Neural Signatures of User-Centered Security: An fMRI Study of Phishing, and Malware Warnings

Ajaya Neupane[1], Nitesh Saxena[1], Keya Kuruvilla[2], Michael Georgescu[1], and Rajesh Kana[2]
[1] Department of Computer and Information Sciences
University of Alabama at Birmingham
{aneupane, saxena, mgeorges}@uab.edu
[2] Department of Psychology
University of Alabama at Birmingham
{rkana, keyak}@uab.edu

Abstract— **The security of computer systems often relies upon decisions and actions of end users. In this paper, we set out to investigate user-centered security by concentrating at the most fundamental component governing user behavior – the human brain. We introduce a novel neuroscience-based study methodology to inform the design of user-centered security systems. Specifically, we report on an fMRI study measuring users' security performance and the underlying neural activity with respect to two critical security tasks: (1) distinguishing between a legitimate and a phishing website, and (2) heeding security (malware) warnings. At a higher level, we identify neural markers that might be controlling users' performance in these tasks, and establish relationships between brain activity and behavioral performance as well as between users' personality traits and security behavior.**

Our results provide a largely positive perspective towards users' capability and performance vis-à-vis these crucial security tasks. *First*, we show that users exhibit significant brain activity in key regions associated with decision-making, attention, and problem-solving (phishing and malware warnings) as well as language comprehension and reading (malware warnings), which means that users are actively engaged in these security tasks. *Second*, we demonstrate that certain individual traits, such as impulsivity measured via an established questionnaire, can have a significant negative effect on brain activation in these tasks. *Third*, we discover a high degree of correlation in brain activity (in decision-making regions) across phishing detection and malware warnings tasks, which implies that users' behavior in one task may potentially be predicted by their behavior in the other task. *Finally*, we discuss the broader impacts and implications of our work on the field of user-centered security, including the domain of security education, targeted security training, and security screening.

I. INTRODUCTION

Computing has become increasingly common in many spheres of users' daily lives. At the same time, the need for securing computer systems has become paramount. To enable secure on-line interactions, actions performed and decisions made by human users need to be factored into system design – a principle sometimes referred to as "human in the loop" [9]. Two such prominent *user-centered security* tasks are: (1) distinguishing between a legitimate and a fake web-site (*phishing detection task*), and (2) heeding warnings provided by modern browsers when connecting to potentially malicious web-sites (*malware warnings task*). User attitudes, perceptions, acceptance and use of information technology have been long-standing issues since the early days of computing. This is especially true in secure computing since user behavior can directly or indirectly impact the security of the system. In this light, it is important to understand users' behavior in executing security tasks and their potential susceptibility to attacks.

The field of user-centered security has received considerable attention recently but is still in its infancy. As such, our understanding of end user performance in real-world security tasks is not very precise or clear at this point. A number of computer lab-based studies focusing on security warnings and security indicators (e.g., [10, 12, 13, 14, 15, 16, 17]) came to the conclusion that users hardly perform well at these tasks and often ignore them. This general wisdom in this area has been called into question by a recent large-scale field study of modern browsers' phishing, SSL and malware warnings [11], which showed that users actually heed these warnings with high likelihood.

In this paper, we set out to enhance the current knowledge in, and address fundamental questions pertaining to, user-centered security from a *neuropsychological* standpoint. The primary questions driving our research include: (1) whether or not users actively engage in security tasks; (2) do users ignore or bypass these tasks; (3) what brain regions get activated while performing these tasks; (4) how well users perform at these tasks; (5) whether certain personality traits influence users' security behavior and performance; and (6) is users' behavior in one task related to their behavior in another task.

In an attempt to answer these inquiries, we introduce a novel methodology for studying user-centered security – one that involves *neuroimaging*. By means of this general methodology, our overarching goal is to delineate the nature of cognitive and neural processes that underlie user-centered security decisions and actions. This specific goal in our work reported in this paper is achieved via fMRI (functional

NDSS '14 23-26 February 2014, San Diego, CA, USA
 ISBN I-891562-35-5
http://dx.doi.org/10.14722/ndss.2014.23056

Magnetic Resonance Imaging) scanning. fMRI provides a unique opportunity to examine the brain responses, in-vivo, mediating user decisions during human-computer security interactions. As a first line of investigation into our novel methodology, this fMRI study will shed light on end users' behavior and performance with respect to the important tasks of *phishing detection* and *malware warnings*.

Contributions & Results Summary: Our main contributions in this paper are summarized as follows:

1. Novel Methodology to Study User-Centered Security: We propose a new generalized methodology for studying neurological patterns governing users' performance and behavior with respect to user-centered security tasks.

2. fMRI Study of Phishing, and Malware Warnings: As a specific use case of our methodology, we design and develop in-scanner fMRI experiments for phishing detection and malware warnings tasks (*Section III*), and conduct a user study by recruiting and scanning 25 individuals performing these tasks. (*Section IV*)

3. Comprehensive Neural and Behavioral Analysis: We provide a comprehensive analysis of neuroimaging and behavioral data, not only evaluating the phishing and malware warnings experiments independently but also contrasting them with each other. (*Section V-VII)*

The results of our study provide a largely positive perspective towards users' capability and performance with respect to phishing detection and malware warnings tasks. *First*, we show that users exhibit significant brain activity in key regions associated with decision-making, attention, and problem-solving (phishing and malware warnings) as well as language comprehension and reading (malware warnings), which means that users are actively engaged in these tasks. In case of malware warnings, this level of brain activation matched with users' good task performance reflected by the behavioral data (confirming the findings reported in [11]). In case of the phishing task, however, the behavioral performance was poor despite significant activation in brain regions correlated with higher order cognitive processing. *Second*, we demonstrate that certain personality traits, specifically impulsivity measured via a simple questionnaire [1], can have a significant negative effect on brain activation in these tasks. In other words, impulsive individuals showed lower brain activation and may thus have poor task performance. *Third*, we discover a high degree of correlation in brain activity (with respect to decision-making regions) across phishing detection and malware warnings tasks, which implies that users' behavior in one task may potentially be predicted by their behavior in the other task. *Finally*, we discuss the broader impact and implications of our work to the field of user-centered security, including the domain of security education, targeted security training, and security screening.

II. Background And Related Work

In this section, we provide the background necessary to understand our experiments and study design, and discuss the ethical aspects and prior work relevant to our research.

A. *fMRI Overview*

fMRI (Functional MRI) is a Blood Oxygen Level Dependent function measure, and is derived from a combination of stimulus-induced changes in the local cerebral blood flow, local blood volume, and local oxygen consumption rate [5]. It is assumed that such changes are associated with changes in neuronal activity [6] and thus fMRI provides an indirect measure of the underlying neuronal activity. In contrast to other brain scanning approaches, such as EEG, fMRI has a much better spatial resolution. In an fMRI scan, human participants lie down in the MRI scanner and perform cognitive tasks while their brain activity is being measured. In this way, we can time-lock the participant's brain activity to a certain cognitive event. fMRI is an appealing platform to conduct small-scale studies providing high spatial resolution. In sum, fMRI measures brain activity by detecting related changes in blood flow.

B. *Our Experimental Set-Up*

Throughout the project, the fMRI data was acquired using the 3T Siemens Allegra Scanner available to us at Civitan International Research Center at the University of Alabama in Birmingham, our University (see Figure 1) depicting our scanner and the experimental set-up). All fMRI tasks followed the same data acquisition protocol as follows. For functional imaging, we used a single-shot gradient-recalled echo-planar pulse sequence that offers the advantage of rapid image acquisition (Repetition Time = 1000 ms, Echo Time = 30 ms, flip angle = 60 degrees, Field of View = 24 cm, matrix = 64 x 64). This sequence covers most of the cortex (seventeen 5-mm thick slices with a 1 mm gap) in a single cycle of scanning (1 TR) with an in-plane resolution of 3.75 x 3.75 x 5 mm^3.

Fig. 1 A pilot subject being prepared for the scan

C. *Ethical and Safety Considerations*

Our study was approved by the Institutional Review Board (IRB) at our University. Care was taken to maximize the safety of the participants while being scanned by following standard practices. Their participation in the study was strictly voluntary. The participants signed an informed consent form prior to the study and were given the option to withdraw from the study at any point of time. Best practices were followed to protect the confidentiality and privacy of participants' data acquired during the study by de-identifying the collected data.

D. *Study Limitations, and Sample Size*

In line with any other study involving human subjects, our study also had certain limitations. A primary limitation pertains to the constraints posed by the fMRI experimental set up. Since the participants were performing the tasks inside the

fMRI scanner, the set up did not mimic real-world online browsing experience. The discomfort associated with lying down in a supine position and being stationary may have also impacted participants' brain activity. In addition, just the fact that the participants were being scanned, may have impacted their brain activation and behavioral responses. The constrained interface (image-based display, binary input and no internet connectivity, unlike a modern computer) available during the scans may have limited participants' interactions with the system. For example, the participants were presented with the images of the websites (rather than the websites themselves) in the phishing task. Similarly, the malware warning images that could be shown on the displays were very simplistic and rudimentary. We believe that this may have negatively affected participants' performance in the underlying security tasks. Furthermore, participants' head motion in the MRI scanner, although we have corrected for it (Section V.A), may have impacted the fMRI data quality. Finally, the lab-based environment of the study may have impacted participants' behavior as they may not have felt real security risks during the experiments.

The effective sample size used in our study ranged from 22 (phishing detection task) to 25 (malware warnings task) participants (see Section V.A), which previous power analysis studies have found to be optimal. For instance, statistical power analysis of event-related design fMRI studies has demonstrated that 80% of clusters of activation proved reproducible with a sample size of 20 subjects [56]. Another study [55] found that a sample size of 24 gave an accurate activation map with a sufficient level of power (i.e., an 80% true positive rate). Thus, the number of participants tested in our experiments was optimal for event-related studies.

E. Related Work

Our study centers on phishing detection and malware warnings. Most closely relevant to the phishing component of our study is the lab study reported by Dhamija et al. [10] with 22 participants, which asked the participants to distinguish between real and fake web-sites. Their results indicated that users do not do well at this task and make incorrect choices 40% of the time. Our behavioral data also yielded similar results. However, our neuroimaging data shows that users exhibit significant brain activation during fake or real website identification task. This suggests that although the outcome of the participants' efforts to differentiate between fake and real web sites may not be good (perhaps because they do not know what to look for on the sites to make this decision), they certainly seem to be making a considerable effort in solving these puzzles as reflected by their brain activity in appropriate brain regions during this decision-making process.

The only prior study that focuses on malware warnings is a very recent large scale field study reported by Akhawe and Felt [11]. This study used modern browsers' telemetry frameworks to record users' real-world behavior when interacting with malware (as well as phishing and SSL) warnings. Unlike previously conducted lab-based studies of security warnings and security indicators (see below), this new study demonstrated that users heed warnings most of the time. Specifically, they found that users ignore Chrome's and Firefox's phishing and malware warnings only 9-23% of the time, and ignored Firefox's SSL warnings 33% of the time. These results are very much in line with the results of our study, which provides neurological proof as to the users' capability to process and heed malware warnings.

For over a decade, many lab studies have focused on different browser security indicators (passive indicators, and active warnings for phishing and SSL attacks) [12, 13, 14, 15, 16, 17]. All of these studies suggested that users seldom act upon warnings and security indicators. (We refer to Akhawe and Felt [11] who provide an excellent survey of the results of these studies). Akhawe and Felt attributed the stark difference in the results of prior lab studies focusing on warnings, and their own field study of [11] mainly to the changes in the nature of the browser warnings.

A previous neuroimaging study somewhat relevant to our work was performed by Craig et al. [18]. This study aimed at understanding users' behavior when faced with advertisements, including the level of suspicion aroused by deceptive advertising. Their study found precuneus and superior temporal sulcus activation while participants processed different levels of deceptive stimuli. This has relevance to user-centered online security interactions, as users may become suspicious when they encounter phishing sites or connect to malware-prone websites. While the Craig et al study points to the cognitive dangers associated with moderately deceptive materials, our phishing task presents participants with a real life online security scenario where they have to determine whether the website is malicious or real.

There have been other studies that applied neuroscience principles to computer security problems, e.g., [19, 20, 52, 53]. Bojinov et al. [19] proposed a neuroscience-inspired approach to coercion-resistant authentication. Martinovic et al. [20] explored the feasibility of side channels attacks with commodity brain-computer interfaces. Thorpe et al. [52], and Chung et al. [53] explored user authentication using EEG devices.

III. DESGIN OF EXPERIMENTS

The in-scanner phase of our *within-subjects* fMRI study is comprised of two experiments, one involving phishing detection and one involving malware warnings. In this section, we discuss the methodology, and design and implementation of these experiments. Since these experiments were implemented using *E-Prime* [2], we begin by providing an overview of this software platform.

A. E-Prime Overview

To develop our fMRI experiments, we used the E-Prime software (Psychology Software Tools Inc., Pittsburgh). E-Prime is a framework for designing and implementing experiments, and collecting the participant response data and exporting this data to different formats for analysis. E-Prime is a suite of applications, namely, E-Studio, E-Run, E-Basic, E-Merge and E-DataAid, where E-Studio is a graphical environment, E-Basic is the scripting language for E-studio, E-Run is for running the environments, E-Merge is for merging session data files into Multi-Session data files and E-DataAid

is for managing data [2]. E-Studio supports creation of experimental environment. It consists of:

1. *interface*: a combination of toolbox, workspace, structure and properties. It has drag and drop functionality to use the objects from its framework in the experiment. Inline scripts can be written in E-basic to control the execution flow of those objects.
2. *frames*: an event in the experiment that includes text or images which run for a certain amount of time.
3. *trials*: collection of frames which forms stimuli in the form of images and text.
4. *blocks*: collection of trials.
5. *procedure*: used to arrange the frames, trials and blocks in a specific order, following a linear time-line.

As a general practice, while developing experiments in E-Prime, one starts with a procedure called Session Procedure. It runs for a session and holds all other objects for that session. Instructions, blocks and trials are then included in the procedure. The time duration of all objects can be fixed as per the requirement of the experiment. An MRI compatible IFIS-SA (*Invivo* Corp., Gainesville, FL) auditory and visual system is available for stimulus presentation at the neuroimaging facility at our University. This system consists of two computers: one for stimulus presentation and another for experimental control and analysis. A master control unit is used to interface the two computers. We use E-Prime software run on the IFIS-SA system to present visual and auditory stimuli. The visual display in the magnet utilizes an IFIS-SA LCD video screen located behind the head-coil that is viewed through a mirror attached to the radio frequency (RF) coil. The auditory stimuli, if any, are presented using MR-compatible pneumatic headphones. The auditory stimuli and video display can be controlled using the master control unit within the scanner's control room. MRI compatible response boxes (e.g., joysticks and button boxes) are available within our neuroimaging center. The E-Prime IFIS-SA systems record the reaction times as well as participant response to each stimulus item presented to the subject in the scanner, and creates data files titled *e-dat* and *t-dat*.

The visual display in the MRI scanner used in our experiments (Section II.B) had the resolution of 640*480 and thus the interface for all the experiments were designed to fit that resolution. Moreover, since E-Prime only supports a 16-bit Bitmap Image format, all snapshots used in the experiments were converted to Bitmap keeping the visual integrity of the stimuli intact.

B. *Phishing and Phishing Control*

Phishing is the act of deceiving people by presenting a fake website which looks like a real one. For this experiment, we identified websites which are popular among people, and took the snapshots of the sites' login pages. We modified the login pages of these websites, created fraudulent replications of them and took snapshots of them. The snapshots were then categorized into two types: "real" and "fake." The fake website snapshots were further divided into two categories: "easy" and "difficult." The easy sites are those for which we modified both the URL and the logo of the companies; keeping the layout of webpages intact; or we changed the URL of the webpages to an IP address. The difficult sites were those for which we modified just the URL keeping the security icons and parameters intact. Table I provides a sample list of the websites used in the experiment along with their URLs. The sites were mainly chosen based on their expected popularity among our participants. Figure 2(a) and 2(b) provide a sample of how these websites images looked for easy and difficult trials, respectively. We obtained some of the URLs from the website www.phishtank.com. The design of fake websites, for this experiment, was similar to the design adopted in the previous study on phishing detection reported by Dhamija et al. [8].

Figure 2(a): sample image "difficult" (URL different compared to real)

Figure 2(b): Sample image "easy" (logo and URL different compared to real)

1) Experiment Design (Phishing): The phishing experiment followed an event-related (ER) design. In ER design, each trial is presented as an event with longer inter-trial-interval. This was done with the goal of isolating fMRI response to each item separately. Event-related designs allow different trials to be presented in random sequences, eliminating potential confounds, such as habituation, anticipation, set, or other strategy effects [51]. In this experiment, we had 39 trials, out of which 3 trials (presented at the beginning of the experiment) were considered as practice trials to familiarize the subjects with the task. During the task, the subjects were asked to determine whether a given snapshot of a website was "fake" or "real."

In addition to trials involving real and fake websites, the experiment had a fixation baseline condition, each of which lasted for 10s. Fixations, in the context of an fMRI experiment, are short blocks of time when the participants are asked to look at a cross on the screen and relax. Fixations are considered as windows of baseline brain activity. Each trial

displayed a website snapshot for 6s followed by a gap of 6s. There were 12 trials involving "easy" fake websites, 13 trials involving "difficult" fake websites, and 14 trials involving real websites. The experiment started with the set of instructions followed by a fixation for 10s, and after every 6 trials, a fixation of 10s was displayed on the screen. Thus, there was a fixation at the beginning of the experiment, at the end, and after every 6 trials. The trials were presented to each participant in a randomized order and the participants had to express whether the site depicted in the snapshot was "real" or "fake" by pressing the designated button. We recorded the response given by users and the corresponding response time.

2) Experiment Design (Phishing Control): The phishing control experiment was designed as a control for the stimuli presented in the phishing experiment. This experiment was identical to the phishing experiment, except that participants were instructed to just look at the images displayed on the screen, and not to engage in an active task. Thus, this experiment had all the visual demands of the phishing experiment except for the decision-making (real or fake website) aspect. In this experiment, 20 snapshots of login pages of different websites, including: *Citibank, USPS, orkut, hi5, 6pm.com,google, bankofAmerica, LinkedIn, chase, instagram, coupons, spotify, onlineshoes, Hotmail, BestBuy, yahoo, discover, AT&T, and Apple* and a portal of our University, were shown to the participants. In this experiment, we were examining the brain responses when users were just looking at the webpages, and we subtracted those signals from the signals we captured from the phishing experiment.

TABLE I. SAMPLE LIST OF WEBSITES USED IN THE PHISHING EXPERIMENT (NOT SHOWN IN THE TABLE ARE OUR UNIVERSITY RELATED SITES, INCLUDING BLACKBOARD AND OTHER PORTALS)

Website	URL
Amazon	http://www.amazon.1click.com/exec/flex-sign-in.com.ch
WellsFargo	www.vvellsfargo.com
eBay	http://91.109.13.183/~ebay/security/
PayPal	http://paypal-verification.com.us.cgi-bin.webscr.cmd.login-submit.dispatch.5885d80a13c17421571527861751275287527525.hargaperumahan.com/
Regions Bank	https://bank.secured/regions-bank/login/index.html
Twitter	https://twitter.login.com
NetFlix	https://signup-netfiix.com/#do-login
Facebook	http://securitycenter.3dn.ru/facebook/warning/account/suspend/index.html
Gmail	https://accounts-google.com/servicelogin?service=mail

C. Malware Warnings

Malware is software created to obtain unauthorized access to computer resources and collect private information. We wanted to identify the neural patterns when people respond to warnings associated with malware. Modern browsers use these warning mechanisms to alert users in case they visit a likely suspicious web site [11] and rely upon users' input to proceed. Our malware warnings experiment consisted of several snapshots of news samples and pop-ups of two types: *non-warnings* and *warnings*. A non-warning pop-up contained casual information or questions in it, and a warning pop-up contained details about the malware threat. In this way, the non-warning pop-up served as a control condition for the warning pop-up. The article itself served the purpose of a primary task the user is engaged in. The news samples were collected from popular news websites such as *CNN, BBC, LA Times, ABC News, and Slashdot.org*. We collected news items from all major categories like entertainment, sports, politics, and general news. We recreated the web pages on our own as the E-Prime software only supports a resolution of 640*480 formatted in Bitmap configuration. We simulated the real-website by showing abstracts of the news first. The participants can click on the *read more* option to read the full news item. When they clicked on "full news" we showed them a pop-up with a warning/non-warning asking them if they wanted to proceed. Depending upon their response, the next populated page was either a full news article or a blank page.

Experiment Design (Malware Warnings): This task required that the subject read a series of articles. While they are reading the articles, they were randomly interrupted by a pop-up asking a specific question (non-warning), or by a pop-up warning (about a malicious threat).

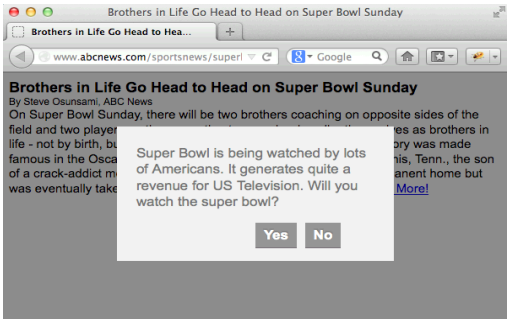


Fig 3(a). A Snapshot of Non-Warning

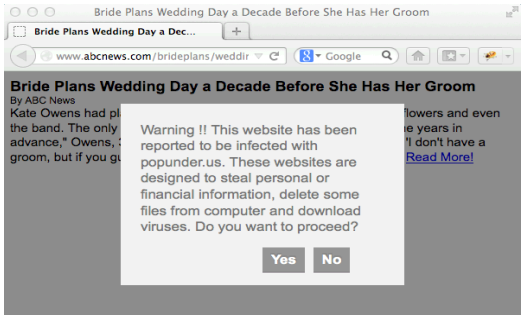


Fig 3(b). A Snapshot of Warning

The experiment started with the instructions set followed by a fixation trial of 10s. After the fixation, the abstract with "read more" link was presented for 10s and when the user clicked on read more, a pop-up (warning or non-warning) was generated asking the user if he/she wanted to proceed. If the user chose not to proceed, a blank screen was displayed for 10s, otherwise, a full news article was displayed for 10s. This was an event-based design and the user gave his input of yes/no by pressing the button. We incorporated the malware warnings of popular web browsers like Chrome, Internet Explorer, Opera, and Mozilla [11]. It was difficult to display all the details of warnings that are shown by these browsers but we kept, to the extent possible, the excerpts similar to the

warnings of these browsers. And the pop-ups contained casual messages, such as, "We are collecting the details about the type of news you like. Do you like these sorts of articles." Figure 3(a) and 3(b) are samples of how the snapshots of the pop-ups looked.

IV. Study Procedures

Our fMRI study followed a within-subjects design, whereby each participant performed all the three tasks, phishing control, phishing, and malware warnings. All tasks were performed in one single fMRI scanning session. The study, including participant recruitment and MRI scanning, ran for a period of about 6 months. In the rest of this section, we present the details of our study protocol, including the recruitment and demographics of our study participants, and the procedures involved during the pre-scanning and scanning phases of the study.

A. *Participant Recruitment & Demographics*

Twenty five healthy university students (14 males and 11 females; mean age: 21.5 years) participated in our fMRI study. Participant demographic information is summarized in Table II. The participating students were enrolled in various educational programs, including: Biology, Music, Athletics, Psychology, Communication Studies, Physical Education, Biomedical Engineering, Pathology, Physical Therapy, Mathematics, Medicine, and Computer and Information Sciences, forming a diverse sample.

Table II. Participant Demographics Summary

	N=25
Gender	14 male; 11 female
Age Range	19 – 32 years
Handedness	24 right-handed; 1 left-handed
Race	13 Caucasian; 5 Hispanic; 6 Asian; 1 African American
Non-Native English Speakers	7

Some of these participants were recruited through a screening questionnaire administered to students enrolled in the Introduction to Psychology course in the Department of Psychology at our University. Participants were not included if they indicated having metal implanted in their bodies (either surgically or accidentally), indicated possibly being pregnant or currently breastfeeding, or indicated having had a history of kidney disease, seizure disorder, diabetes, hypertension, anemia, or sickle cell disease. Individuals were also excluded if they were taking psychotropic medications, had claustrophobia, or had hearing problems. Participants were not recruited if they indicated a history of a developmental cognitive disorder, anxiety disorder, schizophrenia, or obsessive-compulsive disorder. Some of the participants were also recruited via flyers posted on our University campus, and prospective participants could call a number listed on the flyer and answer a screening questionnaire by dialing in information through integrated voice response (IVR).

B. *Pre-Scanning Phase*

The scans were performed at the neuroimaging facility available to us at our University. Participants signed an informed consent form approved by our University's Institutional Review Board. In addition, participants filled out an Edinburgh Handedness form [54], an MRI safety questionnaire, and a Barratt's Impulsivity questionnaire [1]. The purpose of the handedness form was to determine handedness because handedness may relate to the lateralization of hemispheric activity in the participants (right-handed individuals may be more left-lateralized). The purpose of the impulsivity questionnaire was to determine the trait impulsivity level of the participants (details provided in Section V.A).

Prior to the scan, each participant was shown example images for both the tasks (phishing detection, and malware warnings) in the form of images on paper. We also explained that the participant was to use the button response system in the MRI scanner during the tasks.

We did not tell the participants before the fMRI scan as to what they are supposed to be doing in the phishing experiment. This was to make sure that they were not influenced by the "real" or "fake" decision-making thoughts while engaged in the phishing control experiment. The image for the phishing control experiment was a screenshot of the Google home page. The image we showed to the participants for the malware experiment is one of the articles displayed in a browser pop-up asking a general question about the article. See Figures 2(a), 2(b), 3(a), and 3(b) for examples shown to the participants. We did not show an example of a malware warning pop–up to the participants before the experiment in order to avoid priming them explicitly.

C. *Scanning Phase*

fMRI data was collected using a Siemens 3.0 T Allegra head-only scanner (as discussed in Section II.B). For structural imaging, initial high resolution T1-weighted scans were acquired using a 160-slice 3D MPRAGE (Magnetization Prepared Rapid Gradient Echo) volume scan with TR = 200 ms, TE = 3.34 ms, flip angle = 1210, FOV = 25.6 cm, 256 x 256 matrix size, and 1 mm slice thickness. A single-shot gradient-recalled echo-planar pulse sequence was used to acquire functional images (TR = 1000 ms, TE = 30 ms, flip angle=600, FOV = 24 cm, matrix = 64 x 64). Seventeen adjacent oblique axial slices were acquired in an interleaved sequence with 5 mm slice thickness, 1 mm slice gap, a 24 x 24 cm field of view (FOV), and a 64 x 64 matrix, resulting in an in-plane resolution of 3.75 x 3.75 x 5 mm3.

For each participant, we set the order of the phishing and malware warnings tasks randomly, but always left the phishing control as the first task. We gave appropriate instructions to the participants via an intercom before each experiment started. Instructions were also provided visually on the display screen in the MRI scanner at the beginning of each task. Each task was run through the *IFIS System Manager*. Participants made their responses using a fiber optic button response system that had a button for each finger on both hands. They indicated a "yes" response using their right index finger and a "no"

response using their left index finger. The session start time, end time, session id, and the order that each task was performed were recorded. For each task, except the phishing control experiment, reaction times and answers were automatically recorded and saved as dat files. The total duration for the phishing control, phishing and malware warnings tasks were 268s, 553s, and 751s.

After the scanning phase was over, we compensated the participant with Psychology course credits or a $50 cash reward depending on their status.

V. Analysis and Study Results

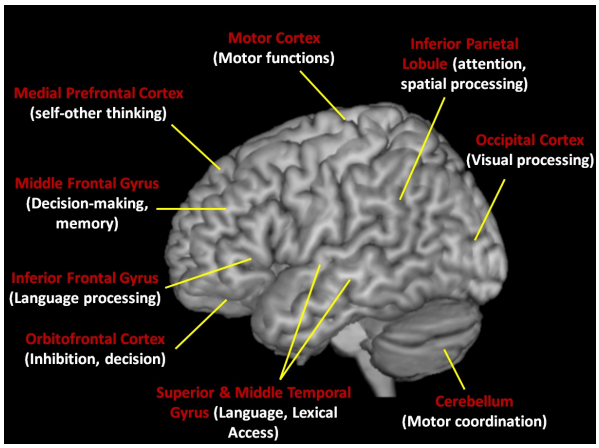
In this section, we provide a comprehensive analysis of the data acquired during our study and attempt to interpret the results. We report on the neuroimaging data analysis followed by the behavioral data analysis

A. *Neuroimaging Data Analysis*

All acquired fMRI images were converted from DICOM (Digital Imaging and Communications in Medicine) format to NIFTI (Neuroimaging Informatics Technology Initiative) format using the Free Surfer software (http://surfer.nmr.mgh.harvard.edu/). Data was preprocessed using SPM8 software (Wellcome Trust Centre for Neuroimaging, London, United Kingdom) within MATLAB and an in-house software. Functional data preprocessing started with slice time correction to account for the interleaved pattern of scan slice acquisition. All slices were realigned to the mean image in the scan. All images were then normalized to the EPI template provided by SPM8 using a $2mm^3$ resampling voxel. Head motion was examined in three translational directions x, y, and z, and three rotations: pitch, roll, and yaw. A cut off point of 1 mm in any direction was kept as the criteria for motion. After these quality control measures, data from three participants from the phishing experiment were discarded resulting in 22 usable datasets for that experiment. Lastly, all normalized images were smoothed using a Gaussian filter of 8mm full width half maximum.

Statistical analyses were performed on individual data and group data using the General Linear Model (GLM). In GLM analysis, each voxel in the brain will have a signal time-series for a given experiment based on how that voxel behaves in response to a specific task. The GLM formula is $Y = X*\beta+\varepsilon$, where Y is the fMRI signal at various time points at a single voxel, X is several components (the design matrix with different conditions, such as real, fake, malware) that can explain the observed fMRI signal, β is the parameter that defines the contribution of each component of the design matrix to the value of Y, and ε is the difference between the observed data (Y) and that predicted by the model ($X*\beta$). Group analyses were performed using a random-effects model. Regions of interest (ROIs) with statistically significant activation were identified using a *t*-statistic on a voxel by voxel basis. Separate regressors were created (for real, fake, and fixation stimuli in phishing experiment, and abstract, warning, and no-warning for malware experiment) by convolving a boxcar function with the standard hemodynamic response function as specified in SPM. Statistical maps were superimposed on normalized T1-weighted images. All data were intensity-thresholded at p=0.001, with a cluster size correction per region for a family wise error (FWE) rate of 0.05. To determine the voxel threshold for significance, a minimum cluster thresholding operation was performed using the AlphaSim software package in AFNI (Analysis of Functional Neuroimages) [57]. Ten-thousand Monte Carlo simulations were generated to maintain the family wise error (FWE) rate at 0.05 for the whole brain. Thus, in order for a given region to be considered significantly active, it should have a minimum cluster size of $64mm^3$.

Table III lists the acronyms we will be using in the rest of the paper for the brain regions associated with our experiments. The figure below provides a higher level overview of these different regions and their general role.



Overview of various brain regions associated with our experiments

TABLE III. ABBREVIATIONS FOR BRAIN REGIONS ACTIVATED DURING OUR EXPERIMENTS

Acronym	Brain Region
MPFC	Medial Prefrontal Cortex
LIFG	Left Inferior Frontal Gyrus
RIFG	Right Inferior Frontal Gyrus
LMFG	Left Middle Frontal Gyrus
RMFG	Right Middle Frontal Gyrus
LOFC	Left Orbitofrontal Cortex
ROFC	Right Orbitofrontal Cortex
LMTG	Left Middle Temporal Gyrus
RMTG	Right Middle Temporal Gyrus
LSTG	Left Superior Temporal Gyrus
RSTG	Right Superior Temporal Gyrus
LIPL	Left Inferior Parietal Lobule
RIPL	Right Inferior Parietal Lobule
LOC	Left Occipital Cortex
ROC	Right Occipital Cortex

(1) PHISHING DETECTION EXPERIMENT RESULTS

In this section, we present the phishing experiment results. We first report the raw analysis results, and then interpret and discuss all of the findings.

In the phishing task, participants were asked to make a decision as to whether a snapshot of a website presented to

them was real or fake. The participants could be looking at the website address or the symbols or logos on the snapshot to make their decision. During this task (in contrast with fixation), we found statistically significantly increased brain activity in the bilateral frontoparietal network along with activation in bilateral occipital areas ($p = 0.001$; cluster size = 64 mm^3 determined by Monte Carlo simulations to be equivalent to a family wise error corrected threshold of $p < 0.05$) [21]. There was increased activity in the bilateral inferior, middle and orbital frontal areas, in the bilateral inferior parietal lobule, and in the bilateral occipital extending to ventral temporal areas. This pattern of activity was similar across different conditions: real, fake or real+fake, all contrasted with fixation (see Figure 4 for a map of brain activity in different contrasts).

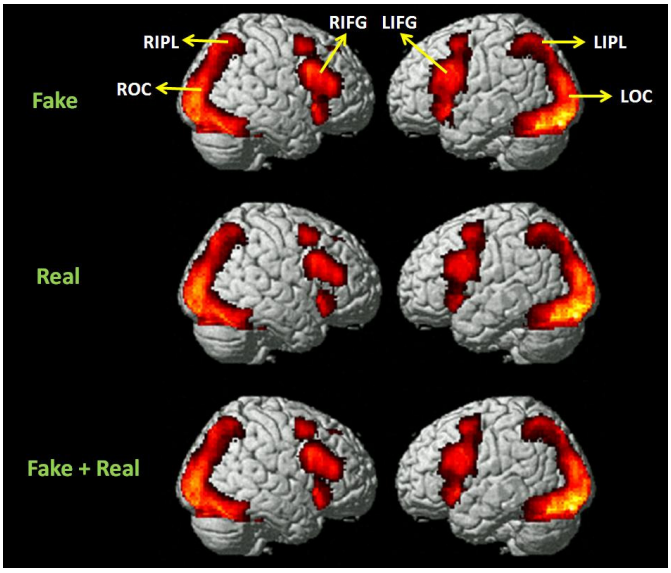


Fig 4. "**Fake", "Real", and "Fake+Real" Activation**. Activation regions include bilateral frontoparietal along with bilateral occipital areas (LOC/ROC), as well as bilateral inferior, middle and orbital frontal areas (LIFG/RIFG; LMFG/RMFG) and bilateral inferior parietal lobule (LIPL/RIPL), and bilateral occipital extending to ventral temporal areas.

Direct subtraction of real trials from fake trials, and fake trials from real trials revealed statistically significant activity in several areas of the brain that are critical and specific to making "real" or "fake" judgments. For websites that the participants identified as fake (contrasted with real), participants activated right middle, inferior, and orbital frontal gyri, and left inferior parietal lobule (see Figure 5). On the other hand, when real websites were identified, participants showed increased activity in several regions, such as the left precentral gyrus, right cerebellum, left cingulate gyrus, and the occipital cortex.

All participants of this study also completed the Barratt's Impulsiveness Scale (BIS) [1]. BIS is a 30 item self-report instrument designed to assess the personality/behavioral construct of impulsiveness. It is perhaps the most commonly administered self-report measure specifically designed for the assessment of impulsiveness in both research and clinical settings. Impulsive responding can result in behavioral errors, and such responses can be critical in computer security interactions where the consequences can be costly. Thus, our goal was to examine the impact of impulsive decisions on phishing task performance and the neural circuitry underlying such behavior. A regression analysis involving BIS scores from participants as a covariate with whole brain activation revealed a statistically significant negative relationship in the MPFC ($p < 0.001$; cluster size = 64 mm^3) (See Figure 6). In other words, more impulsive individuals had less activity in MPFC.

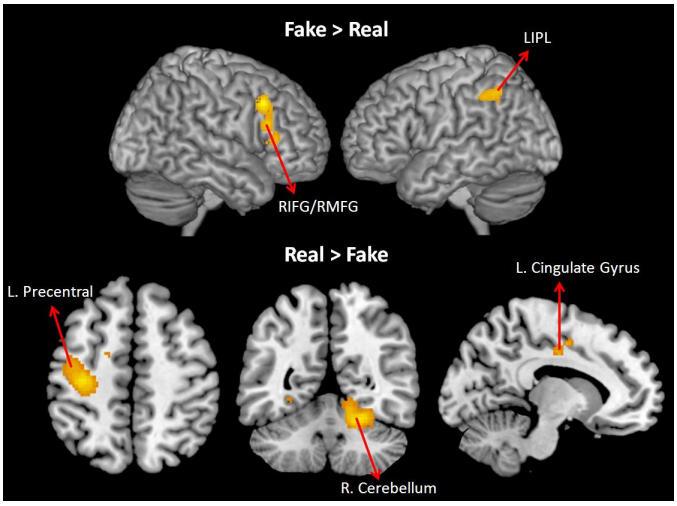


Fig 5. **Contrast of "Real" and "Fake" Activation**. Fake vs. Real activation regions include right middle, inferior, and orbital frontal gyri (LIFG/LMFG), and left inferior parietal lobule (LIPL). Real vs. Fake activation regions include left precentral gyrus, right cerebellum, left cingulate gyrus, and the occipital cortex.

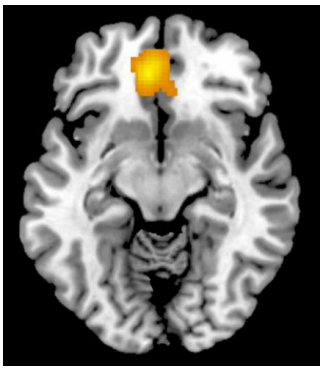


Fig 6. **Impulsivity vs. MPFC Activation in Malware Warnings**. There exists a negative relationship between impulsivity and brain activity in medial prefrontal cortex (MPFC).

Interpretation and Discussion (Phishing Detection): Searching, attention shift, and decision-making are critical components of higher cognitive functions. The phishing task in the present study involves all these elements in helping participants decide which website was real and which was fake. At the neural level, the statistically significant activity we see in frontoparietal network (Figure 4) may be indicative of the involvement of a top-down control and attention modulation system and a bottom-up control system in this task. The top-down system consists of regions, such as the intra-parietal sulcus (IPS), and superior parietal lobule (SPL), and the bottom-up system includes temporoparietal junction

(TPJ) and inferior fontal gyrus (IFG) [22, 23]. Many of these regions that are part of top-down and bottom-up control showed significant activity in the phishing task in the present study. Such increased control may be critical in making important judgments about the legitimacy of a website.

Increased activation was found in the right frontal and left parietal regions while deciding that a given website is fake (Figure 5). At one level, this is another evidence of a more strategic and controlled approach to a more difficult task (identifying fake websites). These findings are consistent with a previous fMRI study [24], where the participants were asked to identify whether a series of Rembrandt paintings were real or fake. This study found increased activity in RMFG while participants identified fake paintings. Fake websites may pose more challenge to the participants as they may have to spend more time thinking about different attributes, sometimes recalling from memory. Middle frontal, inferior frontal, and inferior parietal areas have also been implicated in working memory [25]. Identifying real websites activated precentral, cerebellum, cingulate and visual areas (Figure 5). In addition to their motor functions, the cerebellum and precentral gyrus have topographically organized feedforward and feedback projections [26]. This network may be mediating the decision-making process as to whether a given website is real.

Yet another finding from the present study pertains to a brain-behavior relationship. Personality traits, such as impulsivity may prove vital in the way an individual approaches a cognitively demanding task. Impulsive individuals may seek immediate gratification and may make quick decisions without much thought. Such behavior can affect online computer security behavior. The present study found an inverse relationship between impulsivity and MPFC activity during real or fake phishing decisions (Figure 6). Evidence from previous studies suggest MPFC's executive/regulatory function in that it mediates competing and conflicting cognitive operations and scenarios [27, 28, 29, 30, 31]. Studies involving animal models suggest a pivotal role of MPFC in impulsive decision-making [32]. Functional MRI studies of delay discounting have found inverse correlation between participants' impulsive choice of decisions and activity in regions like MPFC [33, 34]. Delay discounting refers to giving future consequences less weight relative to more immediate consequences (e.g., [35]). In other words, delay discounting can be construed as the tendency to choose a smaller, sooner reward over a larger, later reward. Similar finding of inverse correlation in the present study suggests the conflict and difficulty involved in making real or fake decisions during the phishing task for impulsive individuals.

(2) MALWARE WARNINGS EXPERIMENT RESULTS

In this section, we present the results from our malware warnings experiment. Similar to the phishing detection presentation, we first report the raw results and then interpret and discuss all of our findings.

To recall, in the malware warnings task, a section of a news item (abstract) was presented to the participants. As the participants read the news item, a message popped up on the screen which either cautioned them about a malicious computer attack (warning condition) or a casual pop-up (non-warning condition) asking them a question or seeking information from them. The participants were asked to decide "yes" or "no" before proceeding. Thus, there were three experimental conditions: *abstract*, *warning*, and *non-warning*. Reading the news item, relative to warning and non-warning taken together, elicited statistically significant increase in brain activity in several regions primarily associated with language comprehension. These regions include LMTG/LSTG, LIFG, bilateral inferior parietal (IPL), and bilateral occipital cortex ($p < 0.001$ with a cluster threshold of $80mm^3$ that is equivalent to a family wise error correction of $p < 0.05$ determined by Monte Carlo simulations [21], as in the phishing data analysis). There was also some activity in the right inferior frontal (RIFG) and middle temporal (RMTG) regions, perhaps not to the same extent as their left hemisphere homologues. This pattern of activity was also seen in the contrasts: Abstract > Non-warning and in Abstract > Warning (see Figure 7).

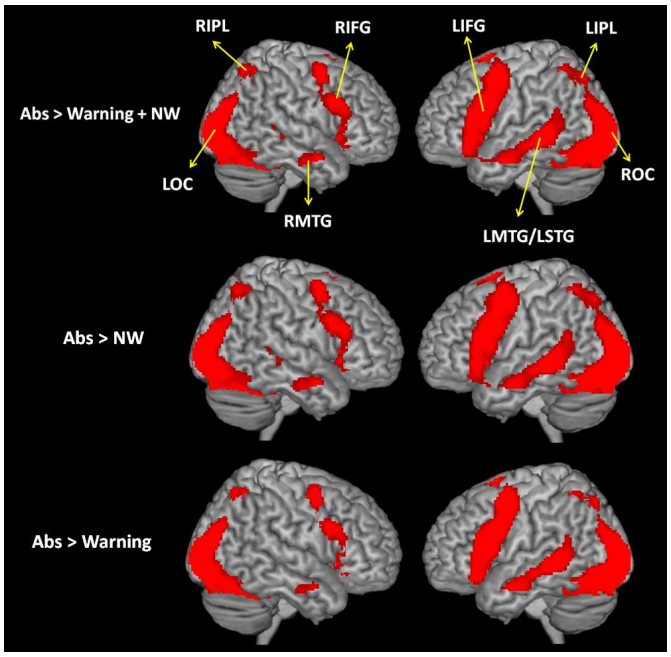


Fig 7. **Abstract vs. (Warning or Non-Warning) Activation**. Activation regions include left middle and superior temporal gyri (LMTG/LSTG), left inferior frontal gyrus (LIFG), bilateral inferior parietal (LIPL/RIPL), and bilateral occipital cortex (LOC/ROC), as well as right inferior frontal (RIFG) and middle temporal (RMTG)

The warning and non-warning conditions showed more activity compared to reading the abstracts of the news items. Comprehending warning, relative to abstract, elicited statistically significant increase in activation in several regions of the right hemisphere, such as the RIPL, RMTG/RSTG, and cuneus (see Figure 8). Processing non-warning pop-ups, relative to news item abstracts, also elicited similar patterns of activation, albeit with some differences. There was bilateral activation in middle/superior temporal cortex in this contrast. In addition, the right parietal activation was relatively more anterior, in the postcentral gyrus.

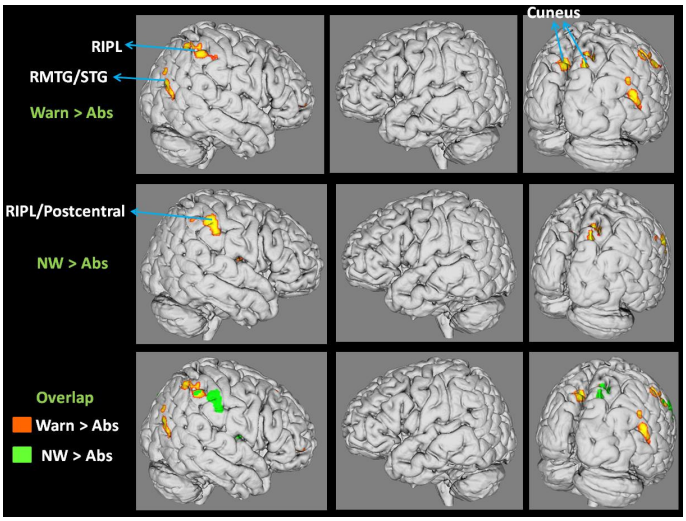


Fig 8 (**Warning or Non-Warning) vs. Abstract Activation**. Activation regions include right inferior parietal lobule (RIPL), right middle/superior temporal gyrus (RMTG/RSTG), and cuneus, as well as bilateral middle/superior temporal cortex, and right parietal in the postcentral gyrus. (The second column brain images do not show any activation; they are included for the sake of completeness)

One of the main goals of this study was to examine the brain areas that may mediate how people approach malware warnings. The participants showed significant increase in brain activity in several areas while processing warnings, relative to non-warnings. These regions include LIFG and LMTG, both primarily associated with processing language. There was also increase in activity in regions, such as the MPFC, and in the bilateral occipital cortices (see Figure 9). On the other hand, we did not find any increase in brain activity for the non-warning condition, relative to the warning condition.

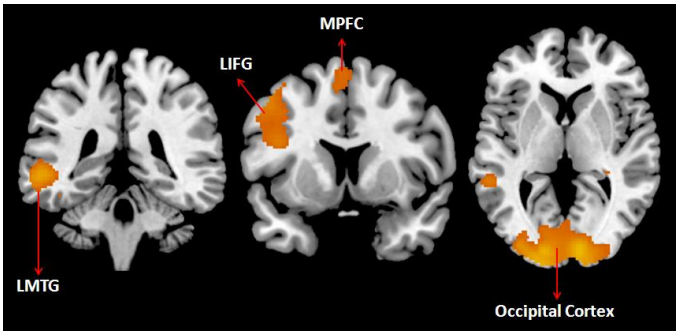


Fig 9. **Warning vs. Non-Warning Activation.** Activation regions include left middle temporal gyrus (LMTG), left inferior frontal gyrus (LIFG) as well as medial prefrontal cortex (MPFC), and bilateral occipital cortices.

In order to examine personality traits and their impact on computer security decisions, as in the phishing data analysis, we used impulsivity scores as a covariate in a regression analysis with brain activity while reading security warnings. This analysis revealed significant negative relationship between impulsivity and brain activity in MPFC and precuneus (see Figure 10).

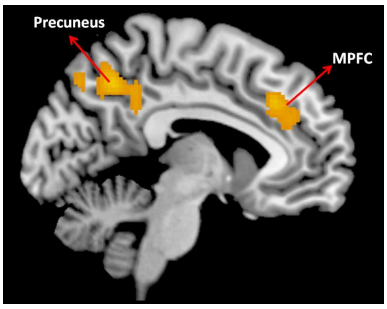


Fig 10. **Impulsivity vs. Activation in Malware Warnings**. There is a negative relationship between impulsivity and brain activity in medial prefrontal cortex (MPFC) and precuneus

Interpretation and Discussion (Malware Warnings):

Language comprehension is a critical component in a user's online interactions. In this study, participants had to read several news items and make appropriate responses to malware warnings (and non-warnings). Reading the news items as well as reading warnings generated significant brain activity in regions such as the LIFG and LMTG/LSTG (Figure 7). This activation pattern provides further evidence of the role of these regions in different aspects of language comprehension (see [36, 37, 38]). The LIFG, in particular, has been implicated in the unification of lexical information stored in the temporal cortex [39]. Activation in these areas suggests that the participants in the present study were going through the warnings to understand the conveyed message and make a decision.

There were also qualitative differences in activation between processing warning and non-warning pop-ups. Warnings generated statistically significant increase in activity in language areas of the brain, such as LIFG and LMTG (Figure 9). In addition, there was statistically significant activation in bilateral occipital cortices, which may provide evidence of how much visual attention and inspection participants were engaging in during warnings. On the other hand, non-warnings, which usually were not a threat did not generate any extra activation when compared with the warning condition.

High impulsivity, the tendency to act quickly without considering the broader, especially future, consequences of one's actions, has been found to be associated with several psychopathological conditions [40, 41]. Impulsive decisions can affect user safety and security in a computer security interaction (as we demonstrated in the case of phishing). We found trait impulsivity in our participants, measured by the Barratt's Impulsivity Scale, to negatively predict brain activity in MPFC and precuneus while paying attention to security warnings (Figure 10). Thus, more impulsive participants had less activity in these regions during the malware task. This finding is consistent with findings from several previous neuroimaging studies. For example, the precuneus was found to be negatively correlated with measures of impulsivity in a response inhibition task [42]. MPFC grey matter volume has also been found to be negatively correlated with impulsivity [43].

B. Behavioral Data Analysis

Phishing Detection Experiment: During the phishing experiment, we instructed the participants to give their response during the trials by pressing the buttons using their index fingers. Each participant was asked to use his/her left index finger to respond that the website is fake, and the right index finger to respond that the website is real. Our E-Prime interface automatically recorded the response made by the participants and the corresponding response time.

Based on this recorded data, we collected statistics for participant accuracy (acc) and response time (time) for different types of trials, as summarized in Table IV**.** Accuracy is defined as the fraction of times a particular trial was correctly identified out of the total number of occurrences for that trial.

We can observe that, on an average across all trials, the participants spent 3.35 seconds to make a decision but the accuracy was only around 60%, i.e., 40% of the times, they misidentified a fake website as a real website and a real website as a fake website. This accuracy is only slightly better than making a random guess. Prior work by Dhamija et al. [8] also reported very similar results with their computer-based lab study. As mentioned in Section V.A, our neuroimaging data indicated strong brain activation in regions associated with decision making during both real and fake judgment, particularly more during fake judgment. Although it is unclear why fake trials would show increased brain activity in RMFG but poor accuracy, it is quite possible that the participants were putting more effort to identify fake websites but still not able to come up with the correct answers.

TABLE IV: STATISTICS FOR ACCURACY AND RESPONSE TIME – PHISHING EXPERIMENT

Trials	μ_{acc} (σ_{acc})	μ_{time} (σ_{time})
Real	76.68% (18.84%)	3323 ms (1066 ms)
Fake	46.48% (20.58%)	3276 ms (584 ms)
Easy Fake	56.57% (23.29%)	3077 ms (625 ms)
Difficult Fake	33.98% (23.61%)	3538 ms (645 ms)
All	60.42% (13.99%)	3347 ms (654 ms)

The average accuracy of identifying real websites was the highest (about 77%) and much higher than the average accuracy of identifying fake websites (about 46%). Similar to the findings of our neural imaging data, this suggests that real website detection is an easier task compared to fake website detection. Intuitively, the accuracies for easy fake trials were higher than for difficult trials (57% vs. 39%).

The average response time for all types of trials was similar, over 3 s. The average reaction time for identifying easy fake website was the lowest and for identifying difficult fake website was the highest. We also found a statistically significant difference between the response time of easy fake and difficult fake trials (paired t-test; p-value = 0.006; CI = 95%). Easy fake websites had differences easier to distinguish than the difficult fake websites, so participants might have noticed them pretty early when making decision. Moreover, the average time spent on deciding the difficult fake websites was comparatively more than the time spent for other trials, but still the accuracy rate for difficult fake was the lowest among all trials. The participants may have spent more time on the difficult fake trials but still could not detect them as fake due to the level of difficulty associated with these trials. However, we did not find significant correlation between the response time and accuracy for any of the trials.

Malware Warnings Experiment: The data acquisition approach for the malware experiment was similar to that for the phishing experiment. The participant pressed a button using his or her left index finger to indicate "No" and a button using his or her right index finger to input "Yes." For the warnings conditions, pressing "No" was equivalent to heeding the warning and "Yes" was equivalent to ignoring the warning.

TABLE V: STATISTICS FOR ACCURACY AND RESPONSE TIME – MALWARE EXPERIMENT

Conditions	μ_{acc} (σ_{acc})	μ_{time} (σ_{time})
Non-Warnings	67.49 % (26.57%)	4228 ms (664 ms)
Warnings	88.71% (28.62%)	3715 ms (1141 ms)
All	81.05% (19.59%)	4022 ms (588 ms)

Similar to the phishing experiment, we collected statistics for subjects' accuracy (acc) and response time (time) for the different conditions, as summarized in Table V**.** Accuracy is defined as the fraction of times a participant pressed "No" for a particular condition out of the total number of occurrences of that condition.

An important observation is that the accuracy for heeding the warnings was quite high (about 89%), which means that participants paid attention to these warnings and chose not to "click-through" most times. This result is in line with the results from a recent large-scale field study of Akhawe and Felt [11]. It is also validated by the high brain activation in brain regions associated with problem solving and decision making as shown by our neuroimaging analysis (Section V.A).

We can also see that participants spent shorter amount of time reacting to warnings than non-warnings. This difference was also found to be statistically significant (paired t-test; p-value = 0.04; CI = 95%). Also, the probability of overriding the pop-up was less for warnings than non-warnings. A similar effect was observed via our neural data analysis. Overall, this indicates that the content of warnings might have been prominent enough to raise suspicion in the minds of the users, resulting in their pressing "No" more quickly.

VI. CROSS-EXPERIMENT ANALYSIS

So far, we have presented the findings drawn from the independent analysis of different experiments. One important

feature of our study design, involving multiple experiments in the same scanning session, was to facilitate measuring users' performance across these experiments. In this section, we present the results from such a cross-experiment analysis.

A. *Phishing vs. Phishing Control*

In order to examine the overlapping and unique activity associated with the phishing task and a visual control task, we compared the phishing with the phishing control experiment. While the phishing task involved participants making judgments about whether the websites were "real" or "fake", the phishing control task presented the same websites as the stimuli, but the participants were told to relax and view them without engaging in any active tasks. Both tasks elicited significantly increased activity in the visual cortex, perhaps in line with the visual demands of the stimuli. However, the phishing task showed significantly greater and unique activation in various regions, such as RMFG and bilateral insula (see Figure 11), a pattern not seen in the phishing control experiment.

The anterior insula has been implicated in a variety of functions, such as affective and cognitive judgments. Activation in anterior insula, along with MFG, has been associated with making choices [44, 45]. The middle frontal gyrus also has been found to be playing a critical role in cognitive control especially in selecting an appropriate choice of action [46]. The activation of these important decision-making regions of the brain in the phishing experiment (vs. the control experiment) demonstrated that the participants were conscientiously making an effort as to differentiate "fake" websites from "real" websites.

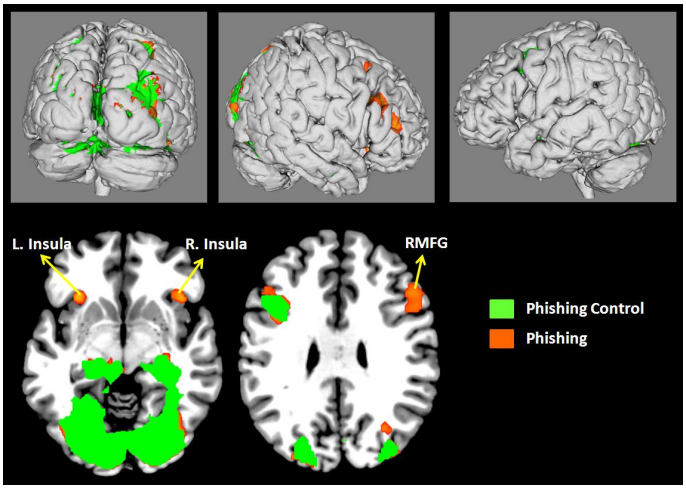


Fig 11. **Phishing vs. Phishing Control Activation**. Both tasks show significant activity in the visual cortex. Phishing shows greater and unique activation in the right middle frontal gyrus (RMFG) and bilateral insula. (The top right corner brain image only shows little activation; it is included for the sake of completeness).

B. *Phishing vs. Malware*

Both phishing and malware tasks in our study involved decision-making, perhaps in slightly different ways. While the malware task tested whether participants were paying attention to security warnings while reading a news item, the phishing task explicitly examined subjects' ability to distinguish between a "real" and a "fake" website. At the neural level, we examined the correlation between these two tasks in terms of the brain activity in two regions, LMFG and RMFG, which are associated with decision-making. We found a significant positive correlation in both LMFG and RMFG activity, particularly in the RMFG region (see Figures 12(a) and 12(b)).

These results suggest that both phishing detection and malware warnings involve similar, higher level cognitive and neural processes. We may also infer that participants' behavior in these two distinct yet related tasks may be well-aligned in that one's ability to heed malware warnings may be associated with his/her decisions about the legitimacy of a website and vice versa. Thus, the quality of online security behavior may be determined by users' cognitive ability and by the selective activation of specific brain areas in appropriate contexts.

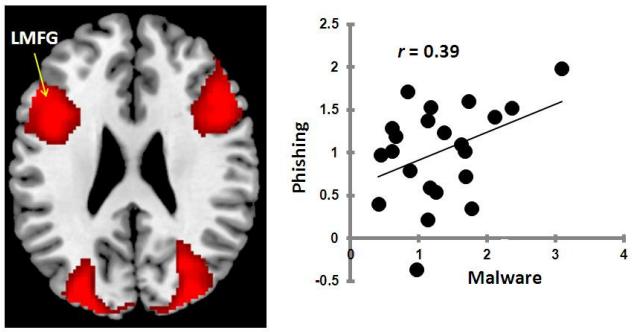


Fig 12 (a). Correlation in Phishing and Malware in LMFG Activation

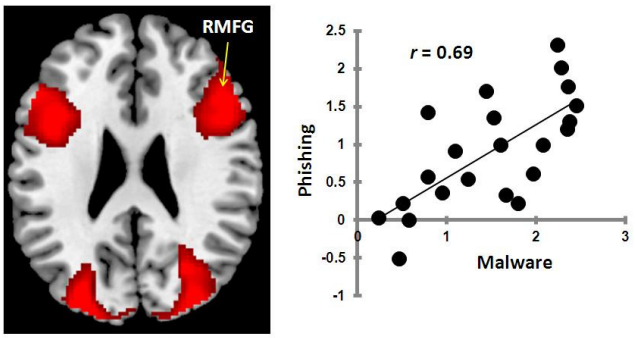


Fig 12 (b): Correlation in Phishing and Malware in RMFG Activation

VII. Discussion: Study Insights and Implications

In this section, we summarize and further discuss the main findings from our study. The neural signatures (activation in unique brain regions) associated with, and across, the phishing detection and malware warnings tasks are summarized in Table VI**.**

Distinguishing between fake and real websites underlying the phishing task produced increased activation in many areas of the brain associated with decision making, problem solving, attention and visual search. This means that the participants were undergoing significant effort in making the "fake" or

"real" decisions. Our neuroimaging results also show extra activation in decision-making areas such as RMFG while identifying a fake website compared to identifying a real website, perhaps reflecting increased task difficulty. Moreover, the comparison between the phishing and phishing control tasks clearly highlighted the decision-making aspect associated with the former (as opposed to just the visual demands associated with both of them).

While our neuroimaging data showed that users exhibited significant brain activation during the phishing detection task, their accuracy in this task, as determined by the behavioral data, was only slightly better than making a random guess (in line with a prior lab study [10]). This suggests that although the eventual decision made by the participants to differentiate between fake and real websites may be far from accurate, they certainly were putting a considerable effort in making this decision as reflected by their brain activity in regions correlated with higher order cognitive processing. Perhaps this was because many of the participants did not know what markers to look for (e.g., URL or logo) on the sites to make their decisions. We note that a large fraction of our participants belonged to a non-technical (non-computer) background. Overall, these findings further justify the need for specialized *education and training* for every day users focusing on phishing in particular (such as the efforts of [47, 48]) and security in general (such as [49, 50]). These training and awareness programs may help improve users' phishing detection performance and reduce the chances of their susceptibility to other attacks. At the same time, the findings also motivate the need for continued research on designing phishing resistant software solutions and user interfaces.

The malware warnings task triggered significant brain activity in regions primarily associated with language comprehension and reading. Importantly, the actual malware warnings, in contrast to casual pop-ups, generated significantly more activation in brain areas governing language comprehension as well as visual attention and inspection. This suggests that the participants were reading through the warnings carefully to understand the message conveyed by the warnings and making an attempt to take an appropriate decision. Indeed, this was validated via our behavioral data which showed that participants heeded warnings about 90% of the times (also in line with the recent large-scale field study of [11]). We therefore believe that our study provides a neurological basis as to the users' capability to process and heed malware warnings, further validating the results of [11]. It should be noted that since our security warnings were quite simplified, our results seem to underestimate users' performance when faced with malware warnings, which could be improved with better warnings (such as those employed by modern browsers and variants thereof [11]).

Another key component of our study was to asses users' performance in user-centered security tasks based on their personality traits. Specifically, we studied the effect of impulsivity measured via a simple questionnaire. We found that, in both phishing detection and malware warnings tasks, impulsive individuals showed significantly less brain activation in regions governing decision-making and problem solving. This implies that impulsive behavior might be counter-productive to phishing detection and malware warnings task performance. A long-term impact of this finding can be in developing *targeted security training programs*. For example, an organization may concentrate their security training efforts on employees who are highly impulsive, as determined by their scores in the impulsivity questionnaire [1]. Similarly, school authorities may focus their online child safety efforts on children with high trait impulsivity.

A unique advantage of our study was that it allowed for a direct comparison between the phishing detection and malware warnings tasks. In this respect, we found significant correlation in participants' brain activity governing decision-making regions (bilateral middle frontal gyri). This suggests that both tasks involve, at a higher level, some similar cognitive processes and that users' performance in the two tasks might be correlated with each other. Note that, although language comprehension is unique to the malware task, both tasks involved a crucial decision making aspect. Broadly, this seems to indicate that the cognitive mechanisms underlying these security tasks are related, which may translate into similarity in users' performance in the two tasks.

Although fMRI scans are expensive, we believe that our methodology could also serve the purpose of *security screening* of individuals involved in high-security operations, such as in the national defense sector.

VIII. Conclusions and Future Work

In this paper, we presented an fMRI study to bring insights into user-centered security, specifically focusing on phishing detection, and malware warnings. Our results provide a largely positive perspective towards users' capability and performance vis-à-vis these crucial security tasks. We found that users show significant brain activity in key regions known to govern decision-making, attention, and problem-solving ability (phishing and malware warnings) as well as language comprehension and reading (malware warnings). This level of activation indicates that users were actively engaged in these tasks, and were not ignoring or bypassing them (as many prior lab studies seem to have concluded [12, 13, 14, 15, 16, 17]). In the case of the malware warnings task, brain activity and behavioral performance (accuracy) were complementing each other validating that users heed malware warnings with a high likelihood (as also shown by a recent field study [11]). For the phishing task, however, task performance was poor despite significant brain activity associated with decision making. This divergent result demands future investigation. It could be attributed to users' lack of knowledge as to the markers for "fake" vs. "real" decisions (e.g., URLs), which may be overcome by user education and training. We also demonstrated that individuals with higher impulsive traits may not utilize their neural resources as efficiently as non-impulsive individuals, and may result in poorer cognitive and behavioral outcomes. This suggests it would be valuable to study whether individual trait characteristics should factor into user-center security design. Finally, we discovered a high degree of correlation in brain activity (with respect to decision-making regions) across phishing detection and

malware warnings tasks. This correlation suggests that users' behavior in one task may be predicted by their behavior in the other task.

We see a clear path-forward for subsequent research using neuroimaging techniques (e.g., fMRI, EEG or fNIRS) to inform the design of user-centered security systems. In the long-run, such studies may provide a neural signature for poor and good security decisions which can be used for predicting as well as correcting users' security behavior. Future research may conduct subsequent evaluation with diverse participant samples, study the effect of warning fatigue or habituation, consider user-centered security domains other than phishing detection and malware warnings (e.g., password memorization and recall), and evaluate the effect of security training and education on users' performance.

TABLE VI: NEURAL SIGNATURES OF PHISHING, MALWARE WARNINGS, PHISHING VS.PHISHING CONTROL, AND PHISHING VS. MALWARE WARNINGS

Task	Condition	Activation Regions (Neural Signatures)	Activation Association
Phishing Detection	Fake; Real; Fake+Real (*Figure 4*)	LOC & ROC, LIFG/RIFG, LIPL/RIPL Bilateral Occipital extending to Ventral Temporal areas	Visual processing, search, attention shift, and decision-making
	Fake vs. Real (*Figure 5*)	RIFG/RMFG, LIPL	Search, attention shift, and decision-making (*fake decisions are harder*)
	Real vs. Fake (*Figure 5*)	Left Precentral Gyrus, Right Cerebellum Left Cingulate Gyrus, Occipital Cortex	Attention, decision-making, and visual processing
Malware Warnings	Abstract vs. Warning; Abstract vs. Non-Warning; Abstract vs. Warning+Non-Warning (*Figure 7*)	LMTG , LSTG ,LIFG, LIPL/RIPL, LOC/ROC, RIFG , RMTG	Language comprehension, visual processing, reading
	Warning or Non-Warning vs. Abstract (*Figure 8*)	Cuneus, Right Middle/Superior Temporal Cortex, RIPL	Language comprehension, visual attention
	Warning vs. Non-Warning (*Figure 9*)	LMTG, LIFG, MPFC, LOC/ROC	Language comprehension, visual attention and inspection (*warnings show more activity than non-warnings*)
Cross-Experiment	Phishing vs. Phishing Control (*Figure 11*)	Visual Cortex (both tasks) RMFG (phishing) Bilateral Insula (phishing)	Visual processing (both tasks) search, attention shift, and decision-making (phishing)
	Phishing vs. Malware (*Figure 12*)	LMFG RMFG	Decision-making, visual attention

ACKNOWLEDGMENTS

We thank N. Asokan, Cali Fidopiastis, Lauren Libero, Ivan Martinovic, Paul Van Oorschot, and John Sloan for their feedback on a previous draft version of this paper, and Rishi Deshpande for initial help with the experimental set-up. We also thank David Wagner (our shepherd) and NDSS'14 anonymous reviewers for their constructive input and guidance.

REFERENCES

[1] Barratt, E.S. (1994). Impulsiveness and Aggression. In Monahan, J. and H. J. Steadman (Eds.), Violence and Mental Disorder: Developments in Risk Assessment (pp. 61-79). University of Chicago Press, Chicago, IL.

[2] An introduction to E-Prime, Laurence Richard, Miami University, Dominic Charbonneau, Université de Montréal, Tutorials in Quantitative Methods for Psychology 2009, vol 5(2), p. 68-76.

[3] http://step.psy.cmu.edu/materials/manuals/users.pdf

[4] Robert M. Joseph, Brandon Keehn, Christine Connolly, Jeremy M. Wolfe, and Todd S. Horowitz.Why is visual search superior in autism spectrum disorder? Developmental Science, 12(6):1083–1096, 2009.

[5] S. Ogawa, TM Lee, AR Kay, and DW Tank. Brain magnetic resonance imaging with contrast dependent on blood oxygenation. Proceedings of the National Academy of Sciences, 87(24):9868, 1990.

[6] R.S. Menon, J.S. Gati, B.G. Goodyear, D.C. Luknowsky, and C.G. Thomas. Spatial and temporal resolution of functional magnetic resonance imaging. Biochemistry and cell biology, 76(2-3):560–571, 1998.

[7] Robert M. Joseph, Brandon Keehn, Christine Connolly, Jeremy M.Wolfe, and Todd S. Horowitz. Why is visual search superior in autism spectrum disorder? Developmental Science, 12(6):1083–1096, 2009.

[8] 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), 581-590, 2006.

[9] Lorrie Faith Cranor. A framework for reasoning about the human in the loop. In Proceedings of the 1st Conference on Usability, Psychology, and Security, UPSEC'08, pages 1:1–1:15, 2008.

[10] Rachna Dhamija, J. D. Tygar, and Marti Hearst. Why phishing works. In CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems, pages 581{590, 2006.

[11] Devdatta Akhawe, and Adrienne Porter Felt. Alice inWarningland: A Large-Scale Field Study of Browser SecurityWarning Effectiveness.Usenix Security, 2013.

[12] 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)

[13] Joshua Sunshine, Serge Egelman, Hazim Almuhimedi, Neha Atri, and Lorrie Faith Cranor. Crying wolf: An empirical study of ssl warning e ectiveness. usenix security. In Usenix Security Symposium, 2009.

[14] Stuart E. Schechter, Rachna Dhamija, Andy Ozment, and Ian Fischer. The emperor's new securityindicators. In IEEE Symposium on Security and Privacy, 2007.

[15] Batya Friedman, David Hurley, Daniel C. Howe, Edward Felten, and Helen Nissenbaum. 2002. Users' conceptions of web security: a comparative study. In *CHI '02 Extended Abstracts on Human Factors in Computing Systems* (CHI EA '02).

[16] 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).

[17] Cristian Bravo-Lillo, Lorrie Faith Cranor, Julie Downs, Saranga Komanduri, "Bridging the Gap in Computer Security Warnings: A Mental Model Approach," IEEE Security & Privacy, vol. 9, no. 2, pp. 18-26, March/April, 2011.

[18] Adam W. Craig, Yuliya Komarova Loureiro, Stacy Wood, Jennifer M.C. Vendemia (*2012*) Suspicious Minds: Exploring Neural Processes During Exposure to Deceptive Advertising. Journal of Marketing Research: June 2012, Vol. 49, No. 3, pp. 361-372.

[19] Hristo Bojinov. Daniel Sanchez, Paul Reber, Dan Boneh, Patrick Lincoln. Neuroscience Meets Cryptography: Designing Crypto Primitives Secure Against Rubber Hose Attacks. In *21st USENIX Security Symposium*. USENIX Association. August, 2012.

[20] Ivan Martinovic, Doug Davies, Mario Frank, Daniele Perito, Tomas Ros and Dawn Song. On the Feasibility of Side–Channel Attacks with Brain–Computer Interfaces. In *21st USENIX Security Symposium*. USENIX Association. August, 2012.

[21] Ward, B.D. (2000). Simultaneous inference for fMRI data. Available at: http:// homepage.usask.ca/~ges125/fMRI/AFNIdoc/AlphaSim.pdf

[22] Shomstein,Sarah, Cognitive Functions of the Posterior Parietal Cortex: Top-down and bottom-up attentional control, Frontiers in Integrative Neuroscience, Volume 6, July 2012 , ISSN 1662-5145.

[23] DiQuattro, N. E., Geng, J.J.,Contextual knowledge configures attentional control networks, The Journal of neuroscience : the official journal of the Society for Neuroscience, Volume 31, Issue 49, December 2011, Pages 18026-1803.

[24] Huang Mengfei, Bridge Holly, Kemp Martin J, Parker Andrew J, Human cortical activity evoked by the assignment of authenticity when viewing works of art, Frontiers in Human Neuroscience, Volume 5, 2011, Number 00134, ISSN 1662-5161.

[25] Juliana V. Baldo, Nina F. Dronkers, The Role of Inferior Parietal and Inferior Frontal Cortex in Working Memory, Neuropsychology, Volume 20, Issue 5, 2006, Pages 529-538.

[26] Nigel Blackwood, Dominic ffytche, Andrew Simmons, Richard Bentall, Robin Murray, Robert Howard, The cerebellum and decision making under uncertainty, Cognitive Brain Research, Volume 20, Issue 1, June 2004, Pages 46-53, ISSN 0926-6410.

[27] Buckner, Randy L., and Daniel C. Carroll, Self-projection and the brain, Trends in Cognitive Sciences Volume 11 Issue 2, 2007, Pages 49-57.

[28] Burgess GM, Mullaney I, McNeill M, Dunn PM, Rang HP. Second messengers involved in the mechanism of action of bradykinin in sensory neurons in culture. J Neurosci. 1989;9:3314–3325.

[29] Gallagher, HL and Frith, CD (2003) Functional imaging of 'theory of mind'. Trends Cogn Sci 7:77–83.

[30] Gilbert DL, Wang Z, Sallee FR, Ridel KR, Merhar S, Zhang J, et al. Dopamine transporter genotype influences the physiological response to medication in ADHD. Brain.Volume 129, Issue 8, 2006, 2038–2046.

[31] Ramnani, Narender., owen,Adrian M., Anterior prefrontal cortex: insights into function from anatomy and neuroimaging, Nat Rev Neurosci, Volume 5, Issue 3, march 2004, Pages 184-194, ISSN 1471-003X.

[32] Loos M, Pattij T, Janssen MC, Counotte DS, Schoffelmeer AN, Smit AB, Spijker S, van Gaalen MM, Dopamine receptor D1/D5 gene expression in the medial prefrontal cortex predicts impulsive choice in rats, Cerebral cortex (New York, N.Y. : 1991), Volume 20, Issue 5, May 2010, Pages 1064-1070.

[33] Luhmann, J. G.,Curtis, D. W.,Schroeder, P.,McCauley, J.,Lin, R. P.,Larson, D. E.,Bale, S. D.,Sauvaud, J. A.,Aoustin, C.,Mewaldt, R. A.,Cummings, A. C.,Stone, E. C.,Davis, A. J.,Cook, W. R.,Kecman, B.,Wiedenbeck, M. E.,Rosenvinge, T.,Acuna, M. H.,Reichenthal, L. S.,Shuman, S.,Wortman, K. A.,Reames, D. V.,Mueller-Mellin, R.,Kunow, H.,Mason, G. M.,Walpole, P.,Korth, A.,Sanderson, T. R.,Russell, C. T.,Gosling, J. T., STEREO IMPACT Investigation Goals, Measurements, and Data Products Overview, Space Science Reviews, Volume 136, Issue 1-4, April 2008, Pages 117-184.

[34] Sripada, C. S., Gonzalez, R., Phan, K. L., Liberzon, I., The neural correlates of intertemporal decision-making: contributions of subjective value, stimulus type, and trait impulsivity, Human brain mapping, Volume 32, Issue 10, October 2011, Pages 1637-1648.

[35] Frederick, S. Valuing future life and future lives: A framework for understanding discounting. *Journal of Economic Psychology*, 27, 667-680, 2006.

[36] Price, C.J. The anatomy of language: contributions from functional neuroimaging. *Journal of Anatomy,* 197, 335–359, 2000.

[37] Bookheimer, S., Functional MRI of language: New approaches to understanding the cortical organization of semantic processing, Annual Review of Neuroscience, Volume 25, 2002, Pages 151-188.

[38] Friederici, A.D., Towards a neural basis of auditory sentence processing. *Trends in Cognitive Sciences,* 6, 78–84, 2002.

[39] Hagoort, P. On Broca, brain, and binding: a new framework. *Trends in Cognitive Sciences,* 9, 416-423, 2005.

[40] Krueger RF, Markon KE, Patrick CJ, Iacono WG. Externalizing psychopathology in adulthood: dimensional-spectrum conceptualization and its implications for DSM-V. *Journal of Abnormal Psychology,* 114:537–550, 2005.

[41] Swann AC, Lijffijt M, Lane SD, Steinberg JL, Moeller FG. . Trait impulsivity and response inhibition in antisocial personality disorder. *Journal of Psychiatry Research,* 43:1057–1063, 2009.

[42] Mortensen M, Ebert B, Wafford K & Smart TG, Distinct activities of GABA agonists at synaptic- and extrasynaptic-type GABAA receptors, The Journal of Physiology, Volume 588, 2010, Pages 1251-1268.

[43] Moreno-López L, Catena A, Fernández-Serrano MJ, Delgado-Rico E, Stamatakis EA, Pérez-García M, Verdejo-García A. (2012). Trait impulsivity and prefrontal gray matter reductions in cocaine dependent individuals. Drug Alcohol Depend. 2012 Oct 1;125 (3):208-14.

[44] Ernst, M. & Paulus, M.P. Neurobiology of decision making: a selective review from a neurocognitive and clinical perspective. *Biological Psychiatry*, 58, 597–604, 2005.

[45] Sanfey, A.G., Rilling, J.K., Aronson, J.A., Nystrom, L.E. & Cohen, J.D. The neural basis of economic decision-making in the Ultimatum Game. Science, 300, 1755–1758, 2003.

[46] Miller, E. K., & Cohen, J. D. An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, 24, 167-202, 2001.

[47] Ponnurangam Kumaraguru, Steve Sheng, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. Teaching johnny not to fall for phish. ACM Trans. Internet Technol., 10(2):1{31, 2010.

[48] Steve Sheng, Bryant Magnien, Ponnurangam Kumaraguru, Alessandro Acquisti, Lorrie Faith Cranor, Jason Hong, and Elizabeth Nunge. Anti-phishing phil: the design and evaluation of a game that teaches people not to fall for phish. In SOUPS '07: Proceedings of the 3rd symposium on Usable privacy and security, pages 88{99, 2007.

[49] Sukamol Srikwan and Markus Jakobsson. Using cartoons to teach internet security. Cryptologia, 32(2):137{154, 2008.

[50] Alain Forget, Sonia Chiasson, and Robert Biddle. Lessons from brain age on password memorability (poster). In Future Play '08: Proceedings of the 2008 Conference on Future Play, pages 262{263, 2008.

[51] Rosen BR, Buckner RL, Dale AM. Event-related functional MRI: past, present, and future. Proc Natl Acad Sci USA 95:773– 780, 1998.

[52] Julie Thorpe, P. C. van Oorschot, and Anil Somayaji. Pass-thoughts: authenticating with our minds. In the workshop on New security paradigms (NSPW '05), 2005.

[53] John Chuang, Hamilton Nguyen, Charles Wang, and Benjamin Johnson. I Think, Therefore I Am: Usability and Security of Authentication Using Brainwaves. In Workshop on Usable Security (USEC), 2013.

[54] Oldfield, R.C. The assessment and analysis of handedness: The Edinburgh inventory. Neuropsychologia, 9, 97-113, 1971.

[55] Desmond, J.E., Glover, G.H. Estimating sample size in functional MRI (fMRI) neuroimaging studies: statistical power analyses. Journal of Neuroscience Methods, 118, 115–128, 2002.

[56] Murphy, K., & Garavan, H. An Empirical Investigation into the number of subjects required for an event-related fMRI study. Neuroimage, 22, 879-885, 2004.

[57] Cox, R.W. AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, 29, 162–73,1996.