Page!

This web page at www.people.com has been reported as an attack page and has been blocked based on your security preferences.

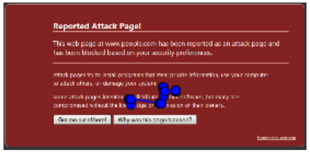
Figure D.5: Gaze plot Frame 5



Reported Attack Page!

This web page at www.people.com has been reported as an attack page and has been blocked based on your security preferences.

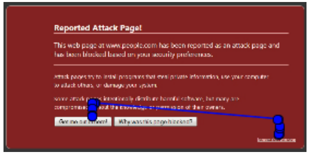
Figure D.6: Gaze plot Frame 6



Reported Attack Page!

This web page at www.people.com has been reported as an attack page and has been blocked based on your security preferences.

Figure D.7: Gaze plot Frame 7



Reported Attack Page!

This web page at www.people.com has been reported as an attack page and has been blocked based on your security preferences.

Figure D.8: Gaze plot Frame 8