



Understanding the Effects of Cyber Attacks on Human Operators

Leanne Hirshfield
SYRACUSE UNIVERSITY

07/13/2018
Final Report

DISTRIBUTION A: Distribution approved for public release.

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE					Form Approved OMB No. 0704-0188	
The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing the burden, to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.						
1. REPORT DATE (DD-MM-YYYY) 03/15/2018		2. REPORT TYPE Final Report			3. DATES COVERED (From - To) 10/31/14-10/31/17	
4. TITLE AND SUBTITLE Understanding the Effects of Cyber Warfare on Human Operators				5a. CONTRACT NUMBER		
				5b. GRANT NUMBER FA9550-15-1-0021		
				5c. PROGRAM ELEMENT NUMBER		
6. AUTHOR(S) Leanne Hirshfield				5d. PROJECT NUMBER		
				5e. TASK NUMBER		
				5f. WORK UNIT NUMBER		
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Syracuse University					8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Office of Scientific Research					10. SPONSOR/MONITOR'S ACRONYM(S) AFOSR	
					11. SPONSOR/MONITOR'S REPORT NUMBER(S)	
12. DISTRIBUTION/AVAILABILITY STATEMENT approved for public release						
13. SUPPLEMENTARY NOTES						
14. ABSTRACT The goal of this YIP research effort was to develop custom machine learning classifiers that take input from fNIRS and other physiological sensors in order to predict the changing mental states of computer operators (e.g., predict cognitive load, levels of trust and suspicion, stress). We tested our models in a variety of use case scenarios that are relevant to the cyber domain, in order to better understand the cognitive and emotional states that are related to operator performance in the presence of cyber threats. Over the course of the effort we collected fNIRS data on 160 human subject participants, created and shared custom machine learning code with AFRL and academic colleagues, and published 15 research articles.						
15. SUBJECT TERMS						
16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON	
a. REPORT	b. ABSTRACT	c. THIS PAGE			19b. TELEPHONE NUMBER (Include area code)	

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATE COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33315-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report. e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

Understanding the Effects of Cyber Warfare on Human Operators

PI: Dr. Leanne Hirshfield
Associate Research Professor
SI Newhouse School
Syracuse University
lmhirshf@syr.edu

ABSTRACT

The goal of this YIP research effort was to develop custom machine learning classifiers that take input from fNIRS and other physiological sensors in order to predict the changing mental states of computer operators (e.g., predict cognitive load, levels of trust and suspicion, stress). We tested our models in a variety of use case scenarios that are relevant to the cyber security domain, in order to better understand the cognitive and emotional states that are related to operator performance in the presence of cyber threats. Throughout the course of the effort we collected fNIRS and other physiological sensor data on roughly 160 human subject participants. All data has been saved and formatted in a way that it has become invaluable for training new generalizable models for robust predictions of mental state. We also developed a suite of preprocessing and machine learning techniques in Matlab and Python throughout this effort. The code has been shared with academic partners at AFRL, Charles River Analytics, and Aptima Inc. in order to help advance state-of-the-art of fNIRS-based predictive modeling. Fifteen peer-reviewed publications have resulted from this research (listed in Appendix A), and the work produced during this effort was directly responsible for a new grant from the National Science Foundation entitled ‘Applying Advanced Deep Learning Techniques to Functional Near-Infrared Spectroscopy Data for Improved Cognitive and Emotional State Classification Across Participants’.

INTRODUCTION

Joint Force Publication 3-13 defines the information environment as the “aggregate of individuals, organizations, and systems that collect, process, disseminate, or act on information. This environment consists of three interrelated dimensions which continuously interact with individuals, organizations, and systems. These dimensions are the physical, informational, and cognitive.” JP 3-13 describes the cognitive dimension as the most important of the three.

“The cognitive dimension encompasses the minds of those who transmit, receive, and respond to or act on information. It refers to individuals’ or groups’ **information processing, perception, judgment, and decision making**. These elements are influenced by many factors, to include individual and cultural **beliefs, norms, vulnerabilities, motivations, emotions**, experiences, morals, education, **mental health**, identities, and **ideologies**. Defining these influencing factors in a given environment is critical for understanding how to best influence the mind of the decision maker and create the desired effects. As such, this dimension constitutes the most important component of the information environment [emphasis added].”

Despite the recognized importance that the cognitive dimension plays in information operations, the scope of cyber warfare research to date has focused on the effects of cyber techniques on the target computer

system. Little empirical work has been done to explore, articulate, and measure the effects that cyber warfare techniques have on the cognitive dimension of human operators. In particular, the U.S. Air Force has invested considerable resources on defensive and offensive cyber warfare techniques. The development of D5 (Deny, Disrupt, Deceive, Degrade, Destroy) capabilities has resulted in a large collection of tools which can be used by cyber-operators to conduct offensive operations. Also, there is a reasonable expectation that many of the D5 techniques contained in our cyber arsenal may mirror the cyber capabilities of adversarial operators who seek to infiltrate US computer systems and target our personnel. There is the potential for a dramatic payoff by undertaking a formal evaluation of these D5 capabilities based on how they affect both adversarial operators (offensively) as well as blue force operators under cyber attack (defensively).

With these limitations and goals in mind, our YIP proposal outlined our high level research goals:

1. Develop machine learning classifiers that take input from behavioral and physiological sensors that can:
 - a) Predict when a cyber attack is occurring in real-time even if the blue force operator remains unaware of the attack.
 - b) Predict the changing mental state (e.g., predict cognitive load, levels of trust and suspicion, affect, etc...) of a computer user during a cyber attack in order to assess the effect of a given D5 technique on a target operator in real-time.

RELATED TEAM EXPERTISE AND FACILITIES

We are well positioned to carry out this research, as Dr. Hirshfield and her collaborators have already established foundational research in the social science and human-computer interaction (HCI) domain relating to defining, manipulating, and measuring relevant user states such as trust, suspicion, workload, and emotional states [1-11]. We also have the facilities and physiological measurement equipment needed to carry out the proposed work. The 3,000 square foot Newhouse MIND Lab features configurable data collection set-ups that include (i) customizable computer workstations and cubicles designed to run experiments with many participants, and (ii) a separate experimentation room designed for running experiments on individuals using some subset of the lab's cognitive, physiological, and behavioral sensors available in the lab. The lab contains over \$500K of non-invasive cognitive, physiological, and behavioral measurement devices. This includes a 52-channel functional near-infrared spectroscopy (fNIRS) device from Hitachi Medical (ETG 4000), Advanced Brain Monitoring's b-alert wireless EEG, faceLab and Tobii eyetrackers, and a Biopac MP150 sensor suite equipped with EEG, galvanic skin response, respiration, and ECG sensors. The lab also contains large-format new media technology, virtual reality, and augmented reality equipment.

Notably, our prior research has included an extensive exploration into the suspicion domain [1]. In our recent review of the minimal suspicion literature, Bobko, Barelka, and Hirshfield (2014) synthesized efforts across the social sciences, including management, marketing, communication, human factors, and psychology. Similar to Levine and McCornack (1991), they defined suspicion as the simultaneous occurrence of three factors: (i) uncertainty, (ii) malintent, and (iii) cognitive activation (e.g., thoughts about possible motivations underlying observed anomalies). We therefore defined "state suspicion" as the simultaneous occurrence of these three components. Our analysis also suggested that trust inhibits

suspicion, while distrust can be a catalyst of state-level suspicion. Based on a three-stage model of state-level suspicion (Figure 1), associated research propositions and questions were developed.

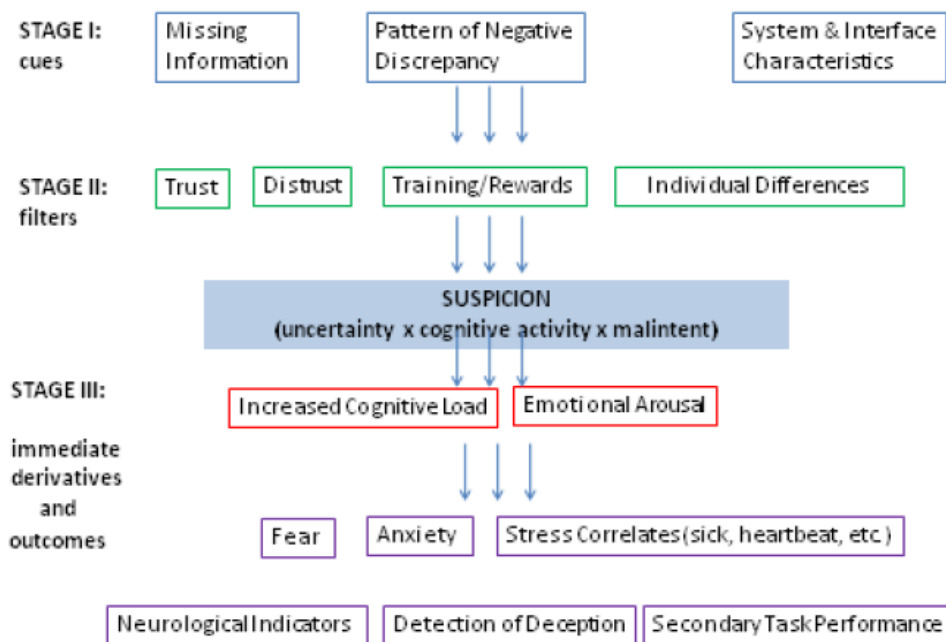


Figure 1: Our three stage model of suspicion [1].

These propositions and questions were intended to help guide future work on the measurement of suspicion (self-report and neurological), as well as the role of the construct of suspicion in models of decision-making and detection of deception[1].

REVIEW OF YEAR 1 RESEARCH (October 2014-October 2015)

The first year of this effort included three primary tasks. These include 1) developing the software and hardware synchronization techniques for syncing up the physiological sensors in the lab with the STIM computer, 2) developing machine learning code for analyzing cognitive and physiological data from our experiments, and 3) running three separate data collections in order to test our ability to predict emotion and suspicion using fNIRS (and other sensors). Specifically, the data collection during the first year resulted in fNIRS data for 48 participants during two studies designed to elicit emotional state changes and one study designed to manipulate user states of trust and suspicion.

TASK 1: Technical work to synchronize lab sensors

Before beginning data collection, we worked to synchronize our EEG, fNIRS, and Biopac (GSR, ECG, Respiration) sensors with the STIM computer. We developed a hardware/software set-up that involves using an Arduino processor connected to the COM ports of the STIM machine and the measurement sensors. The STIM machine calls custom batch files to send marks (we select the mark 'number' in the batch file) to the sensors.

TASK 2: Machine learning code development

We developed code in python to take our fNIRS data and conduct dimensionality reduction on our channels into Regions of Interest (ROI), as has been done in prior research applying machine learning on brain data [12]. For each ROI our code computes each of these features listed below separately for the (i) first half (ii) second half, and (iii) for the total of each task:

- ☐ **Full-width-at-half max:** *The time difference between the two points where the signal is at half of its maximum value for the data from each 60-second-long video second.*
- ☐ **Slope:** *Slope calculated between the start and ending values of the signal*
- ☐ **Mean:** *Average Signal Value*
- ☐ **Max:** *Maximum Signal Value*
- ☐ **Min:** *Minimum Signal Value*

For each of the first half, second half, and total chunks of time series data noted above, we also further split that data into six equal segments of time and we took average values across those segments:

- ☐ **Piecewise Mean 1:** *Average Signal Value of Segment 1*
- ☐ **Piecewise Mean 2:** *Average Signal Value of Segment 2*
- ☐ **Piecewise Mean 3:** *Average Signal Value of Segment 3*
- ☐ **Piecewise Mean 4:** *Average Signal Value of Segment 4*
- ☐ **Piecewise Mean 5:** *Average Signal Value of Segment 5*
- ☐ **Piecewise Mean 6:** *Average Signal Value of Segment 6*

We implemented a wrapper feature selection scheme (using the *filterbestn* function in the Python scikit learn toolkit) in order to select the top 100 (or an n of our choosing) features that have the highest information gain for model training. We then built in capabilities to train a support vector machine (SVM), Naïve Bayes classifier, and K-nearest neighbor model on participants' resulting data using a *within-participant blocked cross-validation* scheme and a *leave-one-participant-out cross-validation scheme*, which is done to prevent any model overfitting to any one participant's brain. Our code then stores the resulting accuracy, area under the curve (AUC), and F1-scores, sensitivity, and specificity for these tests.

TASK 3: Data Collections

Once the sensor synchronization and machine learning code was in place, much of this year focused on the design and collection of data for three experiments using our lab's sensors. As noted above, our primary goal is to predict the presence of mental states such as trust, suspicion, and their related cognitive and emotional correlates. As noted in our prior review on suspicion in the social sciences, we proposed a definition of state suspicion as a "*simultaneous state of cognitive activity, uncertainty, and perceived malintent about underlying information*" from an external agent (p. 493). Note that a key component of suspicion is uncertainty – and it occurs with a negative frame ("concern") that is linked to cognitive load regarding the uncertainty (e.g., what are the possible meanings of the unusual events being observed). We suggest that fNIRS can be used to measure the neural correlates of suspicion, and the inversely related construct of trust. However, we also caution that the accurate measurement of emotional states and

cognitive load is essential to measuring trust and suspicion, as these mental states are direct neural correlates of trust and suspicion. Therefore, our data collection efforts incorporate experiments that focus on low level emotional state measurements, measurements of a variety of types of cognitive load, as well as more complex and ecologically valid experiments focused on measuring suspicion and trust.

Experiment 1: Measuring Emotion via Music Videos

Protocol: Our goal in this experiment was to induce a variety of emotional states in participants while measuring the hemodynamics of their brain with fNIRS, and to demonstrate the use of the resulting fNIRS brain data to identify emotional state. A subset of music videos from the Database for Emotional Analysis using Physiological Signals (DEAP) database [13] were selected as stimuli to elicit participants' emotions. As described before, the DEAP dataset stimuli materials consisted of 40 music videos that had been found to induce consistent self-report scores in the outer boundaries of the four quadrants of the circumplex model (Figure 2), representing four experimental conditions of High Valence Low Arousal (HVLA), High Valence High Arousal (HVHA), Low Valence High Arousal (LVHA), and Low Valence Low Arousal (LVLA). We selected fifteen of these videos-- three videos to represent each of the five conditions. Videos were intentionally selected to maximize the expected emotional reaction of participants; that is, the videos were handpicked from the results reported by Koelstra et al. that were as far away from circumplex model's "neutral" center as possible. The purpose of this selection was to ensure that each participant's brain state was maximally representative of the quadrants in the valence/arousal space. The protocol followed a block design format. The music videos were separated into three blocks, each containing videos from the DEAP dataset with five unique emotion labels, and included the conditions of Low Valence/Low Arousal (LVLA), High Valence/High Arousal (HVHA), Low Valence High Arousal (LVHA), High Valence/Low Arousal (HVLA), and Neutral Valence, Neutral Arousal. A Note that the 'neutral' condition was included with music videos that were found by Koelstra et al. to be neutral on both the valence and arousal dimensions. The order within blocks was selected to ensure that within each block of videos the stimuli were presented in a random manner to the participant (so that the participants would not be able to easily guess which type of video would be played next, and to avoid the possible confounding effects of having the same-emotion-inducing video in a row, which would result in an intensified emotion effect), while still ensuring that each block in the experiment contained one video from each of the five conditions above.

Table 1: Block design of experiment. Videos within each block were randomized.

Music video block 1	
1.	A Fine Frenzy, <i>Almost Lover</i> : Low Valence Low Arousal (LVLA)
2.	Black Eyed Peas, <i>My Humps</i> : High Valence Low Arousal (HVLA)
3.	Blur, <i>Song 2</i> : High Valence High Arousal (HVHA)
4.	Smashing Pumpkins, <i>1979</i> : Neutral (NN)
5.	Stigmata, <i>В отражении глаз</i> : Low Valence High Arousal (LVHA)

Music video block 2

1. Sia, *Breathe Me* : Low Valence Low Arousal (LVLA)
2. Christina Aguilera, Lil' Kim, Mya, Pink, *Lady Marmalade*: High Valence High Arousal (HVHA)
3. Napalm Death, *Procrastination on the Empty Vessel*: Low Valence High Arousal (LVHA)
4. Madonna, *Rain*: Neutral (NN)
5. Taylor Swift, *Love Story*: High Valence Low Arousal (HVLA)

Music video block 3

1. Glen Hansard, *Falling Slowly*: Neutral (NN)
2. White Stripes, *Seven Nation Army* (HVHA)
3. Trapped Under Ice, *Believe*: Low Valence High Arousal (LVHA)
4. Wilco, *How to Fight Loneliness*: Low Valence Low Arousal (LVLA)
5. Louis Armstrong, *What a Wonderful World*: High Valence Low Arousal (HVLA)

Twenty healthy, college age participants from a university in the Northeast took part in the experiment (13 male, 7 female). Upon arrival to the lab, participants provided informed consent and completed a pre-questionnaire to obtain their demographic data. Then they were provided with instructions explaining the different experiment and how to fill out the post task surveys.

Equipment Set-Up: The experiment was performed in a controlled laboratory environment. The fNIRS signals from participants' brains were recorded using a Hitachi ETG-4000 fNIRS device with a sampling rate of 10Hz. The device provides 52 channels of brain activity data from the frontal region of the participant's brain. Each participant was seated on the experiment chair and the chair was adjusted to his or her comfort level. The fNIRS probe (Figure 2) was a 3x11 probe with 17 light sources and 16 detectors, resulting in 52 locations measured on the head. Once the fNIRS probe was in place, a 3d digitizer was used to record the locations of each fNIRS channel on that subject's head.

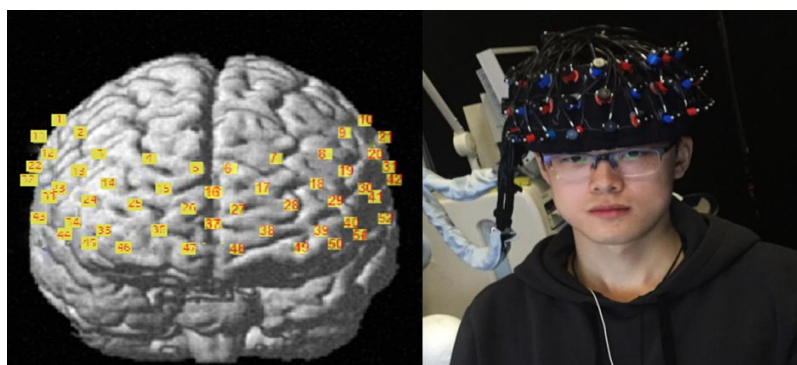
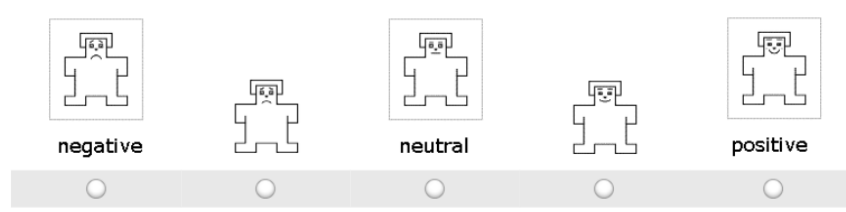


Figure 2: A participant wearing the fNIRS sensors

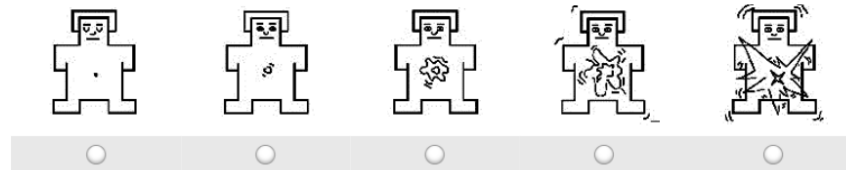
Experiment 2: Measuring emotion via advertisements

Protocol: The stimuli in this experiment consisted of advertisements in a video format, ranging from 15 seconds to 1 minute. The ads were selected by a domain expert¹ to elicit a range of viewer emotions, and videos were pilot tested on an introduction to Mass Communications class at the Newhouse School to ensure they elicited a range of emotions. When presenting the advertisements to participants, two different presentation orders were used, in order to reduce order effects. After watching each ad, participants filled out a series of survey questions designed to self-assess their reaction to the ad. They then rested for 15 seconds to allow their brains to return to baseline, before viewing the next ad. The survey questions are shown in Figure 3. The first two questions are the Self-Assessment Manikin Survey instrument [14], designed to assess valence (happy-unhappy scale) and arousal (excited vs calm dimension) in participants. The third question was designed to gauge overall effectiveness of the advertisement.

happy-unhappy scale



excited vs. calm dimension



Please indicate overall effectiveness of the advertisement

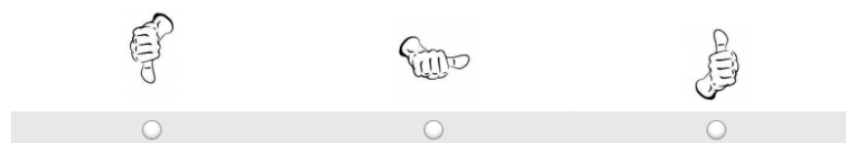


Figure 3: Survey items administered after each advertisement.

Eighteen participants participated in this experiment (9 female). The participants each viewed a total of 30 advertisements from 26 different brands, with study participants wearing the fNIRS device. Upon arrival to the lab, participants provided informed consent and completed a pre-questionnaire to obtain their demographic data. Then they were provided with instructions explaining the different experiment and how to fill out the post task surveys.

¹ The domain expert was a prior CEO at one of the top advertising firms in the country.

Equipment Set-Up: fNIRS data was collected using the Hitachi ETG-4000 near-infrared spectroscopy device with a sampling rate of 10Hz. As shown in Figure 4, a custom probe design, measuring 46 channel locations, was created to cover the frontal cortex, right temporoparietal junction (TPJ), and left TPJ. Once the probe was placed on each participant's head, a Patriot Polhemus 3d digitizer was used to measure the location of each source/detector on that participant's brain.



Figure 4: The custom probe design placed on participants was designed to cover the frontopolar and orbitofrontal brain regions, as well as the temporoparietal junction.

Experiment 3: Measuring suspicion via a team-work scenario and deception manipulation

The final data collection we ran this year was designed to attempt our ability to predict suspicion in a more ecologically valid setting than that used in experiments 1 and 2. As noted above, in our prior review on suspicion in the social sciences, we collated the definitions of suspicion that have been proposed in the literature [15]. This synthesis led to the definition of state suspicion as a “*simultaneous* state of cognitive activity, uncertainty, and perceived malintent about underlying information” from an external agent (p. 493). Note that a key component of suspicion is uncertainty – and it occurs with a negative frame (“concern”) that is linked to cognitive activity regarding the uncertainty (e.g., what are the possible meanings of the unusual events being observed). We suggest that fNIRS can be used to measure the neural correlates of suspicion. Based on the theoretical definition of suspicion as a simultaneous state of cognitive activity, uncertainty, and perceived malintent about underlying information from an external agent, we hypothesize that suspicion, relative to trust, will be associated with three measurable facets:

H1: Suspicion will be associated with an increase in cognitive load. More specifically, suspicion will be associated with increased brain activation in the frontopolar region and the orbitofrontal cortex.

The uncertainty of a suspicion-inducing situation will cause individuals to devote cognitive resources to generating hypotheses about the true nature of the situation.

H2: Suspicion will be associated with an increase in activation in the temporoparietal junction (TPJ), which is involved in thinking about other people's intentions. A suspicious individual will devote cognitive resources to interpreting the 'intentions' of the other individual, and consistent with ToM, this will be associated with an increase in activation in the TPJ.

H3: Suspicion will be associated with an increase in negative affect, causing activation in the dorsolateral prefrontal cortex (DLPFC), which is involved in regulation of negative emotion. A suspicious individual will have concern about the intent of the other individual in the interaction, leading to a negative emotional reaction that will involve emotion regulation in the DLPFC.

Protocol: To test these hypotheses, an experiment was designed to induce suspicion (and, in contrast, trust) in participants while their brain activity was measured using fNIRS. Participants worked on a task called the 'Winter Survival Problem'. This problem is a variation of a commonly used teamwork exercise called the Desert Survival Problem [16]. Traditionally, teams are given a hypothetical scenario where their plane has crash-landed in the middle of dense wilderness, with very few survivors. Team members are given a list of items that survived the crash. They are asked to discuss the list of items among themselves and to create their own list prioritizing each item based on its usefulness in helping the survivors to endure the elements and find, or attract, help. In our experiment participants were told that they would be working together in dyads. They were introduced to their research partner, who was actually a research confederate. Participants were told that their partner had been given 30 minutes prior to the experiment to use the internet and look up the items on the list in terms of their usefulness for survival. During the experiment, the participant and his/her partner would be given a list of three items, randomly generated from the list of items in Table 2. Their goal was to work together (although, as noted below, their interactions were controlled via video-feeds) to order the three items from greatest priority to least priority. They were given an added monetary incentive for placing items in the correct prioritized order (assessed via an answer key created by survival experts from the military).

Table 2: List of the 25 items that might be available after the plane crash.

Compress kit (with 28 ft. of 2-inch gauze)	30 feet of rope
Cigarette lighter without the fluid	Flashlight with batteries
Newspaper (one each)	Extra shirt & pants (one each)
Two ski poles	Individual bottle of water (one each)
Sectional air map made of plastic	1 box of plastic garbage bags
Family-sized chocolate bar (one each)	A mirror
Quart of 85-proof whiskey	Matches
Can of shortening	Crackers (one each)
20 x 20 ft. piece of heavy-duty canvas	Saw
Individual bag of peanuts (one each)	Hat and mittens (one each)
Ball of steel wool	1 pair of reading glasses
Loaded .45-caliber pistol	Knife
Compass	

Each participant was equipped with the fNIRS headset during the experiment. While wearing the fNIRS, participants viewed a series of 75-second long web-videos. Each web-video was created by the partner. During each video, the partner described the three items, and explained her rationale for ordering/prioritizing the three items to maximize survival. After each video ended, the participant had the opportunity to make a final list of the items in order of priority (i.e., using or ignoring the partner's advice). The participant then filled out a series of surveys (described in detail next) to self-assess her/his cognitive and emotional state while watching the video. Then a new video and task would begin.

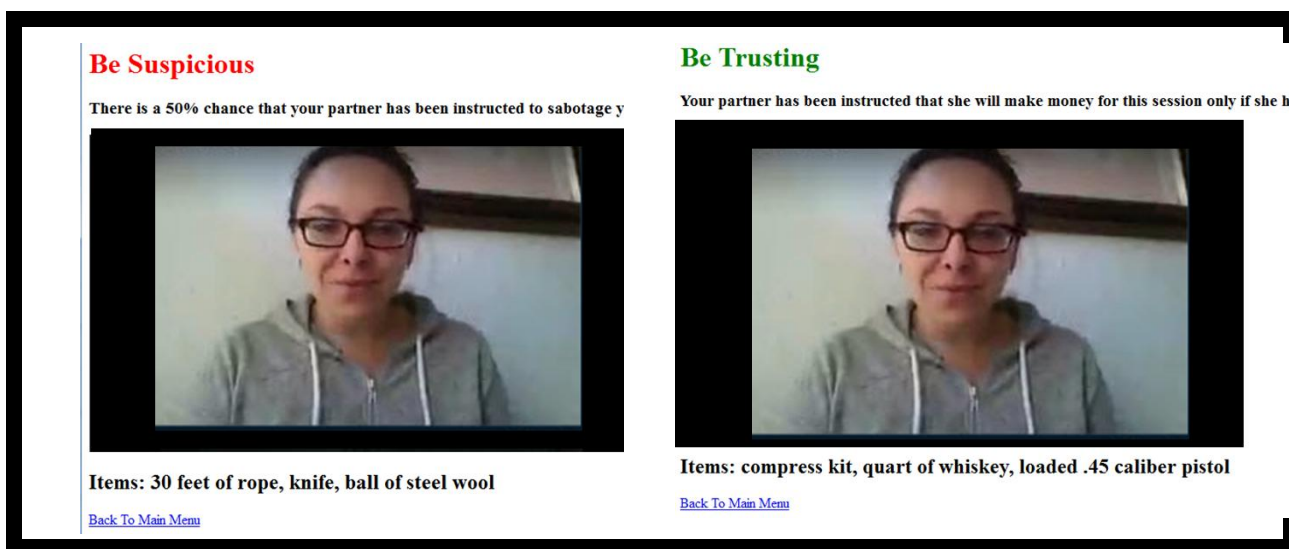


Figure 6: The two conditions represented in the experiment. The only modification was the instructions given to participants regarding the veracity of their partner's current information.

The two states of trust and suspicion were manipulated via instructions given to the participants on their computer screen above the partner's video feed (as shown in Figure 6). During the suspicion condition participants were told: *"There is a 50% chance that your partner has been instructed to sabotage you, convincing you to make an incorrect order of items. In this type of session, your partner will receive extra compensation for every item that you have wrong in the list. Of course, there is a 50% chance that your partner is not sabotaging you during this time, and is behaving like (s)he would during the 'be trusting' scenario."* During the 'trust' condition, participants were told: *"Your partner is always instructed to do the best job that he or she can to help you maximize accuracy. You and your partner will both receive the same amount of \$\$ at the end based on your accuracy."* Although the participant saw instructions about the veracity of their partner's advice, in actuality, the partner was **not** instructed to deceive the participant (i.e., the partner was always telling the truth), ensuring that the only difference between the two conditions was the framing of the instructions to participants on whether or not to be suspicious.

Ten participants took part in the experiment (6 female), all of whom were undergraduate students from a university in the Northeast. After giving informed consent and going over experiment instructions, the fNIRS device was placed on their heads. Each participant completed the task a total of 12 times. The

experiment followed a randomized block design format, where one example of each condition was randomly presented in each block. For example, the ordering of conditions for the first participant was: [TS|ST|ST|TS|ST|TS], where T = trust and S = suspicion condition. After watching each video clip, participants were asked to list the three items in order from highest to lowest priority toward survival. Next, they filled out the NASA Task Load Index (TLX), which is a five-item self-report survey used for self-assessment of workload [17], and the Self-Assessment Manikin (SAM), a self-assessment of emotional state based on pictorial representations of the valence and arousal dimensions of emotion [14]. Emotional state is often assessed on two dimensions - valence and arousal. Valence describes the positive or negative feeling associated with a stimulus, while arousal indicates the level of excitedness associated with that feeling.

Equipment Set-Up: fNIRS data was collected using the Hitachi ETG-4000 near-infrared spectroscopy device with a sampling rate of 10Hz. We used the custom probe design from experiment 2 (Figure 4) in this experiment. This probe design, measuring 46 channel locations, was created to cover the frontal cortex, right TPJ, and left TPJ. Once the probe was placed on each participant's head, a Patriot Polhemus 3d digitizer was used to measure the location of each source/detector on that participant's brain.

Experiment 4: Measuring the effects of phishing detection and malware warnings on computer users using EEG, eyetracking, and behavioral data

In this experiment, which is detailed in our recent publication [18], we pursued a comprehensive three-dimensional study of phishing detection and malware warnings, focusing not only on what users' task performance was but also on how users processed these tasks based on: (1) neural activity captured using Electroencephalogram (EEG) metrics, and (2) eye gaze patterns captured using an eyetracker. Our primary novelty lies in employing multi-modal neurophysiological measures in a single study and providing a near realistic set-up. Specifically, in the context of phishing detection, we showed 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 showed that users are frequently reading, possibly comprehending, and eventually heeding the message embedded in the warning. This publication is included as a supplementary upload with our annual report [18].

REVIEW OF YEAR 2 RESEARCH (October 2015-October 2016)

Building on our research from Y1, the second year of this effort included four primary tasks. This included 1) finishing the results and analysis on the data acquired during Y1 and writing manuscripts for publication 2) re-orienting our technical approach based on results thus far and other recent findings in the domain of machine learning on brain data, 3) designing and collecting data for five new experiments based on our re-oriented approach, and 4) updating our machine learning code to do data fusion on data from multiple sensors.

TASK 1: Finishing Data Analysis and Preparing Manuscripts for Year 1 fNIRS Studies

During the past year we submitted three manuscripts for publication, based on the fNIRS data collected in the three data collection efforts from Year 1. **All three of these manuscripts have been included as supplementary material with this report.** These include machine learning results on data from:

- The experiment measuring emotion with fNIRS using music videos from the DEAP dataset. This manuscript (lead author Bandara) was submitted to International Journal of Human Computer Studies. At the time of this report, Journal reviewers have requested minor revisions to the paper, which we are currently making.
- The experiment measuring emotion with fNIRS using advertisements. This manuscript (lead author Sommer) was submitted to IEEE transactions on Affective Computing. At the time of this report, we are waiting the outcome of the review process.
- The experiment measuring trust and suspicion with fNIRS during a teamwork scenario (lead author Hirshfield). This was submitted to the Journal of Social, Cognitive, and Affective Neuroscience. At the time of this report, we recently heard back from the editor of SCAN that the topic of the paper was not the right fit for the Journal. Based on this feedback, we are revising the paper to submit to a venue more aligned with the human-computer interaction domain.

TASK 2: Re-Orienting Technical Approach Based on Results Thus Far

Our high level goal is to make accurate, real-time predictions of people's cognitive and emotional states, with an emphasis on people's levels of trust and suspicion while they experience cyber attacks. Although the machine learning results from our three manuscripts in Task 1 show promising accuracies, we are becoming concerned about the issues relating to model overfitting in the applied neuroscience domain. Recent advancements in biotechnology have been the catalyst for a dearth of research applying machine learning techniques to non-invasive brain measurement devices in order to predict cognitive and emotional states [6, 19-22]. Although many early successes were achieved using machine learning on brain data, several notable challenges have arisen, which significantly limit the impacts of these early successes. In particular, much early research trained models per individual, on very small datasets, which led to model overfitting and inflated accuracies. When these models were tested on new participants, or even on the same participant during a different measurement session, the model performance degraded significantly. Therefore, we view our successes from our year 1 studies with the knowledge that it is quite possible that these models would degrade in performance when predicting the states of interest in different task environments, and on different users, or even on the same users during different sessions.

Therefore, we shifted our approach applying machine learning on brain data during Y2, with the belief that we need to collect data on many participants, across a range of tasks, so that we can adequately train models that do not overfit to any given participant or task type. Therefore, the second year of this effort focused heavily on data collection (lots of participants, across a range of tasks), and on updating our code to take in a range of physiological sensor formats.

TASK 3: Data Collections

During Year 2 we conducted four data collection efforts. Three of the experiments used just fNIRS, while one of the experiments used fNIRS and EEG. We have submitted manuscripts for two of these data collections, and the manuscripts are included as supplementary material with this report.

Experiment 1: Using fNIRS to Explore the Neural Underpinnings of Website Legitimacy and Familiarity Detection.

This experiment involved measuring users' brains with fNIRS while they surfed the web and experiences mock phishing attaches and malware warnings. We used our custom machine learning code developed through this research effort to analyze our results. See supplementary manuscript (Neupane, Saxena, Hirshfield, 2016) for detailed information.

Experiment 2: Neural Fingerprints of Vocal Security: An fNIRS Study of Speaker Authenticity and Familiarity

In this experiment we measured people's brains with fNIRS while they engaged in a speaker verification experiment, which was designed to manipulate trust and suspicion of caller's voices. We have attached the manuscript created for this experiment as well. The manuscript was recently rejected from the USENIX security conference, and we are re-working it based on reviewer feedback for submission elsewhere. See supplementary manuscript (Neupane, Saxena, Hirshfield, 2016) for detailed information.

Experiment 3: Traditional Workload Experiment

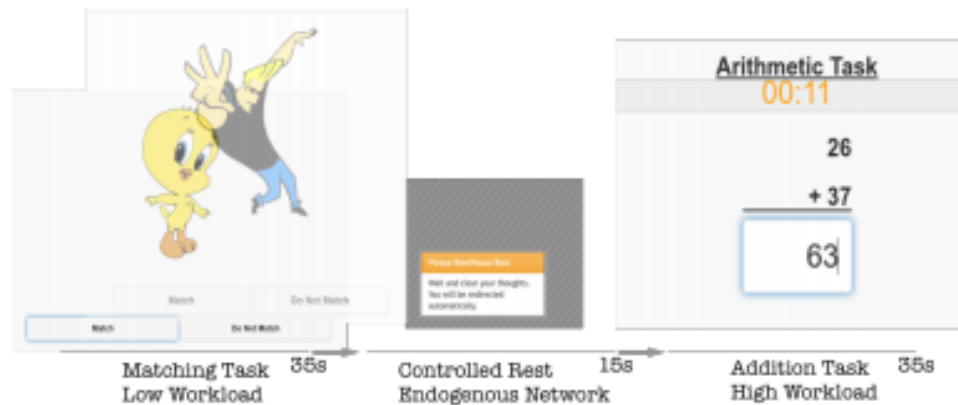
Much of Dr. Hirshfield's prior research involved workload measurement with fNIRS. Cognitive load is essential for successfully measuring cyber users' experiences while engaged in cyberwarfare. With the goal of building a large database on a range of tasks, this experiment involved data collection on 20 participants, where participants used an easy, medium, and difficult version of tetris, as well as an easy, medium, and difficult version of the Multi-Attribute Task Battery. In addition to recording participants' accuracy, we also administered the Nasa TLX and Self Assessment Manikin after each task, to gauge participants' self report workload and emotional state during each task.

At the time of this report the results from this data collection have yet to be analyzed.

Experiment 4: Traditional Workload Experiment focused on Machine Learning and Multiple Sessions with fNIRS and EEG

The EEG used in this experiment was Advanced Brain Monitoring's b-alert X10, a 9 channel wireless EEG system with a linked mastoid reference, sampling at 256hz. The EEG headset was placed on users using standard 10-20 measurement set-up techniques. The fNIRS device used in this experiment was the Hitachi ETG-4000 fNIRS device with a sampling rate of 10Hz. The fNIRS probe was a 3x11 probe with 17 light sources and 16 detectors, resulting in 52 locations measured on the head. 3.2 Method Twenty-three subjects (8 female) between the ages 18 and 49 participated in the experiment. Upon arrival, subjects consented to the experiment and were fitted with the fNIRS or EEG sensors. The prepackaged b-alert and Hitachi software calibrated itself to the detected connection with the user's scalp. Then, the subject alternated between 8 instances of an arithmetic task and 8 instances of an image-matching task, performing each task for 35 seconds, with 15-second controlled rest periods in between. For the image

matching task, users indicated whether sequences of images matched each other, as in an n- with n permanently set to 1, similar to the low cognitive workload condition used in previous implicit BCI work (Afergan et al., 2014a). For the arithmetic task, users added two two-digit integers to each other, entering the response into a text-box. We included workload in two separate modalities in order to, potentially, induce two separate states of workload.



In the figure above, we show each of these experimental conditions. We were thus interested in the transition between task and rest, and how that would change in the first and second session for both EEG and fNIRS. Our interest in inter-session comparison was born out of a consideration for how machine learning algorithms might decay over time if they do not account for updates to the user's cognition. We were especially interested in corroborating a previous small long term pilot (Hincks et al., 2016), where we tracked one of the author's fNIRS data over the course of several months as he made himself an expert at the cognitive workload tasks typically used in implicit BCI (the nback). In later sessions, we noticed that fNIRS failed to register a strong effect unless the difficulty placed him at the edge of his ability, and we hypothesized that the task had over time generated too efficient topdown schemes for solving it, thereby leaking less prediction error, the information theoretical construct to which we hypothesize these cortically-sensing instruments are principally sensitive. If that is the case, then machine learning algorithms operating from either EEG or fNIRS data should degrade in a second session of the experiment. The experiment was thus repeated four times in two sessions for each device on two separate days. The second session for a given participant was at most 27 days later and at least 23 days later. For each of the 23 subjects, we therefore had four datasets (two sessions for each device). Each dataset included 8 trials for the math task, 8 for the image task, and 16 for the resting task. We were interested in whether or not we could build a machine learning algorithm to separate the image and math task using both fNIRS and EEG data, and how this algorithm might change in the second session.

At the time of this report the results from this data collection have yet to be analyzed.

TASK 4: Updating the Machine Learning Code to Work with Data from Multiple Sensors and Standardizing our Data Formats

We also worked this year to expand on the python code developed in Year 1 of this effort. We added in the capability to take in data from multiple sensors (fNIRS, EEG, GSR, ECG, and Respiration), from multiple sessions, and we standardized the format for ‘conditions files’, which are files created by experimenters that detail all possible ‘labels’ that can be applied to data for supervised machine learning. For example, the Tetris medium task described above can be labeled as ‘tetris medium’, or it can be labeled with the person’s performance during that task, or it can be labeled by the person’s NASA-TLS or Self Assessment Manikin self reports filled out after the task of interest.

A description of the code, and the standardized file formats are included in the ‘Python Machine Learning Manual’ as supplementary material with this report.

REVIEW OF YEAR 3 RESEARCH (October 2015-October 2016)

As summarized above, throughout this research we undertook a large fNIRS data collection effort that involved collecting supervised machine learning data across many participants and many different task types, in order to build accurate models of cyber operators cognitive and emotional state changes without overfitting to participants or to task types. To facilitate quick implementation of machine learning models across participants and across task type (to build models that do not overfit to an individual or to a specific task types), we designed the studies, and organized the data, in a systematic way:

Data from our experiments is stored in.csv format, and organized into three main file types to reflect the three types of data collected in all studies. This includes the fNIRS oxy- and deoxy-hemoglobin datafiles acquired during an experiment, as well as what are known as ‘*conditions files*’ and a ‘*meta-data file*’, as explained next.

The *conditions file* acts as both a reference to and a descriptor of the fNIRS datafiles. The conditions file accomplishes this by indicating where in a data file an event began, as well as for how long an event lasted (event onsets and durations). The conditions file also contains information about the type of task the participant was engaged in during the event, her performance during the event, as well as other information the participant self reported about the task immediately following the completion of the task. All of this information can be used to label segments of fNIRS data for supervised classification, as research applying machine learning to brain data has used self-report values, stimulus/task type, and performance as labels [23]. Since workload and emotional state are intertwined [24], we always administer the NASA Task Load Index [17] for self-report values of workload as well as the Self-Assessment Manikin [14] for self-report values of affect (valence and arousal) after every experimental task. Thus, all datasets collected thus far have associated values of participants’ TLX and SAM throughout that experiment, along with other performance or self-assessment measures we may have recorded during the experiment. We also can label machine learning data based on task type, which has been done in a good deal of prior research [3, 6, 21, 22, 25]. The types of tasks participants completed in our current datasets are varied, and can be classified on a spectrum from those tasks that are considered "ecologically valid," to those that are more "construct valid"[26]. Tasks considered to be more ecologically valid in our dataset include tasks such as playing Tetris, viewing emotionally salient videos, or working with the Multi-Attribute Task Battery [27]. Tasks that are considered more construct valid in our dataset include validated measures for testing psychological constructs such as the n-back task for measuring working memory [9, 28] , or the Posner cueing paradigm for measuring visual attention [29].

Tasks in the dataset will be organized into task vectors, which numerically describe each task. The task vectors will be modeled off of Wickens' Multiple Resource Theory [30, 31], allowing for a value to be assigned to the engagement of each mental resource involved in accomplishing a particular task. MRT demand vectors are broken up into ten different scalar values. Each of these scalar values represents a level of engagement of a mental resource. The mental resources that are described by a demand vector are as follows: *Visual-Ambient*, *Visual-Spatial*, *Visual-Verbal*, *Visual-Focal*, *Auditory-Spatial*, *Auditory-Verbal*, *Cognitive-Spatial*, *Cognitive-Verbal*, *Response-Spatial*, *Response-Verbal*. The benefit of this approach is that it allows a researcher to organize the tasks in terms of the level of engagement of a particular mental resource, rather than organizing the data in terms of psychological constructs (i.e., a Tetris task could be described by the semantic term 'tetris', or by a Wickens' (MRT) demand vector that describes the task as high in both Visual-Spatial, as well as Cognitive-Spatial demand). We are choosing this system of task organization because querying the dataset for tasks that are 'high' or 'low' examples in the use of one particular mental resource can return novel combinations of tasks that fit the criteria defined by the researcher. For instance, if one were to only want to train a model on tasks that were high in Visual-Ambient processing then the model would be trained on data from participants who completed the Tetris tasks, as well as participants who completed the Posner cueing paradigm tasks, as their demand vectors both indicate that those tasks involve high amounts of visual-ambient processing.

The **meta-data file** contains all of the demographics for each participant, the probe configuration used for that participant, along with a reference to where that file is located in the database. This organization of the database helps with not only being able to organize the data in terms of variables relating to a participant's biological sex, age, handedness, or level of education (demographics data), but also it further allows the organization of the data by variables such as the type of task a participant completed, as well as their self-reported levels of workload and affect (or other self-reported state measures given to participants) with the task.

To benefit from this organization we have developed Matlab scripts that allow researchers to query the dataset with the parameters they want with reference to demographics, task type, as well as self-reported levels of mental engagement of the task, and they can quickly train and test models based on their input parameters (using across-participant-cross-validation, or by specifying different train/test sets). With this infrastructure in place, a researcher can quickly train and test cross experimental models on existing as well as future datasets included in the database.

Sample queries:

Train model on data from all female participants who conducted the Multi-Attribute Task Battery
Use labels of participants' performance for supervised classification
Run a leave-one participant out cross validation
Train model on data from all right handed people who have task vectors indicating high spatial engagement, excluding tetris
Use labels of participants' self-report NASA-TLX value for supervised classification
Test model on data from all right handed participants who completed on the tetris task

In this way during Year 3 have been able to quickly generate models on all of the datasets collected during Y1 and Y2 of this research, resulting in publications demonstrating the use of advanced machine

learning on fNIRS data for predicting emotional states, attentional state changes, as well as changing levels of trust during online phishing attacks, voice impersonation attacks, and changing levels of suspicion during HCI [32-36]. We've also received a new award from the National Science Foundation building on this research, which will involve applying Long Short Term Memory Networks to our extensive fNIRS dataset.

APPENDIX A: All Publications from this effort

Bobko, P., Barelka, A., **Hirshfield, L.**, Lyons, J. How the Study of the Construct of Suspicion Can Benefit Theories and Models in Organizational Science. *Journal of Business and Psychology*. 2014.

Bobko, P., Barelka, A., **Hirshfield, L.M.** A Review of the Construct of “Suspicion” with Applications to Automated and Information Technology (IT) Contexts: A Research Agenda. *the Human Factors and Engineering Society Journal (HFES)* 2014).

Hirshfield, L.M., Barelka, A. Bobko, P. Paverman, D., Hirshfield, S., Using Non-Invasive Brain Measurement to Explore the Psychological Effects of Computer Malfunctions On Users During Human-Computer Interactions. *Journal of Advances in Human-Computer Interaction* , 2014.

Solinger, C. **Hirshfield, L.** Hirshfield, S., Friedman, R., Lepre, C. Beyond Facebook Personality Prediction: A *Multidisciplinary Approach in Predicting Social Media Users' Personality*. Proc. of the International Conference of Human Computer Interaction, Crete, Greece. (2014).

Bandara, D., **Hirshfield, L.**, Velipasalar, S., Insights into User Personality and Learning Styles through Cross Subject fNIRS Classification. *Proceedings of the International Conference of Human Computer Interaction, Crete, Greece. (HCII 2014)*.

Sommer, N., **Hirshfield, L.**, Velipasalar, S., *Our Emotions as Seen Through a Webcam*. Proc. of the International Conference of Human Computer Interaction, Crete, Greece. (HCII 2014).

Neupane, A., Rahman, L., Saxena, N., **Hirshfield, L.M.** A Multi-Modal Neuro-Physiological Study of Phishing Detection and Malware Warnings. In *ACM Conference on Computer and Communications Security (CCS)*, 2015

Serwadda, A., Poudel, S., Phoha, V., **Hirshfield, L.**, Bratt, M., Bandara, D., Costa, M., *Biometric Authentication using Functional Near-Infrared Spectroscopy*. the IEEE Biometrics, Theory, Applications, and Systems (BTAS) conference (2015).

Hirshfield, L., Costa, M., Bratt, S., Bandara, D. *Measuring Situational Awareness Aptitude Using Functional Near-Infrared Spectroscopy*. in Proc. of the International Conference of Human Computer Interaction, Los Angeles, CA (2015).

Hirshfield, L.M. Bobko, P., Barelka, A., Toward Interfaces that Help Users Identify Disinformation Online: Using fNIRS to Measure Suspicion. *ACM Transactions on Human-Computer Interaction*. (under review). 2017.

Sommer, N., **Hirshfield, L.M.**, Velipasalar, S. Improving Emotion Classification with Convolutional Neural Networks on Functional Near-Infrared Spectroscopy Data. *Journal of Brain-Computer Interfaces* (under revision). 2017.

Bandara, D., **Hirshfield, L.**, Velipasalar, S. *Building Predictive Models of Emotion with Functional Near-Infrared Spectroscopy*. Vol 110. *International Journal of Human-Computer Studies* (2017).

Neupane, A., Saxena, N., **Hirshfield, L.**, Bratt, S. *The (Neuro)Science of Voice Security: An fNIRS Study of Speaker Legitimacy and Familiarity Detection* (submitted) *USENIX Security Conference*. 2018.

S.W. Hincks, S. Bratt, S. Poudel, V. Phoha, R.J.K. Jacob, D.C. Dennett, and **L.M. Hirshfield**, “*Entropic Brain-computer Interfaces - Using fNIRS and EEG to Measure Attentional States in a Bayesian Framework*,” *Proc. PhyCS 2017 International Conference on Physiological Computing Systems*. 2017.

Neupane, A., Saxena, N., **Hirshfield, L.**, Neural Underpinnings of Website Legitimacy and Familiarity Detection: An fNIRS Study. In *Security and Privacy Track, the World-Wide Web Conference (WWW)*, 2017.

REFERENCES

1. Bobko, P., A. Barelka, and L.M. Hirshfield, *A Review of the Construct of "Suspicion" with Applications to Automated and Information Technology (IT) Contexts: A Research Agenda*. Human Factors 2013.
2. Escalante, J., L.M. Hirshfield, and S. Butcher. *Evaluating Interfaces Using Electroencephalography and the Error Potential*. in *International Conference of Human-Computer Interaction*. 2013. Las Vegas, NV: Springer.
3. Girouard, A., E. Solovey, L.M. Hirshfield, K. Chauncey, A. Sassaroli, S. Fantini, and R. Jacob. *Distinguishing Difficulty Levels with Non-invasive Brain Activity Measurements*. in *Proc. INTERACT Conference*. 2009.
4. Haas, M., L. Hirshfield, P. Ponangi, P. Kidambi, D. Rao, N. Edala, E. Armbrust, F. Fendley, and N. S., *Decision-making and emotions in the contested information environment*. . ICST Journal Transactions. , 2012.
5. Hirshfield, L., M. Costa, D. Paverman, E. Murray, and S. Hirshfield. *Measuring the Trustworthiness of Ebay Seller Profiles with Functional Near-Infrared Spectroscopy*. in *International Conference on Human Computer Interaction*. 2013 Crete, Greece.
6. Hirshfield., R. Gulotta, S. Hirshfield, S. Hincks, M. Russell, T. Williams, and R. Jacob. *This is your brain on interfaces: enhancing usability testing with functional near infrared spectroscopy*. in *SIGCHI*. 2011. ACM.
7. Hirshfield, L.M., K. Chauncey, R. Gulotta, A. Girouard, E.T. Solovey, R.J.K. Jacob, A. Sassaroli, and S. Fantini. *Combining Electroencephalograph and Near Infrared Spectroscopy to Explore Users' Mental Workload States in HCI International*. 2009 Springer.
8. Hirshfield, L.M., A. Girouard, E.T. Solovey, R.J.K. Jacob, A. Sassaroli, Y. Tong, and S. Fantini. *Human-Computer Interaction and Brain Measurement Using Functional Near-Infrared Spectroscopy*. in *Symposium on User Interface Software and Technology: Poster Paper*. 2007. ACM Press.
9. Hirshfield, L.M., E.T. Solovey, A. Girouard, J. Kebinger, R.J.K. Jacob, A. Sassaroli, and S. Fantini. *Brain Measurement for Usability Testing and Adaptive Interfaces: An Example of Uncovering Syntactic Workload in the Brain Using Functional Near Infrared Spectroscopy*. in *Conference on Human Factors in Computing Systems: Proceeding of the twenty-seventh annual SIGCHI conference on Human factors in computing systems*. 2009.
10. Sassaroli, A., F. Zheng, L.M. Hirshfield, A. Girouard, E.T. Solovey, R.J.K. Jacob, and S. Fantini, *Discrimination of mental workload levels in human subjects with functional near-infrared spectroscopy*. accepted in the Journal of Innovative Optical Health Sciences 2009.
11. Solovey, E., A. Girouard, K. Chauncey, L. Hirshfield, A. Sassaroli, F. Zheng, S. Fantini, and R. Jacob. *Using fNIRS Brain Sensing in Realistic HCI Settings: Experiments and Guidelines* in *ACM UIST Symposium on User Interface Software and Technology*. 2009 ACM Press.
12. Wang, X., R. Hutchinson, and T. Mitchell. *Training fMRI Classifiers to Discriminate Cognitive States across Multiple Subjects*. in *Advances in Neural Information Processing Systems (NIPS)*. 2004.
13. S. Koelstra, C. Muehl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, *DEAP: A Database for Emotion Analysis using Physiological Signals*. IEEE Transaction on Affective Computing, Special Issue on Naturalistic Affect Resources for System Building and Evaluation, 2014.
14. Bracley, M. and P.J. Lang, *Measuring Emotion: The Self-Assessment Manikin and the Semantic Differential* Journal of Behavior, Therapy, and Experimental Psychiatry, 1994. **25**(1): p. 49-59.

15. Bobko, P., A. Barelka, and L.M. Hirshfield, *The construct of state-level suspicion: a model and research agenda for automated and information technology (IT) contexts*. Human Factors 2014. **56**(3): p. 489-508.
16. Lafferty, J.C., P. Eady, and J. Elmers, *The desert survival problem*. 1974.
17. Hart, S.G. and L.E. Staveland, *Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research*, in *Human Mental Workload*, P. Hancock, Meshkati, N., Editor. 1988: Amsterdam. p. pp 139 - 183.
18. Neupane, A., L. Rahman, N. Saxena, and L. Hirshfield. *A Multi-Modal NeuroPhysiological Study of Phishing Detection and Malware Warnings*. in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. 2015. ACM.
19. John, M.S., D. Kobus, J. Morrison, and D. Schmorow, *Overview of the DARPA Augmented Cognition Technical Integration Experiment*. International Journal of Human-Computer Interaction, 2004. **17**(2): p. 131-149.
20. Tan, D. and A. Nijholt, *Brain-Computer Interfaces: Applying our Minds to Human-Computer Interaction*. 2010: Springer.
21. Solovey, E., D. Afergan, E. Peck, S. Hincks, and R. Jacob, *Designing Implicit Interfaces for Physiological Computing: Guidelines and Lessons Learned Using fNIRS*. ACM Transactions on Human Computer Interaction, 2015. **21**(6).
22. Grimes, D., D. Tan, S. Hudson, P. Shenoy, and R. Rao. *Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph*. in *CHI 2008 Conference on Human Factors in Computing Systems*. 2008. Florence, Italy.
23. Costa, M. and S. Bratt. *Truthiness: Challenges Associated with Employing Machine Learning on Neurophysiological Sensor Data*. in *In Proceedings, Part I, 10th International Conference on Foundations of Augmented Cognition: Neuroergonomics and Operational Neuroscience*. 2016. Springer-Verlag.
24. Parasuraman, R. and D. Caggiano, *Neural and Genetic Assays of Human Mental Workload in Quantifying Human Information Processing*. 2005, Lexington Books.
25. Izzetoglu, K., S. Bunce, B. Onaral, K. Pourrezaei, and B. Chance, *Functional Optical Brain Imaging Using Near-Infrared During Cognitive Tasks*. International Journal of Human-Computer Interaction, 2004. **17**(2): p. 211-231.
26. Burgess, P.W., N. Alderman, and C. Forbes, *The case for the development and use of "ecologically valid" measures of executive function in experimental and clinical neuropsychology*. Journal of the International Neuropsychological Society, 2006. **12**(2): p. 194-209.
27. Arnegard, R. and R. Comstock, *The multi-attribute task battery for human operator workload and strategic behavior research*. 1992: NASA Technical Report.
28. Zhou, Y., Y. Zhu, and T. Jiang. *Ventrolateral prefrontal activation during a N-back task assessed with multichannel functional near-infrared spectroscopy*. in *The International Society for Optical Engineering*. 2007.
29. Posner, M.I., C.R. Snyder, and B.J. Davidson, *Attention and the detection of signals*. . Journal of Experimental Psychology, 1980. **109**(2): p. 160-174.
30. Wickens, C., *Multiple resources and mental workload*. Human Factors., 2008. **50**(3): p. 449-55.
31. Wickens, C.D., *Multiple resources and performance prediction*. Theoretical Issues in Ergonomics Science, 2002. **3**(2): p. 159-177.
32. Bandara, D., S. Velipisalar, S. Bratt, and L. Hirshfield, *Building Predictive Models of Emotion with Functional Near Infrared Spectroscopy*. International Journal of Human-Computer Studies (accepted), 2017.

33. Hincks, S., S. Bratt, S. Poudel, V. Phoha, R. Jacob, D. Dennett, and L. Hirshfield. *Entropic Brain-Computer Interfaces: Using fNIRS and EEG to Measure Attentional States in a Bayesian Framework*. . in *4th International Conference on Physiological Computing Systems*. 2017. Madrid, Spain.
34. Hirshfield., Bobko, Barelka., and Sommer., *Toward Interfaces that Help Users Identify Misinformation Online: Using fNIRS to Measure Suspicion*. (under review), 2018.
35. Neupane, A., N. Saxena, and L. Hirshfield. *An fNIRS Study of Website Legitimacy and Familiarity Detection*. in *In Security and Privacy Track, World Wide Web Conference (WWW)*. 2017.
36. Neupane, A., N. Saxena, S. Bratt, and L. Hirshfield. *Neural Fingerprints of Vocal Security: An fNIRS Study of Speaker Legitimacy and Familiarity*". in *under submission at the IEEE European Symposium on Security and Privacy (submitted)* 2017.