AFRL-AFOSR-VA-TR-2017-0111

BRI: CYBER TRUST AND SUSPICION

Eunice Santos
UNIVERSITY OF TEXAS AT EL PASO

06/06/2017
Final Report

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

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
06-06-2017	Final Performance	30 Sep 2012 to 31 Dec 2014

4. TITLE AND SUBTITLE
BRI: CYBER TRUST AND SUSPICION

5a. CONTRACT NUMBER

5b. GRANT NUMBER
FA9550-12-1-0457

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Eunice Santos

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
UNIVERSITY OF TEXAS AT EL PASO
500 UNIV ST ADMIN BLDG 209
EL PASO, TX 79968-0001 US

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AF Office of Scientific Research
875 N. Randolph St. Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR RTA2

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-AFOSR-VA-TR-2017-0111

12. DISTRIBUTION/AVAILABILITY STATEMENT
A DISTRIBUTION UNLIMITED: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT

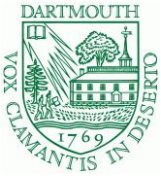
For the Cyber Trust and Suspicion (CTS) research project, the project proposal identified a substantial undertaking in advancing our knowledge and understanding of both trust and suspicion in the cyber domain, particularly as they apply to insider behaviors and insider manipulation. In order to quickly progress this effort, five separate thrusts were identified, to be addressed separately by different research teams. These five thrusts and their major findings for the first two years of the CTS project are briefly summarized in the next few paragraphs, and then covered more extensively in the remainder of this report.
THRUST 1 A Social, Cultural, And Emotional Basis for Trust and Suspicion: Manipulating Insider Threat In Cyber Intelligence & Operations: For 2013, the concepts of Predictability, Susceptibility, and Awareness (PSA) as predictors of insider types was explored. Efforts were focused on defining, representing, and validating the PSA modeling framework, as outlined in the Cyber Trust and Suspicion (CTS) research proposal. Synthetic scenarios reflective of realistic situations were created to analyze how PSA might appear in real-world situations, and how they could be measured and applied to identifying potential insider types.

15. SUBJECT TERMS
cyber, trust, suspicion

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			NGUYEN, TRISTAN
Unclassified	Unclassified	Unclassified	UU		

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

				19b. TELEPHONE NUMBER *(Include area code)* 703-696-7796



FINAL PROJECT REPORT

CYBER TRUST AND SUSPICION

Contract/Grant # : FA9550-12-1-0457

Reporting Period : 09/30/2012 to 12/31/2014

TABLE OF CONTENTS

TABLE OF FIGURES

LIST OF TABLES

EXECUTIVE SUMMARY

For the Cyber Trust and Suspicion (CTS) research project, the project proposal identified a substantial undertaking in advancing our knowledge and understanding of both trust and suspicion in the cyber domain, particularly as they apply to insider behaviors and insider manipulation. In order to quickly progress this effort, five separate thrusts were identified, to be addressed separately by different research teams. These five thrusts and their major findings for the first two years of the CTS project are briefly summarized in the next few paragraphs, and then covered more extensively in the remainder of this report.

THRUST 1 – A Social, Cultural, And Emotional Basis for Trust and Suspicion: Manipulating Insider Threat In Cyber Intelligence & Operations: For 2013, the concepts of *Predictability*, *Susceptibility*, and *Awareness* (*PSA*) as predictors of insider types was explored. Efforts were focused on defining, representing, and validating the *PSA* modeling framework, as outlined in the Cyber Trust and Suspicion (CTS) research proposal. Synthetic scenarios reflective of realistic situations were created to analyze how *PSA* might appear in real-world situations, and how they could be measured and applied to identifying potential insider types. In order to accomplish this, a study was made of insider cases, along with an examination of the common indicators found in these cases. Additionally, a survey of existing measurement tools for personality was initiated. Initial definitions of each *PSA* term, and each insider type, were drafted. These were then subsequently related to the synthetic scenarios for examination. Bayesian Knowledge Base (BKB) modeling substantiated the likelihood of conducting measures of potential insiders for their respective types, but also strongly indicated that further refinement were required for *PSA*, insider types, indicators, and measurements of insider behaviors. As a follow-up to the CTS effort in 2013, in 2014 a basis for measuring and quantifying *PSA* through the use of established personality measurement techniques was formed. The scenarios explored in 2013 reinforced the need for a stable platform upon which to base *PSA* measurements. The use of the Five Factor Model (FFM) and the Sixteen Personality Factors (16PF) Model were therefore explored as well-defined approaches to personality measurement. Linkages between those personality models and *PSA* were proposed, and then BKB fragments were constructed to represent their connections. Literature reviews were conducted for each of the components of *PSA* to leverage defined relationships to personality factors. Furthermore, hypotheses were proposed for alternative connections in order to complete the *PSA* models. Finally, the likely role that emotions can have in the dynamic aspects of *PSA* was researched. It is expected that evaluating emotions will be an important aspect of gauging *PSA*. Future work will address the relationships of personality to *PSA*, as well as other dynamic factors, such as emotional triggers, can impact the *PSA* and threat type of insiders.

THRUST 2 – Targeted Interventions Derived From Biomarkers Of Cyber Trust: The primary interest for Thrust 2 is to determine the ability to affect and influence Cyber Trust through via biomarkers. First among these goals was therefore identifying biomarkers for Cyber Trust. This two-year research effort focused on producing a foundation upon which cyber trust biomarkers can be explored. The primary contributions from this research include: the Perceptual Cyber Trust Taxonomy; experimental design for identifying neural correlates of trust decisions in a cyber context; simulation platform for Cyber Trust research ("The CyberPhishing Game"); and data and preliminary analysis of Cyber Trust studies. The results set the stage for deriving a biomarker for Cyber Trust, and conducting additional Cyber Trust studies to evaluate the efficacy of targeted interventions. Additionally, four separate publications were generated from this study.

THRUST 3 – Cyber Trust and Suspicion: A Human-Centric Approach: In this thrust, how uncertainty perception affects human-computer trust in a cyber-attack scenario was investigated. In order to examine the effects of uncertainty on human-computer trust and reasoning, participants'

adherence to cues about the reliability of a radar display that was susceptible to cyber-attacks was measured. Overall, the results of the present study indicate that uncertainty perception plays an important role in human-computer trust. Subjects had the greatest difficulty deciding how to respond when there was an equal (or almost equal) probability of entering a correct response as entering an incorrect response. Although it is unclear how uncertainty perception leads to suspicion (and the resulting effect on performance), it appears that the current findings support the general notion of a trust spectrum with trust on one end, distrust on the other, and uncertainty in the middle.

THRUST 4 – Using Non-Invasive Sensors To Predict Trust And Suspicion In Human Operators: During the first year of the project, the focus of Thrust 4 was on theoretical research as a theoretical understanding of suspicion is needed in order to properly define, manipulate, and measure suspicion. Two data collection experiments were also designed and conducted. These experiments looked at the effects of cyber-attacks on human operators, and at the participants' mouse movements and brain activity as they used the Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU) testbed. During the second year, three experiments relating to the CTS project were designed and implemented. The first experiment was aimed at measuring trust and suspicion using the techniques of keylogging and sentiment analysis. The second experiment used a different data source namely social media posts to determine the users' characteristics such as their personality type and propensity to trust. The third experiment looked at detecting emotional state changes using a webcam. One valuable outcome from the research described above has been a model describing the physiological correlates of trust, distrust, and suspicion, which is being prepared (along with supporting experiment results) for publication.

THRUST 5 – Assessing, Attributing, And Manipulating Operator Suspicion: Thrust 5 consisted of three separate research foci: 1) detecting suspicion, 2) attributing the perceived and actual sources of suspicion, and 3) manipulating suspicion. These separate foci produced feasibility studies that led to prototype sensors for collecting data during experiments. Four distinct cyber sensors were produced to be used in conjunction with an Assured Information Security, Inc. (AIS) platform. As part of the first research focus, a keylogger and mouse logger were developed to investigate the utility of cyber sensors towards suspicion detection in two experiments. The first experiment focused on collecting mouse data while subjects used an unmanned system simulator, and the second experiment allowed subjects to collaborate through an instant messenger to solve the Winter Survival Problem. In the second research focus, additional cyber sensors were developed to determine if both perceived and actual sources can be determined using cyber data. A third experiment was formulated to collect data on suspicion attribution, with the sources of suspicion scoped to only include deny, disrupt, deceive, degrade, and destroy (D5) effects. Unfortunately, the effort was cut short after the second year, before the attribution experiment was performed. As a result, the suspicion attribution analysis and the suspicion manipulation, which was part of the third focus, were not performed.

The overall project lead organization for this research effort was the University of Texas at El Paso (UTEP), with Dr. Eunice E. Santos as the overall lead principal investigator (PI). The organizations and the PIs involved with the various thrusts are given below.

1. Thrust 1 was led by the University of Texas at El Paso (UTEP) and supported by Dartmouth College. The thrust PI was Dr. Eunice Santos.

2. Thrust 2 was led by the University of Tulsa and supported by the Laureate Institute for Brain Research. The thrust PI was Dr. John Hale.

3. Thrust 3 was led by the Texas A&M University, with Dr. Hongbin Wang as the PI.

4. Thrust 4 was led by the Syracuse University, with Dr. Leanne Hirshfield as the PI.

4. Thrust 5 was led by the Assured Information Security, Inc. (AIS), with Dr. John S. Bay, initially as the PI followed by Dr. Barry McKinney.

THRUST 1 – A SOCIAL, CULTURAL, AND EMOTIONAL BASIS FOR TRUST AND SUSPICION: MANIPULATING INSIDER THREAT IN CYBER INTELLIGENCE & OPERATIONS

The research efforts for Thrust 1 was led by the University of Texas at El Paso (UTEP), with support from Dartmouth College. The key personnel are Dr. Eunice E. Santos (thrust principal investigator (PI)), Dr. Eugene Santos Jr. (PI, Dartmouth) and Dr. John Korah (Co-PI, UTEP)[1].

1.1 SUMMARY

Under the Cyber Trust and Suspicion (CTS) effort, in 2013 the viability of the concepts of *Predictability*, *Susceptibility*, and *Awareness* (*PSA*) being sufficient predictors of insider types was explored. For this particular effort, we focused on defining, representing, and validating the *PSA* modeling framework. In line with this approach, we strove to create synthetic scenarios reflective of realistic situations to analyze how the facets of *PSA* might surface in real-world situations, and how they could be measured and applied to identifying potential insider types. This effort was undertaken to better understand the details of the proposed *PSA* framework to insider type correlation, and also to determine if it appeared reasonable to be able to detect such a linkage in common insider scenarios.

In order to accomplish this, a study was made to examine known instances of insider behavior. Related to this study of insider cases, an examination of the common indicators found in these cases was conducted. Additionally, a survey of existing measurement tools for personality and indicators was initiated. Finally, initial definitions of *PSA* and each insider type were drafted, and then subsequently related to the synthetic scenarios for examination.

Initial modeling using Bayesian Knowledge Bases (BKBs) substantiated the likelihood of being able to conduct measures of potential insiders for their respective types, but also strongly indicated that much further refinement and definition were required for *PSA* and insider types, as well as for indicators and measurements of insider behaviors.

As a follow-up to the CTS effort in 2013, in 2014 we endeavored to refine and expand upon our understanding of *PSA*. In particular, we sought to form a basis for measuring and quantifying *PSA* through the use of established personality measurement techniques. The scenarios explored in 2013 informed us on the dynamic qualities of *PSA*, but also reinforced the need for a stable platform upon which to base *PSA* measurements. We therefore explored the use of the Five Factor Model (FFM) and the Sixteen Personality Factors (16PF) as well-understood and well-defined approaches for personality measurement. We established linkages between those models and *PSA*, and then constructed BKB fragments to represent their connections. Extensive studies were conducted for each of the components of *PSA* to leverage defined relationships to personality factors. Additionally, hypotheses for alternative connections were proposed in order to complete the *PSA* models. Finally, the likely role that emotions can have in the dynamic aspects of *PSA* was researched. It is expected that evaluating emotions will be an important aspect of gauging *PSA*. Future work will attempt to validate the relationships of personality with *PSA*, as well as identify what dynamic factors, such as emotional triggers, will also impact the *PSA*, and therefore the threat type, of insiders.

[1] Discussion and results in this section were also provided in the project annual reports.

1.2 INTRODUCTION

Insiders can undoubtedly be the most effective and useful personnel for an organization, as they are, by definition, accepted as an integral part of the organization, and because they have (oftentimes unique) access to critical and confidential information about that organization and its mission. Equally, these same personnel can become absolutely devastating to that same organization, if they decide independently, or are convinced by another, to act contrary to the organization's mission. The promise and threat of insiders remains a conundrum for organizations, particularly for any organization which values trade secrets and competitiveness. This is an acutely sensitive topic for governments and the military, as they are reliant on the integrity and discretion of their employees. Insiders are thus both a vital part of any organization, and simultaneously a critical vulnerability if compromised by outside agencies. The Cyber Trust and Suspicion (CTS) project's goal to develop a model of insiders which can explain the social, cultural, and emotional basis for trust and suspicion is therefore of immense interest.

Trust and suspicion are the very foundations upon which insider exploitation may occur, and thus are central to understanding its impacts on insider threat. By harnessing the powers of computational and social science constructs, a sufficient understanding of insiders should be possible, enough to eventually control and ultimately manipulate insider threats in cyber intelligence and operations.

The CTS project has a number of objectives, each of which is spelled out in the project proposal. However, for the purposes of this particular investigation, the most relevant objective is:

> *Develop a model of insider behavior that accounts for and explains the social, cultural, and emotional basis for trust and suspicion especially its impacts on insider threat.*

The primary reason for this research effort is to better understand the insider. An insider, like any individual, possesses influences, motivations, abilities, beliefs, strengths, and weaknesses. Understanding what drives and influences an insider is critical to advancing our ability to identify insider threats within our own organizations, and, if necessary, to locate and exploit insider threats in the organizations of our opponents. The CTS project proposal put forward a hypothesis that there exist eight different, distinct insider types, and that those types may be identified through the measurement of three qualities: Predictability, Susceptibility, and Awareness (PSA). This report focuses on our efforts to translate that hypothesis into a functional, computational model of insider threat types.

In 2013, concrete definitions for the components of *PSA* were produced. These definitions were:

> *1 An insider's Predictability is based on the ability to foretell that insider's reaction to stimuli that a manipulator could potentially provide.*
> *2 An insider's Susceptibility is the quality or tendency of that insider to become involved in an action that either directly or indirectly affects the organization, due to external or internal influence.*
> *3 An insider's Awareness is the insider's ability to identify manipulative intent behind false and/or partial information.*

It was recognized that each of those components have aspects that are relatively static and unchanging for an individual—the insider's character or nature defining their predisposition towards different levels of *PSA*, and therefore towards a potential insider threat type. This aspect we believe can be captured through the measure of personality, which was the primary focus of our investigations in 2014. We further recognized, though, that there are certainly more dynamic aspects

of *PSA* which are affected by less constant factors, such as emotion. We conducted an initial investigation into emotions, which is outlined in the Background section of this report, and developed a research plan to model how values from personality can be mapped into different insider types.

1.3 BACKGROUND

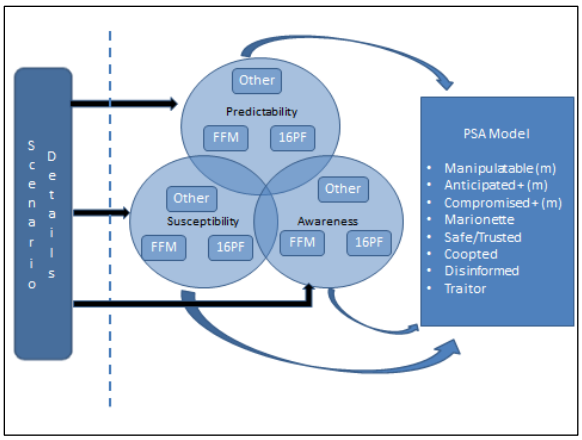


Figure 1: PSA Model with Personality Measures

In this section, we re[illegible]ted the study of *PSA* and the CTS project overall. Emotion appeared to be a key trait that could be related to the more dynamic aspect of *PSA* with regard to insider threat scenarios, while personality measures were identified early on as prospective tools for gaining insight into individual *PSA* profiles. Finally, a number of case studies were researched in order to gain insight into "typical" insider behaviors and backgrounds. A handful of case studies are briefly summarized, while a more complete list is provided in Table 1.

1.3.1 EMOTION

We reviewed the literature to identify previous works that identified aspects of emotions possibly related to *PSA*. In short, both *Susceptibility* and *Predictability* show some promise of being related to emotions. More discussion regarding emotion's effects on *PSA* will be presented later in the relevant portions of Section 1.4, but it is worth noting here that a *Susceptibility* to emotional manipulation can in turn be used to affect *Awareness,* so in essence, all aspects of *PSA* can be impacted by emotion.

According to Zinck [1], *"Basic emotions are mainly characterized by feeling aspects, while an attitudinal component is regarded as a relevant part of the emotion pattern of complex emotions as well"*. This could imply that complex emotions are also related to personality, which is also relevant to our studies. Additionally, fusion and interaction between emotions were identified from literature [2]. Further enquiry about the relationship between emotions and predictability, online emotional transparency, and insider threat specific emotions are avenues for future work.

1.3.2 PERSONALITY MEASURES

Results from the 2013 CTS investigation pointed us towards exploring established personality measures as a foundation upon which to build measurements for *PSA*. Note that it is fully expected that *PSA* will be much more dynamic and variable than personality, but we hypothesize that a major, stabilizing factor in the *PSA* characteristics will be an individual's personality. Thus, we hope to leverage the decades of research into measuring personality for establishing a basis for measuring *PSA* in general.

Several possibilities for standardized personality evaluation methods were investigated. The Five Factor Model (FFM), also known as Big Five, personality traits [3] and the Sixteen Personality Factor (16PF) Questionnaire [4] are two standardized methods that are widely used to measure individual personality traits. We reviewed the literature to understand if individual personality types from these models can be used to model aspects of *PSA*.

FFM measures an individual's personality based on the traits: openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism. 16PF evaluates normal personality based on warmth, reasoning, emotional stability, dominance, liveliness, rule-consciousness, social boldness, sensitivity, vigilance, abstractedness, privateness, apprehensiveness, openness to change, self-reliance, perfectionism, and tension. There have also been studies indicating that the FFM is essentially an orthogonal representation of the 16PF factors [5], [6]. Aspects of these measures will be further explored with each component of *PSA* later in this report.

While considering the impact of personality on *PSA*, a conceptualization was produced of how those personality measures might be incorporated into an overall model of *PSA* and potential insider threat type. That idea is captured in Figure 1, where "Other" indicates that alternative measurements for personality or other static influences can also be incorporated. More dynamic influences, such as emotion and mood, could also be captured in the scenario details and modeled.

1.3.3 INSIDER CASE STUDIES

In an effort to understand the full scope of insider types and situations, we conducted a study of insider cases from publications. In total, 12 separate insider incidents (listed in Table 1) were studied from over 25 different sources. With the wide range of motives discovered during the case study review, it became apparent that modeling such cases presented a number of difficulties. Motives included revenge, nationalism, religion, power, and financial gain. Chief among the difficulties was the lack of detailed character sketches or personality profiles for the actual insiders, which then of course severely limited the ability to tie the characteristics of *PSA* to the actual insider types identified in the cases. Brief summaries of some of the more prominent cases discovered follow, as well as an initial effort to classify the insiders' types as defined in *Table 2*, and thus associate a *PSA* profile with each. Our goal was to learn as much as we could about insider scenarios, in order to leverage this knowledge for computational models. We now describe a sampling of the insider threat scenarios that were studied.

1.3.3.1 Case Study of Insider Sabotage: The Tim Lloyd/Omega Case

Insider type: Disinformed (Not *Predictable*, *Susceptible*, *Aware*)

Traits: Revenge seeking, selfish, emotionally disturbed

Summary: According to court documents[7]: Omega Engineering Corp. ("Omega") is a New Jersey-based manufacturer of sophisticated devices and control equipment for, inter alia, the U.S. Navy and NASA. On July 31, 1996, all of its design and production computer programs were purged, crippling Omega's capabilities and causing a loss of millions of dollars in sales and contracts. The suspect, Tim Lloyd, was the only person responsible for the entire backup. Many warning signs, like being rude with co-workers, were observed but ignored. Lloyd apparently planned the attack for months. [8] In essence, Tim Lloyd changed from a reliable, *Predictable* employee, to a completely unpredictable one. *Susceptibility* was a more difficult thing to ascertain, but something clearly changed his behavior and allegiance. It might be termed as a bruised ego, emotional immaturity, or simply rage, but Mr. Lloyd was clearly susceptible to some form of unhealthy motional influence. His *Awareness* appeared to be unquestionable, based on the premeditated nature of the crime.

Table 1: Full Case Study List

Case Study	Source(s)
"Ring leader" of security breach in NJ banks - Orazio Lembo	▪ http://www.nbcnews.com/id/7670774/ns/nbc_nightly_news_with_brian_williams/t/massive-bank-security-breach-uncovered-nj/#.UeBNAvm1H64
The Omega case - Tim Lloyd	▪ http://craigchamberlain.com/library/insider/Case%20Study%20of%20Insider%20Sabotage.pdf ▪ http://law.justia.com/cases/federal/appellate-courts/F3/269/228/532823/
Ex-Coke workers in trade secret case - Joya Williams and Ibrahim Dimson	▪ http://money.cnn.com/2007/05/23/news/newsmakers/coke/
Wisconsin cancer researcher accused of economic spying - Hua Jun Zhao	▪ http://www.bloomberg.com/news/2013-04-02/wisconsin-cancer-researcher-accused-of-economic-spying-for-china.html ▪ https://di2agwiki.thayer.dartmouth.edu/wiki/ReOXuEWKBcu5kXgr/b/b3/Court_file.pdf
Espionage case - Robert Philip Hanssen	▪ http://www.fbi.gov/about-us/history/famous-cases/robert-hanssen ▪ http://www.fas.org/irp/ops/ci/hanssen_affidavit.html
Erroneous Sandia Labs Chinese espionage case - Wen Ho Lee	▪ http://www.nytimes.com/1999/03/16/world/los-alamos-scientist-admits-contacts-with-chinese-us-says.html ▪ http://www.nytimes.com/2001/02/04/us/the-making-of-a-suspect-the-case-of-wen-ho-lee.html?pagewanted=all&src=pm ▪ https://di2agwiki.thayer.dartmouth.edu/wiki/ReOXuEWKBcu5kXgr/8/8e/Chin_%282003%29_-_Implausible_Denial-_Wen_Ho_Lee.pdf
Attempted espionage by US soldier - William Colton Millay	▪ http://www.fbi.gov/news/stories/2013/april/soldier-receives-16-year-sentence-for-attempted-espionage ▪ http://www.guardian.co.uk/world/2013/apr/16/alaska-policeman-jailed-russian-spy
Fort Hood shooting - Major Nidal Malik Hasan	▪ http://www.cpps.com/blog/wp-content/uploads/2010/08/Fort-Hood-Case-Study-FINAL.pdf ▪ http://www.huffingtonpost.com/kamran-pasha/a-muslim-soldiers-view-fr_b_348973.html
US military contractor divulged confidential information to girlfriend - Benjamin Pierce Bishop	▪ http://www.fbi.gov/honolulu/press-releases/2013/defense-contractor-charged-in-hawaii-with-communicating-classified-information-to-person-not-entitled-to-receive-such-information ▪ http://www.fbi.gov/honolulu/press-releases/2013/defense-contractor-charged-in-hawaii-with-communicating-classified-information-to-person-not-entitled-to-receive-such-information
Industrial espionage - Steven L. Davis	▪ https://www.fas.org/irp/news/1998/01/davispld_htm.html ▪ http://www.wright.edu/rsp/Security/Spystory/Industry.htm ▪ http://tradesecretshomepage.com/indict.html#_Toc9924965
FBI double agent and her US handler - Katrina Leung and James J.Smith	▪ http://www.fas.org/irp/ops/ci/leung.html ▪ http://www.nytimes.com/2006/05/25/washington/25spy.html?ref=katrinaleung
Passing classified information to lobbyists for the American Israel Public Affairs Committee - Lawrence Franklin	▪ http://www.fas.org/irp/ops/ci/franklin0805.pdf ▪ http://forward.com/articles/108778/once-labeled-an-aipac-spy-larry-franklin-tells-his/#ixzz1J9hQ5hKU

1.1.1.1 Wisconsin Researcher Accused of Economic Spying for China

Insider Type: Traitor (*Predictable*, *Susceptible*, *Aware*)

Summary: Insider made contact with an outside organization, and sought position and funding in this organization. Suspect sent a package containing the patented compound back to a family member in China. He also transferred secret research data and references from a lab computer to his personal computer and an external disk. He entered the lab office without permission and stole patent compounds. Incredibly, he claimed on a website that he found the patent compound himself. Once confronted, he refused to answer FBI questions directly. [9] To our best estimation, the insider could be considered *Predictable*, in that his behavior was not necessarily out of character, he was clearly *Susceptible* to greed, and was also quite *Aware* of his deceit.

1.1.1.2 Robert Philip Hanssen Espionage Case

Threat type: Traitor or Disinformed (*Susceptible*, *Aware*)

Traits: Greedy and obsessed with divulging information

Summary: Hanssen was an FBI agent who spied for Soviet intelligence. He went undetected for 22 years, from 1979 to 2001, and held key computer-intelligence positions during that time. Though many warning signs were observed and reported by colleagues, they were not taken seriously by superiors[2]. Hanssen's *Predictability* was particularly difficult to assess, as the case notes indicated that he apparently volunteered to provide classified information to the KGB. He of course was *Susceptible*, perhaps to greed, and also well *Aware* of his crimes.

1.1.1.3 Attempted Espionage By US Soldier William Colton Millay

Threat type: Traitor (*Predictable*, *Susceptible*, *Aware*)

Traits: Greedy, white-supremacist, Nazi tattoos, reckless, and desperate to make money

Summary: Millay was a military police officer stationed at Joint Base Elmendorf-Richardson near Anchorage. He was accused of trying to sell classified information to a person he believed was a Russian intelligence officer. Millay was sentenced a 16 year jail term. In this case his co-workers appeared to either lax or afraid and failed to report the threat. He tried to convince co-workers to perform espionage and this scenario depicts the possibilities of how a malicious intent could spread among co-workers. Millay was *Predictable* and *Susceptible* because of his stereotypical characteristics. He was also *Aware* of his crimes, at the same time appeared to be disillusioned to a certain extent.

1.4 METHODS AND RESULTS

One of the key issues in developing an insider threat model is to understand and be able to represent the relationships between information manipulations, emotional responses, changes in task performance, and changes in situational awareness and its influence on decision-making. *PSA* provides the fundamental constructs required to represent such complex relationships. The insider threat spectrum is often not binary, and it scales from an insider simply being vulnerable to being a severe threat or perhaps having already caused significant damage. *PSA* provides a unifying

[2] USA v. Robert Philip Hanssen: Affidavit in Support of Criminal Complaint, Arrest Warrant and Search Warrant. URL: http://fas.org/irp/ops/ci/hanssen_affidavit.html

framework for representing these various types and levels of insider threat. As key components of the CTS proposal, *PSA* are hypothesized to be good indicators of *potential* insider type. Moreover, a breakdown of eight different *potential* insider types was indicated, based on the existence (or not) of the three *PSA* components in the subject individual. It is critical that one note the term *potential*, as it is not suggested that all individuals with a particular *PSA* profile will actually commit destructive insider actions, but that if they do at some point become harmful insiders, they will be most likely to exhibit behaviors according to the identified type. While the linkage between *PSA* and insider types is proposed in the CTS proposal (and indicated in *Table 2* below), the exact nature of the connections, the ability to measure the *PSA* of individuals, how identifiable each type of insider is, and even the extent to which these connections exist remain to be determined. In the initial exploration of the *PSA* concept during the first year of the project, the ability to make deductions and inferences about the *PSA* of individuals in synthetic scenarios is explored, in order to examine the viability of *PSA* as a means for establishing potential insider types.

To establish the groundwork for producing a synthetic scenario model, three distinct areas of investigation were required:

1. Insider case studies (see Section 1.3.3 INSIDER CASE STUDIES),
2. Insider type definition and distinction, and
3. *PSA* definition and elaboration.

Table 2: The Three Dimensions of an Insider Behavior and Potential Insider Threat [3]

Predictable?	Susceptible?	Aware?	Potential Insider Threat Type
No	No	No	Manipulatable (m)
Yes	No	No	Anticipated + (m)
No	Yes	No	Compromised + (m)
Yes	Yes	No	Marionette
No	No	Yes	Safe/Trusted
Yes	No	Yes	Coopted*
No	Yes	Yes	Disinformed*
Yes	Yes	Yes	Traitor*
(*) denotes exceptional cases especially with aware insiders			

As part of the *PSA* elaboration, synthetic scenarios for detecting each of the *PSA* values were developed to test initial viability of the proposed modeling framework.

1.4.1 *PSA* DEFINITION AND ELABORATION

In studying the cases in Section 1.3.3, it was identified that additional clarification was required for the insider types specified in *Table 2*. In particular, while indicated in the table, attention should be drawn to the "+ (m)", which indicates that for the types *Anticipated* and *Compromised*, they are also considered *Manipulatable*. Additionally, for the *Coopted*, *Disinformed*, and *Traitor* types, the insiders are aware and typically not a threat, but in the rare instance they are a threat, they are aware of the harmful aspects of their actions, as opposed to the unaware types, who may be "tricked" into accomplishing tasks without realizing their deleterious effects. Moreover, *Safe/Trusted* type is

[3] From CTS proposal

exceptionally unlikely to become a threat, but technically is still considered a potential threat type, and one of particularly harmful capacity, due to their trusted status. Beyond these distinctions, the types are distinguishable based upon their particular *PSA* profiles, once precise definitions for *Predictability*, *Susceptibility*, and *Awareness* are agreed. Here, we explore the precise definitions of *Predictability*, *Susceptibility*, and *Awareness*, in the context of the CTS project.

1.4.2 PREDICTABILITY DEFINITION AND INITIAL MODEL

Based on the general concept of *PSA* determining the insider type, it was considered important to tie the definition of *Predictability* to stimuli that a manipulator might provide, such that the insider's reaction is then (more) predictable.

Based on our survey of existing literature in the field, we formulated the following definition for predictability.

> *An insider's Predictability is based on the ability to foretell that insider's reaction to stimuli that a manipulator could potentially provide.*

This makes for a challenge when evaluating *Predictability*, because the insider's *Predictability* is based on contextual information (stimuli), which can of course vary by situation. Bias, i.e. an inclination or prejudice towards or against an entity or idea, could provide the reasoning behind an insider's behavior in a given situation. So in an effort to provide a concrete foundation that was independent of stimuli, we first investigated the possibility that an insider's bias might provide a consistent measure of their *Predictability*. We broadly classified biases based on internal and external factors. Socio-cultural biases are inherent biases that do not require any conscious thinking (age, gender, education, etc.). Emotional biases are those based on the state of the insider (being happy, sad, etc.). Socio-cultural and emotional biases are both based on internal factors. Situational biases (biases invoked due to external events or stimuli) and social network based biases (an insider's biases toward certain groups or organizations) are both based on external factors. This bias-based classification facilitates categorical modeling of an insider's predictability.

1.4.2.1 Initial Model

An initial model for *Predictability* was devised to facilitate the investigation of bias as a prime indicator for measurement. Details of that model are outlined below.

In this work, we consider a scenario to represent a specific behavior aspect relevant to the CTS domain. The scenario may contain multiple time steps and events. Various levels have been formulated to systematically combine multiple scenarios and targets. A target represents a single outcome or a group of closely related outcomes in the scenario. Levels represent the scope of the model. Intuitively, as the level increases, the number of scenarios and targets considered in the model also increase, thereby broadening the scope of the model as shown in Figure 2. We analyzed how the model behaved as more knowledge, biases, and events are incorporated into our system.

> *Target: A single factor or a group of closely related factors--represented as random variables (rvs) in our framework.*
> *Level 1: Single scenario and single Target ($P_{S,S}$)*
> *Level 2: Single scenario, multiple targets ($P_{S,M}$), or*
> *Multiple scenario, single target ($P_{M,S}$)*
> *Level 3: Multiple scenario, multiple targets ($P_{M,M}$)*

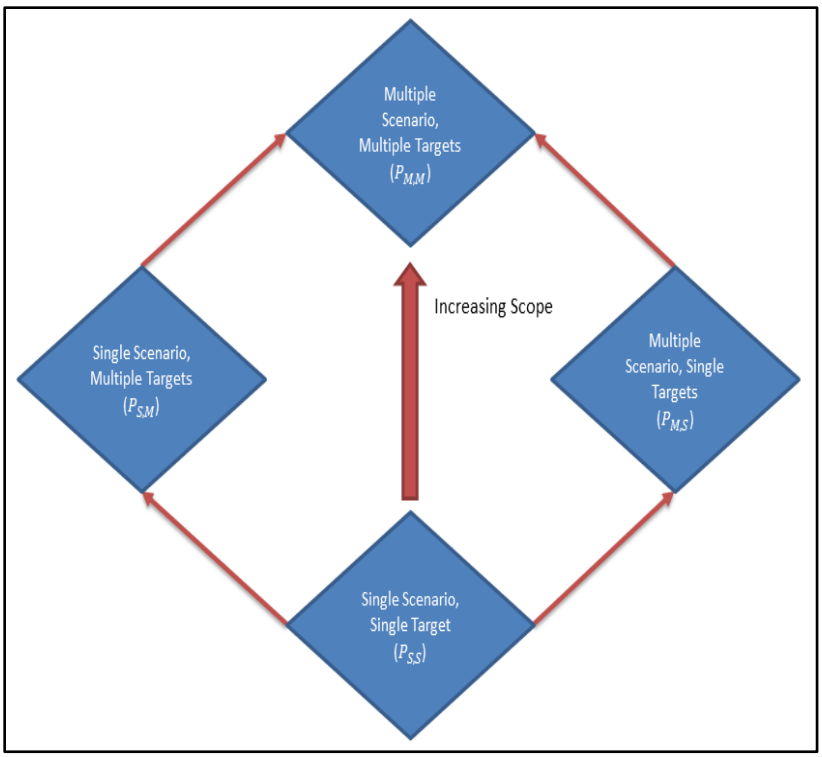


Figure 2: Predictability Levels

Bayesian knowledge bases (B ctability scenario. BKBs are a collection of conditional probability rules based on Bayesian statistics. BKBs can be used to represent various events in our scenario through random variables (rvs), i.e. events connected to each other by probabilistic rules as shown in Figure 3 - Figure 6. We defined a measure of *Predictability* via posterior analysis using a technique known as belief updating [11]. On a high level, belief updating provides the probability of a given event occurring in our scenario. Through belief updating, it is possible to calculate the posterior probabilities of all the states of an rv of interest. Selecting rvs associated with actions or decisions taken by an insider can then lend insight into that individual's *Predictability*. By calculating the standard deviation of the posterior probabilities of the different states of that rv, we have a rough measure of the individual's *Predictability* in that situation.

1.4.2.2 Synthetic Scenario

One of our goals here is to investigate whether bias could serve as a consistent indicator of predictability. For a *Predictability* scenario, we focused on modeling a job recruiter's hiring-bias based on candidates' schooling, past work experience, etc. We developed two synthetic scenarios to analyze the effect of bias on *Predictability*, and also to demonstrate our basic methodology. In scenario 1, the candidate was a Texas A&M graduate, and in scenario 2 the candidate was a Harvard graduate. We modeled how the recruiter's inherent bias on the candidate's alma mater, along with other external factors such as work experience, influences the hiring decision. The *Predictability* levels for the scenario are defined in *Table 3*, while relevant BKBs can be found in Figure 3 - Figure 6. For these BKBs, the goal "(G) Hire candidate" rv represents the unbiased output of the hiring process, while the action "(A) Candidate hired" represents the actual outcome of the situation, once bias has been considered.

Table 3: Example of Scenario Predictability Levels

Levels	Target 1- Action Candidate Hired	Target 2- Goal Hire Candidate
Level 1	Scenario 1: Texas A&M Grad	Scenario 1: Texas A&M Grad
Level2a	Scenario 1: Texas A&M Grad Scenario 2: Harvard Grad	Scenario 1: Texas A&M Grad Scenario 2: Harvard Grad
Level2b	Scenario 1: Texas A&M Grad	
Level3	Scenario 1: Texas A&M Grad Scenario 2: Harvard Grad	

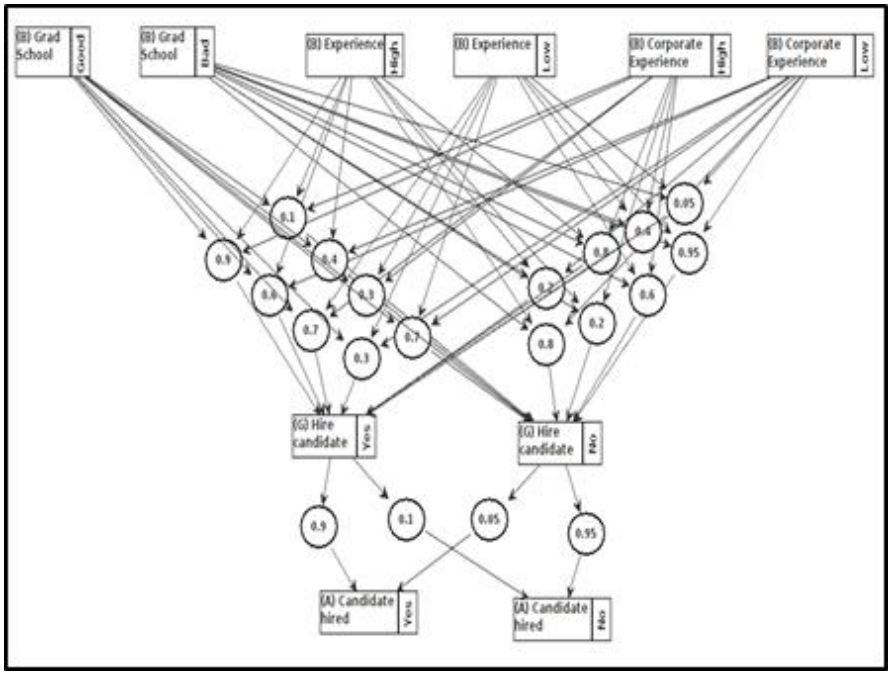


Figure 3: Predictability Hiring Base BKB

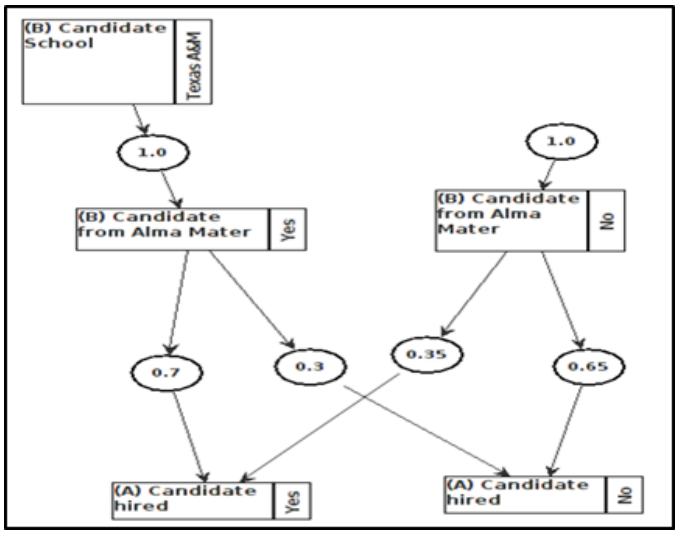


Figure 4: Predictability Hiring Bias BKB

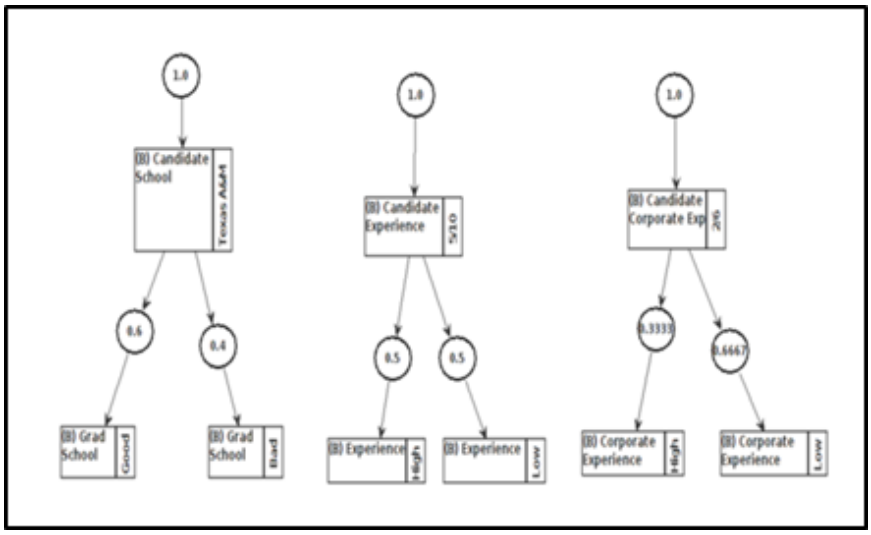


Figure 5: Predictability Scenario 1 (Texas A&M) BKB

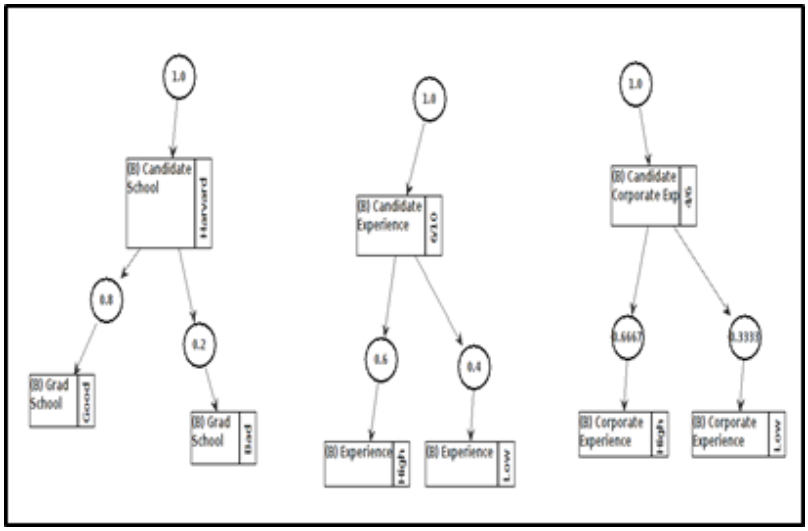


Figure 6: Predictability Scenario 2 (Harvard) BKB

1.4.3 SUSCEPTIBILITY DEFINITION AND INITIAL MODEL

It was determined after thorough research of existing literature, that insiders could be influenced by external manipulation, and/or by their own internal arguments and justifications. *Susceptibility* also helps in incorporating the threat indicators that are usually observed in an insider threat scenario. These indicators could be based on observable changes at an individual level (disgruntlement, disengagement, performance, stress, etc.) and observable events (personal issues, absenteeism, a disciplinary action, etc.).

Based on our survey of the literature, we came up with the following definition for susceptibility.

> *An insider's Susceptibility is the quality or tendency of that insider to become involved in an action that either directly or indirectly affects the organization, due to external or internal influence.*

1.4.3.1 Initial Model

For demonstrating how BKBs can be used to represent the various factors and social processes influencing *Susceptibility*, a personality profile and details from the Aldrich Ames case study[4] was used. Analysis of various case studies and research on people's vulnerabilities revealed that deception, threat, and bribery are some of the most common avenues for *Susceptibility* [12]. As an example, an insider's *Susceptibility* to different tactics, especially to bribery and deception, was modeled in this scenario. In the Aldrich Ames case, his *Susceptibility to bribery* was modeled based on observations that he is a spendthrift character [13] (purchase of new jaguar car) and was going through a divorce which induced the need for money during settlement. *Susceptibility to deception* is modeled based on observations of his drinking problem, and of him being careless and irresponsible in many circumstances. This was evident from the fact that he was involved in a fight during a social gathering. Though the case study was based on Aldrich Ames's character profile, some of the events, such as the inheritance and being demoted, were artificially inserted to demonstrate that our model can handle the behavioral shifts caused by both positive and negative effects of *Susceptibility*.

[4] http://www.wrc.noaa.gov/wrso/security_guide/ames.htm

1.4.3.2 Synthetic Scenario

Similar to our *Predictability* model, we developed BKBs (*Figure 7* and Figure 8) to represent the scenario and time steps in the Aldrich Ames case. In this model, Ames's susceptibility to bribe and deception were used as the target factors. We performed belief updating to analyze how Ames's *Susceptibility* to the target factor changes in response to various events in the scenario. These events can be seen in *Table 4*.

Table 4: Susceptibility Levels

	Target 1- Susceptibility to bribery	Target 2- Susceptibility to deception
Level 1	Scenario: Finance Event1: Buys an expensive car Event2: Divorce Event3: Inheritance	Scenario: Alcoholism Event1: Demoted (Long lunch breaks and working under influence of alcohol) Event2: Paranoid (Misconceives information and has argument in social meeting) Event3: Rehab
Level2a	Scenario 1 & 2: Finance & Alcoholism Event1: Buys an expensive car Event2: Demoted Event3: Divorce Event4: Paranoid Event5: Rehab Event6: Inheritance	Scenario 1 & 2: Finance & Alcoholism Event1: Buys an expensive car Event2: Demoted Event3: Divorce Event4: Paranoid Event5: Rehab Event6: Inheritance
Level2b	Scenario: Alcoholism Event1: Demoted Event2: Paranoid Event3: Rehab	
Level3	Scenario 1 & 2: Finance & Alcoholism Event1: Buys an expensive car Event2: Demoted Event3: Divorce Event4: Paranoid Event5: Rehab Event6: Inheritance	

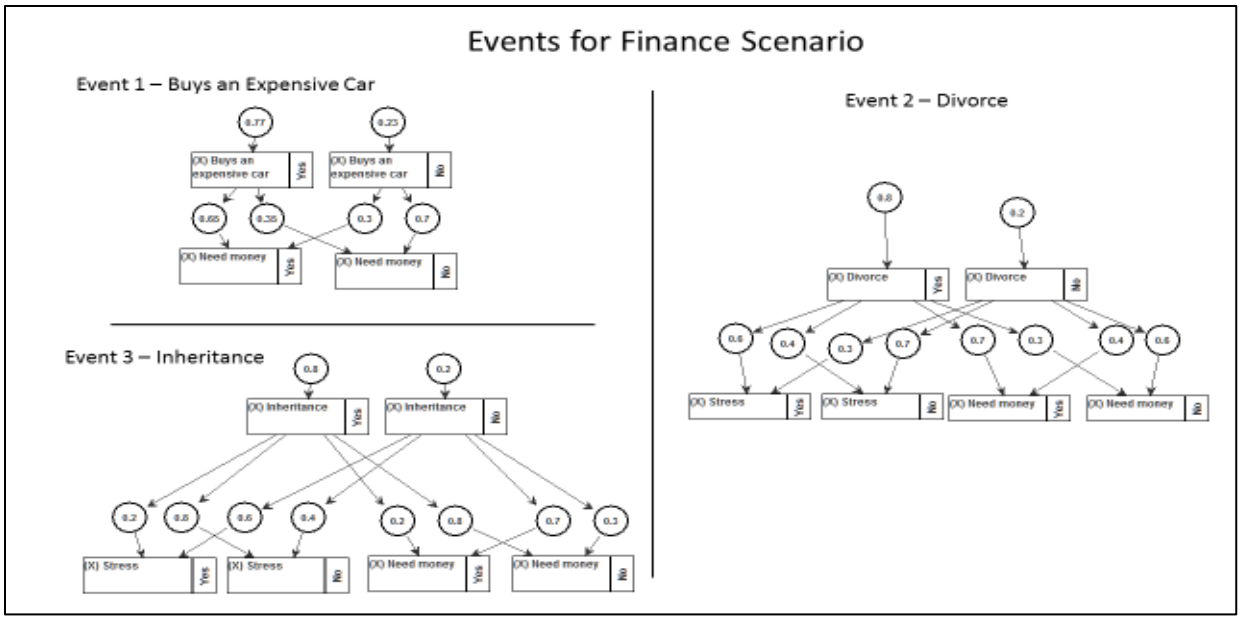


Figure 7: Events for Susceptibility Finance Scenario

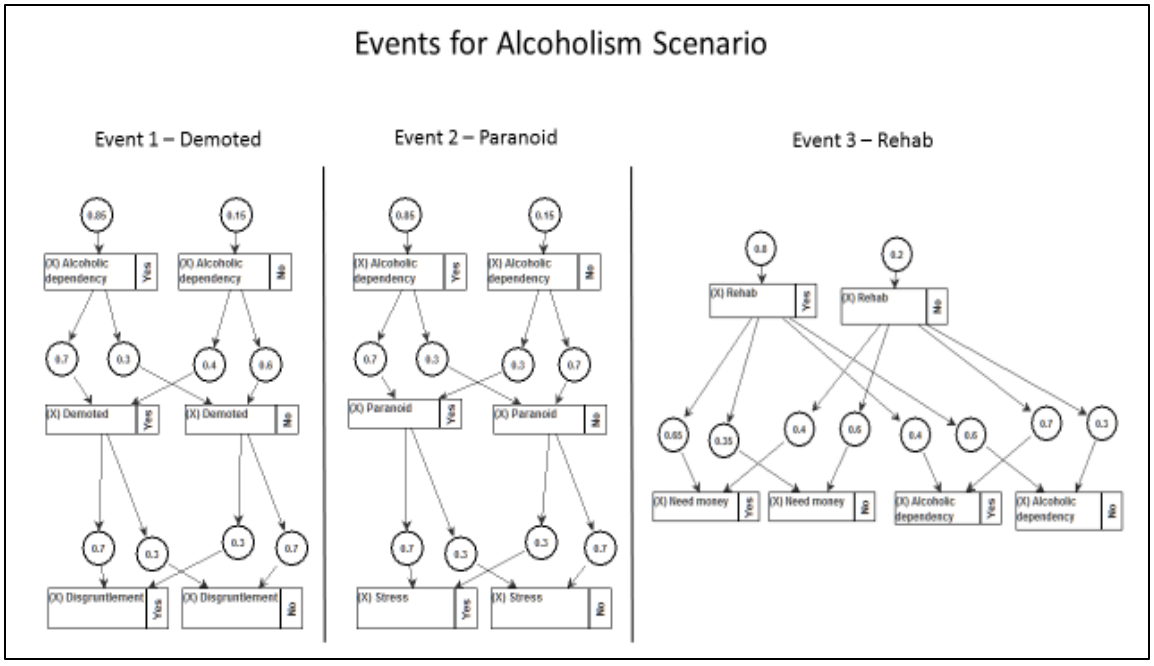


Figure 8: Events for Susceptibility Alcoholism Scenario

1.4.4 AWARENESS DEFINITION AND INITIAL MODEL

Awareness in general is useful and relevant for defeating cyber (and other) attacks, but in the context of identifying the insider's potential threat type, it would be in the context of being manipulated. Here, an insider's *Awareness* was determined to only be relevant if there was an attempt at manipulation. Some of the key factors that we determined to be vital to model an insider's

Awareness were reasoning and understanding ability, memory, training, experience, communication ability, mental state, and information overload.

Based on our survey of the literature, we came up with the following definition for susceptibility.

> *An insider's Awareness is the insider's ability to identify manipulative intent behind false and/or partial information.*

Based on research concerning manipulation techniques, we consider *Awareness* separately for the following different techniques:

- Trust-based Manipulation
  - This type of manipulation usually happens between colleagues or friends. The manipulator observes weaknesses of the victim in daily life, and makes use of them to hide their true purpose. As a result, the victim usually helps the criminal willingly. This type of manipulation often requires a longer time to take effect, but once the criminal gains the trust of the victim, it is often very hard for the victim to discover his/her situation. In fact, the victim may help the criminal even after some dangerous behavior has been caught.
- Empathy-based Manipulation
  - This type of manipulation often happens between strangers. The manipulator pretends to be in a helpless situation while the victim happens to be the only one that can offer help, so that the victim will feel guilty if he/she refuses the manipulator's request. This will often end up with key information leakage, illegal access to critical facilities, or delayed service to normal users. It is usually faster to carry out this type of manipulation, but it is also easier for the victim to find out the manipulator's true purpose. As a result, the manipulator must make several attempts on different targets before he/she can gather all he/she wants.
- False Identity-based Manipulation
  - This type of manipulation often happens online, where the victim cannot verify the manipulator's identity easily and accurately. The manipulator usually pretends to be the victim's friend, supervisor, or customer by forging similar profiles online. The victim can discover it easily if he/she double-checks what has happened with the real person through a different channel. But if the victim is careless and trusts the manipulator without such verification, he/she will follow the manipulator's request and provide classified information willingly. This manipulation type is more difficult than empathy-based manipulation, but easier than trust-based manipulation. It also requires a long time to mimic the other person and forge his/her profiles online. Once the preparation is done, it is faster to lie to the victim and easier to succeed.

We define an individual's overall *Awareness* as the composite of the three different types:

> *Awareness of trust-based manipulation* (A_t)
> *Awareness of empathy-based manipulation* (A_e)
> *Awareness of false identity-based manipulation* (A_f)

A number functions may be considered to best represent the composite *Awareness*. To start, we employ an averaging function $A = (A_t + A_e + A_f)/3$.

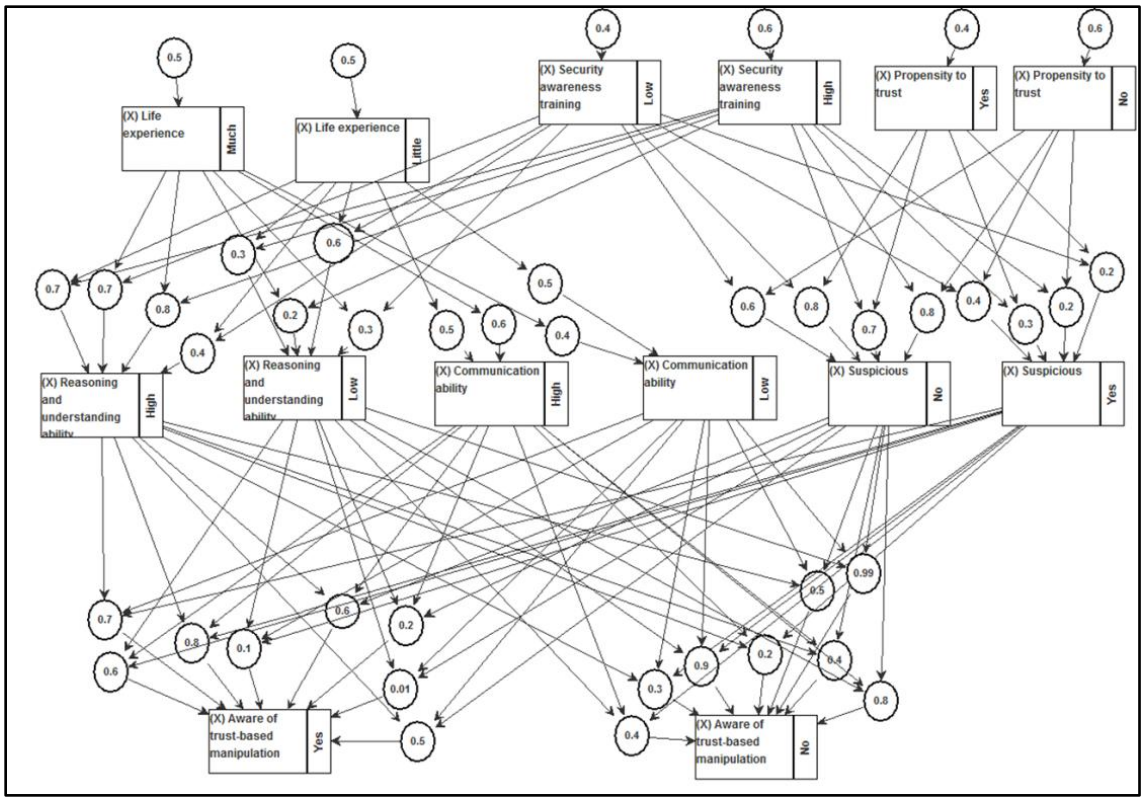


Figure 9: Awareness Baseline BKB

Each type of *Awareness* is influenced by several factors, including reasoning ability, memory, training, experience, communication ability, mental state, and information overload. These factors were built into *Awareness* related BKBs, which give different *Awareness* values based on the current scenario. Due to limited space, only the BKB about trust-based *Awareness* is shown below (*Figure 9*).

To compute each type of *Awareness* value, we build a baseline BKB representing an average *Awareness* value among different people, and event BKBs reflecting a specific person's *Awareness* value at the time of manipulation. In this baseline BKB, we compute the posterior probability of the baseline I-node "(X) aware of trust-based manipulation = Yes", $P(A_{tb})$. Then we fuse the baseline BKB with event BKBs that reflect the scenario's influence on an individual's current *Awareness* state, to get his/her current *Awareness* value, $P(A_{tc})$. The difference between them is the *Awareness* value for that person at that time for that type of manipulation. This value ranges from -1 to +1, where -1 corresponds to totally unaware, and +1 means fully aware, of the manipulation.

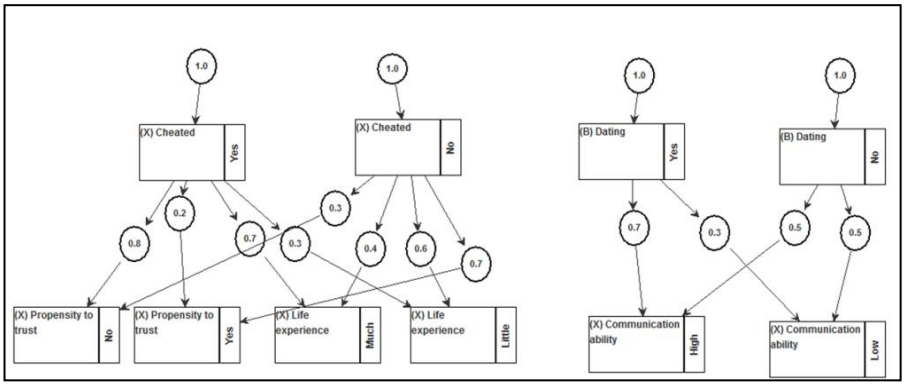


Figure 10: Awareness Scenario 1 Event BKBs

1.4.4.1 Synthetic Scenario

Two basic scenarios were designed to demonstrate a person's *Awareness* for various types of manipulations. These scenarios are summarized in *Table 5*. The event BKBs for each scenario can be found in Figure 10 and Figure 11.

Table 5: Awareness Scenario Descriptions

Scenario #	Brief Description
Scenario I	Dating a new flame, first online, then later in person He spends a lot of money on her Girlfriend deceives him and leaves him He becomes more careful about online dating
Scenario II	Hacker copies victim's Facebook profile Hacker phishes victim's colleagues on twitter using victim's name Hacker requests proprietary documents from colleagues Hacker sells documents to other company

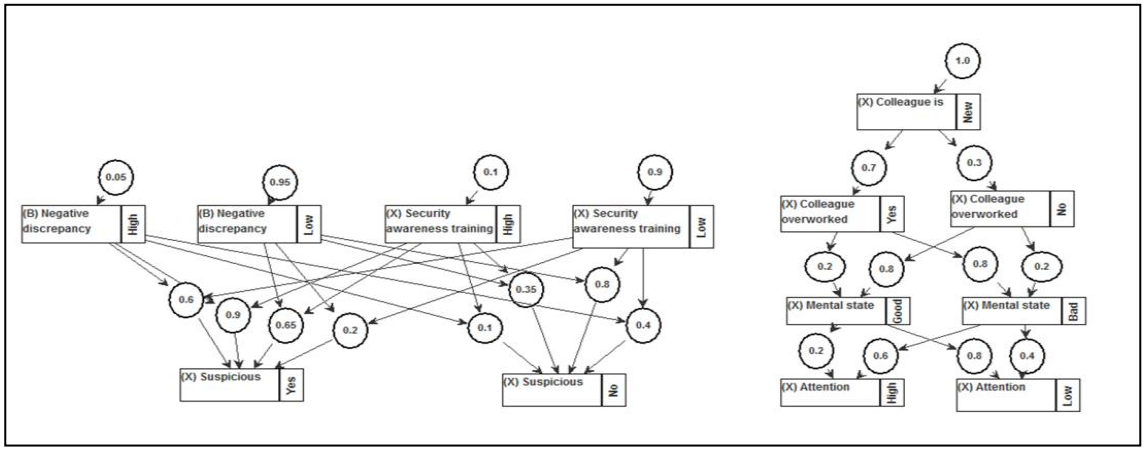


Figure 11: Awareness Scenario 2 Event BKBs

1.4.5 INITIAL PSA MODELING RESULTS

Overall, the results of the first year's research were rewarding in that they provided us additional insights into the level of detail required to usefully inform a model on an insider's actions, thought processes, and background, and thus provide useful insight into the insider's potential type. There are many challenges to be addressed, but we see a way forward. Results for each aspect of *PSA* are summarized below. For necessary brevity, only the results of *Susceptibility* are presented in greater detail.

1.4.5.1 Predictability Results

Our hypothesis was that bias introduces an unknown that results in less *Predictability*, and thus could serve as a baseline measurement for *Predictability*. After applying the calculations from 1.4.1, Scenario 1's results revealed that the outcome was more predictable *without* bias, as we expected.

The Scenario 2 Level 1 calculation yielded a markedly better *Predictability* score than Scenario 1, in direct contradiction to our hypothesis of bias decreasing *Predictability*. It seems that at times, bias can indeed decrease *Predictability*, when the bias is unknown and the outcome of the situation is predictable otherwise. On the other hand, when the outcome is borderline or unclear, and a known bias exists, then *Predictability* may actually increase. Bias does not seem to be the consistent *Predictability* indicator for which we were searching.

The Level 3 calculation reinforced the previous observation. We recognized that bias can both add and subtract from *Predictability*, depending on the level of uncertainty of other factors, and the level of uncertainty with regards to the bias. *Conclusion: Bias cannot serve as a foundation for establishing the Predictability of a potential insider.*

1.4.5.2 Susceptibility Results

On the whole, the results for *Susceptibility* were favorable. For the results of Level 1, where a single scenario is used to model and thereby predict a single target outcome, we found that the Finance Scenario indicated Events 1 and 2 (*Figure 7*) drove his need for money and therefore increased his *Susceptibility to bribery*, whereas the Event 3 inheritance decreased his need for money and therefore reduced his *Susceptibility*.

In Scenario 2, Events 1 and 2 (*Table 6*) increased the insider's *Susceptibility to deception* because of his alcoholism and threat indicators, while Event 3 decreased his alcohol dependency, and therefore reduced his *Susceptibility*.

Table 6: Level 2a - Multiple Scenarios and Single Target

Scenario – Finance and Alcoholism; Target – Susceptible to bribery **or** deception					
		Susceptible to bribery		**Susceptible to deception**	
	Events	**Yes**	**No**	**No**	**Yes**
1	Buys Expensive Car	0.465125	0.534875	0.59875	0.40125
2	Demoted	0.48929	0.51071	0.622915	0.377085
3	Divorce	0.511582	0.488418	0.635665	0.364335
4	Paranoid	0.516098	0.483902	0.640182	0.359818
5	Rehab	0.515083	0.484917	0.590645	0.409355
6	Inheritance	0.483818	0.516182	0.589518	0.410482

Table 6 shows the results of Level 2a where multiple scenarios (Finance and Alcoholism) are modeled together to predict a single outcome (bribery or deception) and *Table 7* shows the results of Level 2b, modeling a single scenario (Alcoholism) to predict the outcome of multiple targets. In *Table 8*, Events 1, 3, and 6 are from the *Finance Scenario*, and Events 2, 4, and 5 are from *Alcoholism Scenario*. Events 1 to 4 increase the insider's threat and *Susceptibility*, because of his increased need for money and alcohol dependencies. Events 5 and 6 reduce both those factors, as the values in the table show. This level shows how interactions between different layers of the model can be captured.

Table 7: Level 2b - Single Scenario and Multiple Targets

Scenario – Alcoholism; Target – Susceptible to bribery **and** deception					
		Susceptible to bribery		**Susceptible to deception**	
	Events	**Yes**	**No**	**Yes**	**No**
1	Demoted	0.495915	0.504085	0.622915	0.377085
2	Paranoid	0.509065	0.490935	0.636065	0.363935
3	Rehab	0.509148	0.490852	0.586398	0.413602

In *Table 7*, the events are from the *Alcoholism Scenario* (contributing directly to deception). Though these events don't contribute directly towards the *Susceptibility to bribery*, they influence the insider's threat level due to Event 1 (demotion causing insider to be disgruntled towards the organization), which in turn affects his *Susceptibility to bribery*. Hence the trend is common for both Susceptibility values.

Table 8: Level 3 - Multiple Scenarios and Multiple Targets

Scenario – Finance and Alcoholism; Target – Susceptible to bribery **and** deception					
		Susceptible to bribery		**Susceptible to deception**	
	Events	**Yes**	**No**	**Yes**	**No**
1	Buys an Expensive Car	0.465125	0.534875	0.59875	0.40125
2	Demoted	0.48929	0.51071	0.622915	0.377085
3	Divorce	0.511582	0.488418	0.635665	0.364335
4	Paranoid	0.516098	0.483902	0.640182	0.359818
5	Rehab	0.515083	0.484917	0.590645	0.409355
6	Inheritance	0.483818	0.516182	0.589518	0.410482

In Table 8, Level 3 captures how multiple scenarios interact and influence the outcome of multiple targets. This level shows the complete trend in the *Susceptibility* values across all scenarios whereas previous levels show more fine grained and selective view of the expected outcome.

1.4.5.3 Awareness Results

The *Awareness* calculation results indicated that, for Scenario 1, the subject's experience of being cheated on increased his awareness in the future. In Scenario 2, due to his bad mental state and lack of training, the individual's *Awareness* was less than the average IT worker. These are encouraging results which we hope to build upon.

1.4.6 PSA FRAMEWORK REFINEMENT

Following the efforts from 2013, where much investigation into the individual components of *Predictability*, *Susceptibility*, and *Awareness* (*PSA*) was conducted, our first priority for 2014 was to

establish how *PSA* related conceptually to our emerging model, including how the *PSA* components related to each other.

It was identified early on that, while the components of *PSA* had many unique qualities that require individual focus, they are also not necessarily orthogonal. They will have some common indicators which should be leveraged to yield the maximum return on very sparse information available for typical insider threat situations.

1.4.6.1 Refinement: Predictability

Impulsivity was identified as a key attribute that could represent *Predictability* in social science literature [14]. Psychiatric aspects of impulsivity were studied and three broad categories were identified: punished or extinction based paradigms (here impulsivity is defined as the perseverance of a response that is punished or unrewarded), reward-choice paradigms (impulsivity is defined as preference for a small immediate reward over a larger delayed reward), response disinhibition and attentional paradigms (here impulsivity is defined either as making responses that are premature or as the inability to withhold a response).

Four factors from the FFM, namely premeditation, perseverance, urgency and sensation seeking, were identified as suitable factors that could be used to estimate impulsivity [15]. The primary factor, (lack of) premeditation, captured the most frequent conceptualization of impulsivity. The inverse of premeditation could be the factor that is most relevant to our need to define, quantify, and measure *Predictability*. The second factor, urgency, appeared to reflect a tendency to commit rash or regrettable actions as a result of intense negative affect. This might also correspond to *Susceptibility* for a potential insider. The third factor, sensation seeking, was comprised of scales measuring the tendency to seek excitement and adventure.

Initially we focused on premeditation and perseverance as possible indicators of *Predictability*. Impulsivity and premeditative variables (see Figure 12) were then used to model predictability levels. We also developed preliminary BKBs based on the related FFM factors, conscientiousness and extraversion, to feed into the impulsivity and premeditative behavior (see Figure 14). The 16PF Questionnaire is also considered as a possible gauge for impulsivity and thus *Predictability*. Figure 13 shows a sample BKB for measuring *Predictability* using 16PF questionnaire.

1.4.6.2 Refinement: Susceptibility

For modeling *Susceptibility* in the context of insider behavior, two avenues were proposed: situation specific and non-situation specific avenues. Situation specific avenues are those that are more dynamic and can be influenced by the environment in which the insider is found. In particular, emotions are dynamic and accessible avenues for manipulating insiders. Emotions such as disgruntlement and anger are commonly found among destructive cyber insiders [16], [17]. Moreover, there have been some clear ties identified between emotions, personality, and insider threats [18]–[20].

To bring all of this together, factors from FFM were examined for modeling *Susceptibility*. Agreeableness and neuroticism were identified to be closely associated in evaluating the *Susceptibility* of an individual. Figure 15 shows a baseline BKB fragment for modeling *Susceptibility* in general, while Figure 16 shows how FFM might factor into a more specific model for *Susceptibility*.

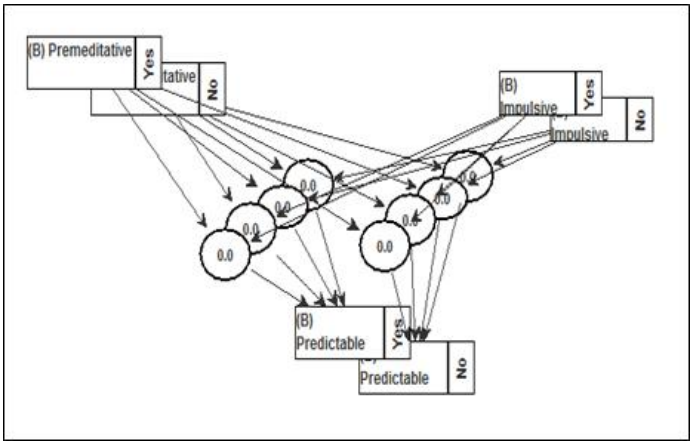


Figure 12: Predictability BKB based on impulsive and premeditative behavior

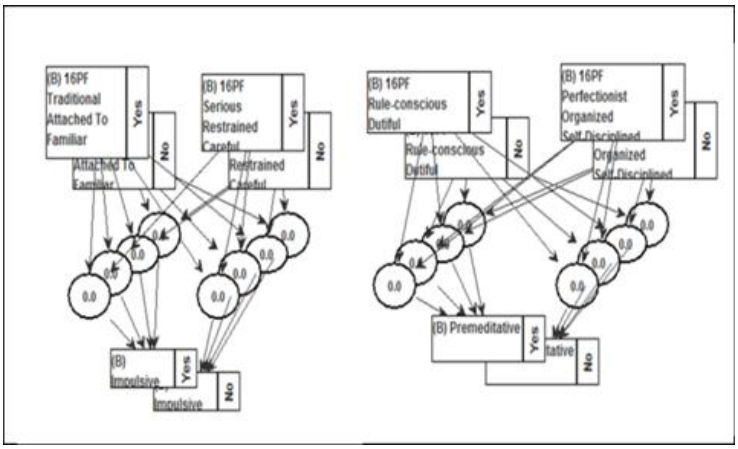


Figure 13: Predictability BKB based on 16PF

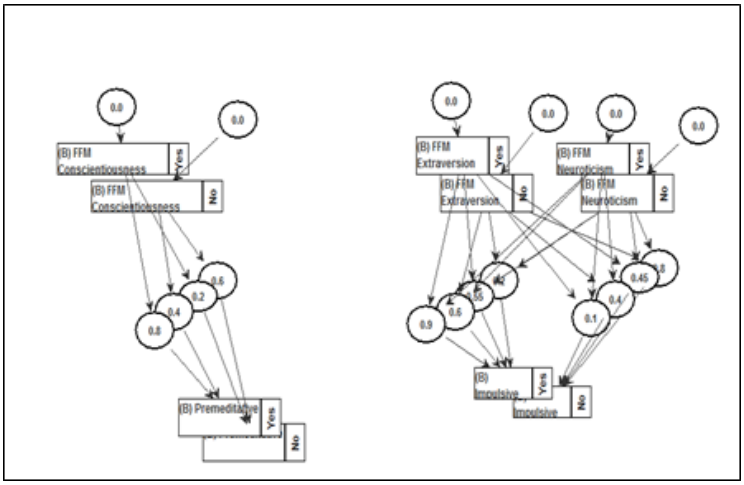


Figure 14: Predictability BKB based on FFM

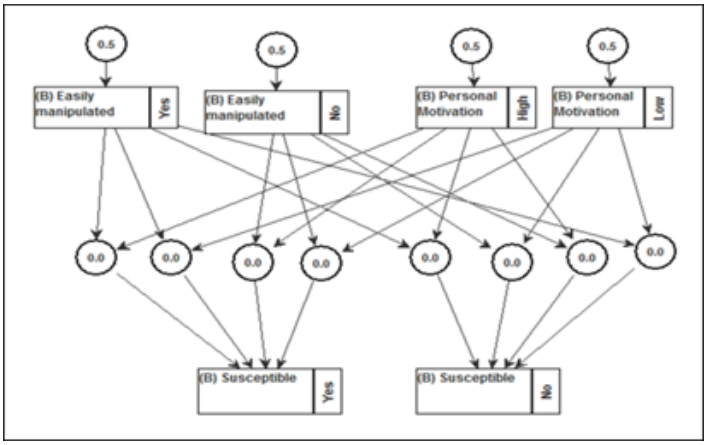


Figure 15: Susceptibility Baseline BKB

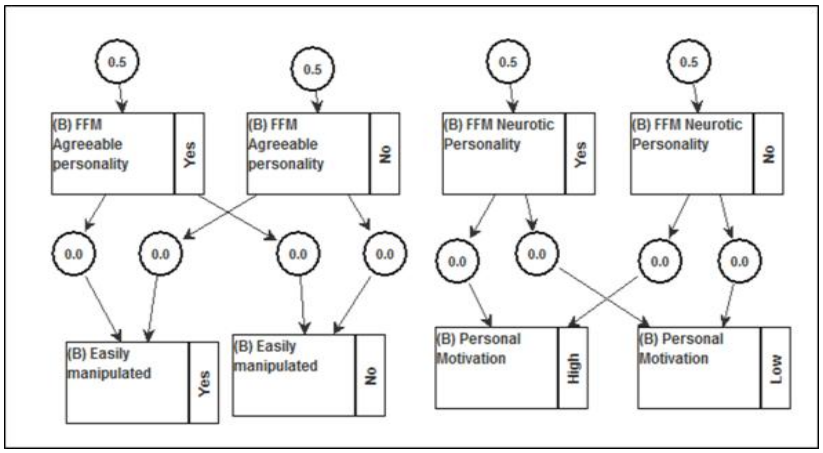


Figure 16: Susceptibility BKB based on FFM

1.4.6.3 Refinement: Awareness

Awareness of cyber-related attack is related to several fields including human behavior patterns, human reasoning patterns, and attacking techniques. In each field, several widely-accepted models have been proposed. Therefore, we leverage factors extracted from these models to measure an individual's *Awareness* of cyber-attacks.

To model human behavior in general, we briefly introduce four typical behavioral models:

- Aggregation;
- Reasoned Action;
- Planned Behavior;
- Affect Infusion.

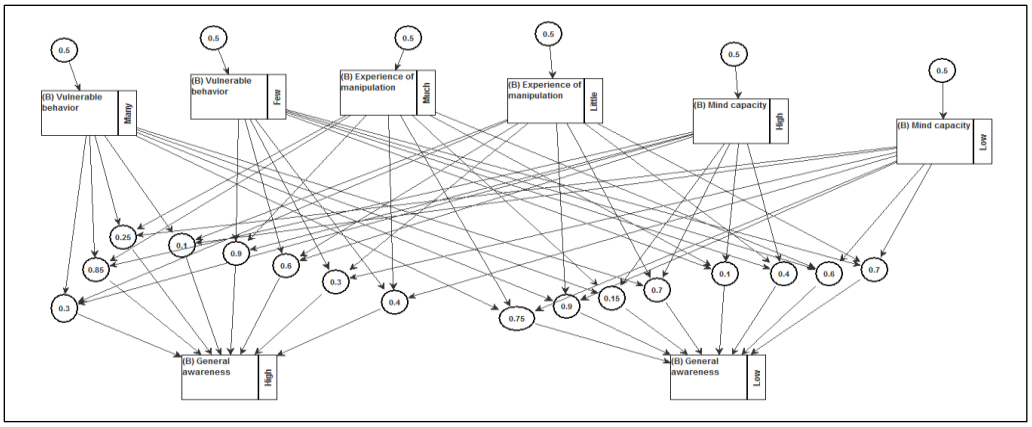


Figure 17: Awareness Baseline BKB

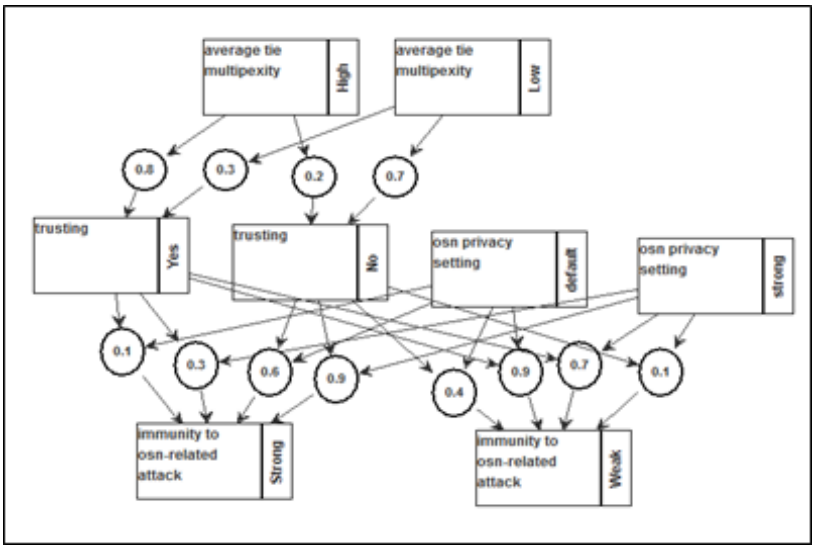


Figure 18: OSN BKB-1

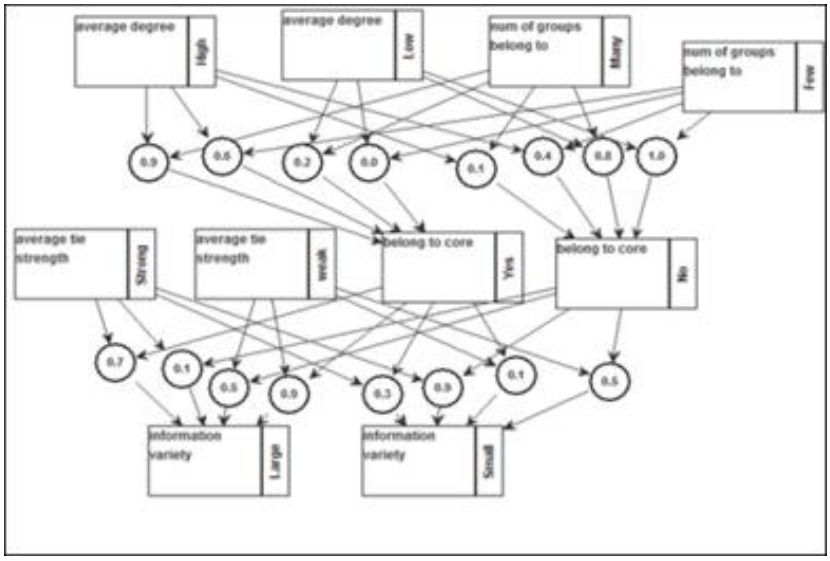


Figure 19: OSN BKB-2

The first model,-Aggregation, assumes that human behavior is formed through aggregating all kinds of stimuli. These stimuli include various occasions, judgments, opinions, and modes. The benefit of aggregation is to cancel out the randomness within each scenario, and extract the common features of all situations. The second model assumes that human behavior is controlled by three factors: attitudes, subjective norms, and behavioral intention. The third model is similar to the second one except that it includes perceived behavior control. The fourth model assumes cognition and affect are related. In this model, people sometimes behave according to experience, and sometimes learn and reason upon the specific situation, based on the complexity and familiarity of the scenario. All of these models have found convincing evidence consistent with observation, but none of them can explain all behaviors sufficiently when applied alone. For instance, altruistic behavior is hard to explain by the reasoned behavior model, and impulsive reaction is difficult to explain by the planned behavior models. Therefore, we introduce all of these models into *Awareness* fragments to cover all situations. Aspects of these processes are reflected in the *Awareness* baseline BKB (see Figure 17).

To model the transmission procedure of information across networks (social networks, computer networks, or a mixture of both), we also consider approaches to dealing with human information processing behaviors. The first method we reviewed is related to the need for cognition, and it claims that an intrinsic motivation is for the process of seeking the truth. It believes that people in low need of cognition prefer heuristic thinking, while those with a high need like to think and consider. The second method concerns information sharing on online social networks (OSNs). People form ties in networks in the form of strong ties and weak ties. Further it claims that weak ties can spread more information to a larger range. At the same time, privacy concerns regarding the sharing process change inside and outside friendship circles. The third model is related to identity theft. Fake profiles on the same OSN or across OSNs are made to steal personal information illegally by criminals. In general, the information diffusion process in cyberspace occurs in both legal and illegal forms, and awareness of information leakage in both forms becomes an essential part of cyber-attack *Awareness*.

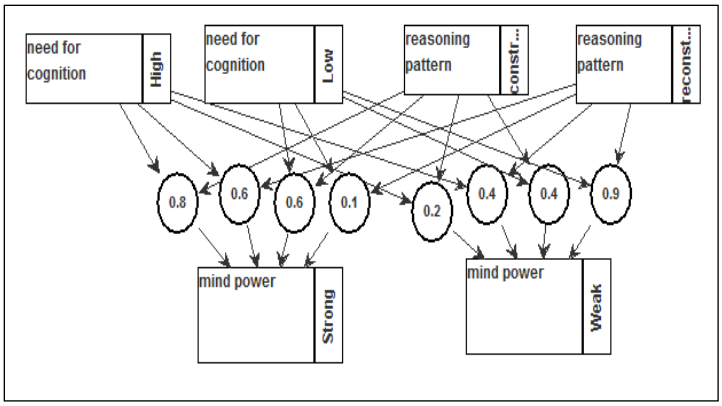


Figure 20: PSY BKB-1

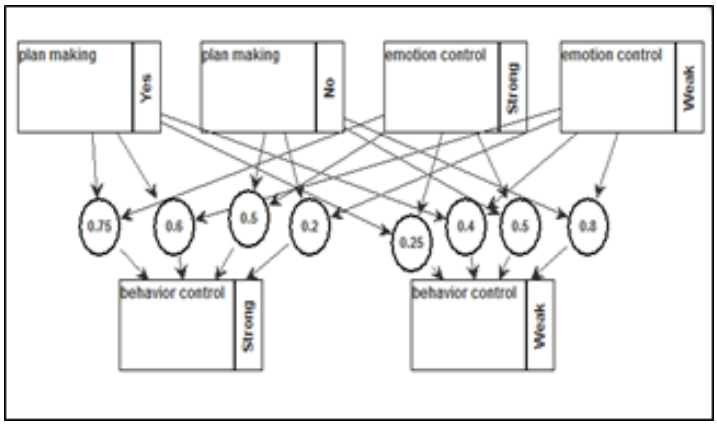


Figure 21: PSY BKB-2

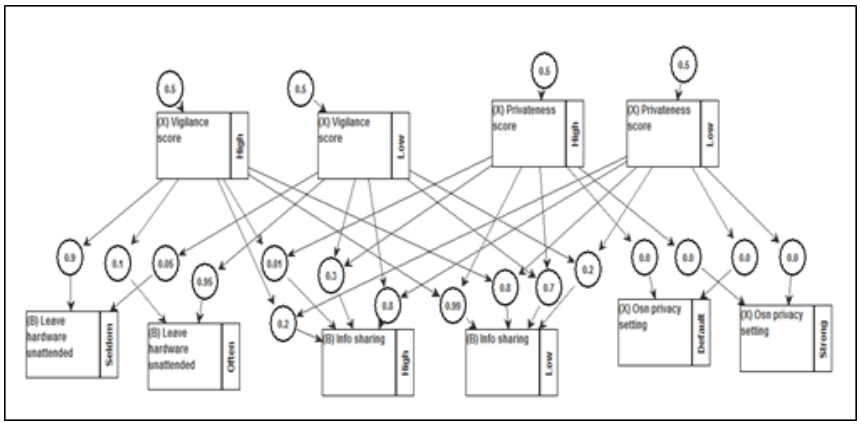


Figure 22: Awareness – Vigilance and Privateness

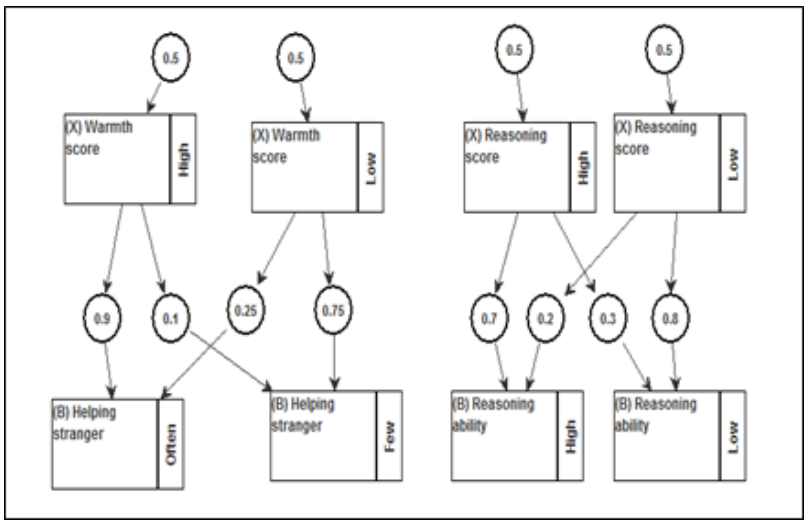


Figure 23: Awareness – 16PF Warmth and Reasoning

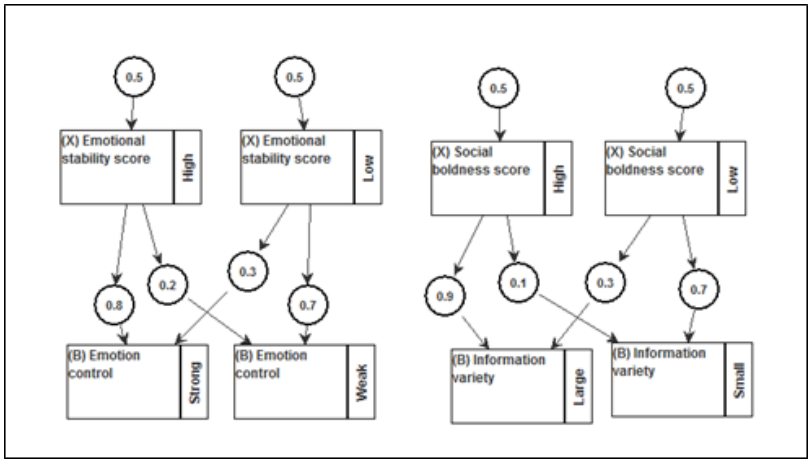


Figure 24: Awareness – 16PF Emotional Stability and Boldness

The third field is about social engineering models. These models demonstrate how criminals make use of all possible weaknesses of a person, such as his empathy to the needy, his lack of defense to colleagues, or even his carelessness to protect devices in the working environment. The first model deals with traditional social engineering approaches, such as information gathering through persuasion. The second model deals with high-tech information elicitation such as USB route, email fraud, or fake websites that propagate viruses. The third model includes false authority and affective commitment. In general, all these models share some features: psychological factors, behavioral factors, technological factors, and social factors. All of these means are considered in the *Awareness* model as well.

Based on the source of cyber-attacks, we divide relevant factors into OSN-related, psychology (PSY) –related, and work-related *Awareness*. We leveraged these concepts in our representations for *Awareness* by constructing BKBs for each type.

OSN-related BKBs represent factors influencing a person's state in OSNs. Relevant factors describe the number of ties built, the number of groups to which belonged, and the average strength of formed social ties. It also includes how personal account security levels are set up, and how friends and strangers are trusted. These factors all influence immunity to OSN-related attacks, and also information diversity and depth. See Figure 18 and Figure 19 for example details.

PSY-related BKBs reflect a person's psychological state, including their need for cognition, reasoning pattern, plan making ability, and emotion control skills. All of these factors will decide how hard the individual is to be manipulated emotionally or psychologically by adversaries. Work-related BKBs describe how likely a person is to give away sensitive information at work, such as a computer password or access to classified facilities. In short, these BKBs represent a person's *Awareness* to all types of cyber-related manipulations in general. See Figure 20 and Figure 21. Likewise, the work-related BKB can be found in Figure 22.

As with *Predictability* and *Susceptibility*, BKB fragments representing personality factors relevant to awareness were developed. Those BKBs may be found in Figure 23 and Figure 24.

1.5 CONCLUSION AND FUTURE DIRECTIONS

This study has helped us to better understand the factors that make a person *Predictable*, *Susceptible*, and *Aware,* and how those traits might be identified. While the *Predictability* model ultimately did not reveal the underlying bias association we had hypothesized, it provided essential insight into the obstacles to be addressed in sufficiently capturing details of an insider's character necessary for measuring all aspects of *PSA*. Our BKB model on *Susceptibility* showed that we can capture interactions between scenarios at various levels and use them to predict the expected outcome of *Susceptibility*. Finally, the *Awareness* scenarios revealed the possibility of measuring different types of *Awareness* influences, both from experience on the job, and in personal life, can affect an individual's *Awareness* to manipulation and deceit.

Our primary objective for 2014 was to build a common, static base for the components of *PSA* upon which to build the more dynamic aspects of the *PSA* components. To accomplish this, we focused primarily on integrating proven measures of personality into our model for *PSA*, which we can then tie to defining potential insider types. While a number of measures were investigated, the Five Factor Model (FFM) and the Sixteen Personality Factors (16PF) were selected as the most promising for inclusion in our insider type model. In addition, we furthered our understanding of the dynamic nature of *PSA* by investigating aspects of emotion which could contribute to determining potential insider types. The models thus produced have provided a firm footing for continuing investigations regarding *PSA*'s relationship to the insider types proposed in the CTS project proposal.

We anticipate that the next steps in this investigation will include producing a model that can incorporate measures from a number of sources such as traditional personality typing instruments and real time information of the environment, generate relevant measures, and map individuals to the potential insider types defined in the CTS project proposal. This will entail devising a method to incite participants to reveal their potential insider type, and to associate those revealed types with personality traits, emotion and other dynamic factors, and *PSA* profiles. Naturally, individuals are reluctant to make known any proclivities towards undesirable insider behavior, so this remains a formidable challenge.

THRUST 2 – TARGETED INTERVENTIONS DERIVED FROM BIOMARKERS OF CYBER TRUST

The research efforts for Thrust 2 was led by the University of Tulsa with Dr. John Hale as the principal investigator (PI). The primary focus for Thrust 2 is to investigate the ability to affect and influence Cyber Trust through interventions targeted at relevant biomarkers. The first goal is to lay the foundation for identifying biomarkers for Cyber Trust. The descriptions and results in this chapter were provided by the Thrust 2 PI and Thrust 2 team.

2.1 INTRODUCTION

It is well-known that humans are the weakest link in any cyber system. Adversaries exploit decision-making processes of users: preying on their propensity to trust based on certain learned contextual cues and environmental influences. Despite this, studies of the psychological, biological, and technological influences on the cognitive phenomena of cyber trust are sparse.

This 2-year research effort built a research foundation upon which biomarkers of cyber trust for targeted interventions can be explored. The primary contributions from this research include:

1. Perceptual Cyber Trust Taxonomy.
2. Experimental design for identifying neural correlates of trust decisions in a cyber context.
3. Simulation platform for Cyber Trust research ("The CyberPhishing Game").
4. Data and preliminary analysis of Cyber Trust studies.

To pursue analysis of data and classification of individuals for the purposes of developing a biomarker called for the construction of a taxonomy with which to capture salient characteristics of trustworthy and untrustworthy digital content. This led to the development of the Perceptual Cyber Trust Taxonomy, which evaluates the trustworthiness of content on the basis of lures and cues as perceived by end users. The taxonomy is embodied as a Document Type Definition file in XML format, and thus can be used programmatically to annotate, interpret, and operationalize trust characteristics of online artifacts.

In Year 1, the neuroimage Cyber Trust study conducted fMRI scans to localize neural networks of interpersonal trust, reward, and theory of mind (ability to infer the mental states of others) in 52 subjects. We hypothesized that the neural cognitive elements of cyber trust are comprised of interpersonal trust, reward decisions, and theory of mind assessments. Moreover, we hypothesized that cyber trust decisions activate a conglomerate of neural networks associated with these component processes. To test this hypothesis, subjects were scanned again as they played a game that confronted them with trust decisions for email, the web, social networks, and applications. This sets the stage for analysis: to assess the neural correlates of cyber trust the activated neural network can be characterized as the weighted sum of the proposed component networks. An additional study made use of an eye tracker to compare the performance of 94 expert and novice users in assessing the trustworthiness of online content and messages.

This project also involved the development of a simulation platform to serve as the basis for a CyberPhishing Game that immersed players in a virtual enterprise, challenging them to make Cyber Trust decisions in the operation of a business. The game was designed to assess the ability of

individuals to make Cyber Trust decision in the presence of various cues and lures from the taxonomy. The simulation platform provides the capability to simulate emails, tweets, and generic web sites with varying levels of trustworthiness. It also functions as a platform for Cyber Trust training and education. Supplementary training content, surveys, and presentations were also developed to fulfill requirements for the Cyber Trust studies.

2.2 BACKGROUND

This effort lays the foundation for measuring an individual's propensity for Cyber Trust using three modalities: 1) behavioral measures including psychological profiling and task response characteristics, 2) biochemical and physiological states, and 3) neuroimaging. Neurocognitive and psychiatric measures are assessed using self-report questionnaires, interviews, a computerized assessment battery, and observation of the individual's actions in different scenarios involving trust. Biochemical and physiological measures include blood serum levels of neuropeptides, inflammatory markers, hormone levels, and glucose metabolism. Additionally, we monitor respiration and heart rate during neuroimaging. This project's neuroimaging modality utilized blood oxygenation level dependent (BOLD) functional magnetic resonance imaging (fMRI) to measure brain activity. Structural MRI was also obtained for quantification of cortical thickness and volume, and to serve as an underlay for functional activation maps.

2.2.1 BEHAVIORAL AND PSYCHOLOGICAL ASPECTS OF TRUST

Behavioral measures meant to predict an individual's likelihood to trust are often grouped into two categories: experience and disposition. Disposition refers to an individual's inherent characteristics which are determined by numerous biological and environmental factors over the course of years. These characteristics include things like propensity to trust, honesty, and risk aversion. In contrast to dispositional characteristics, which broadly apply to different areas, experience is domain specific. For example, expertise with computers is likely to affect an individual's trust in the cyber domain, but it is unlikely to affect trust in face-to-face communication. In one study, 299 subjects were given questionnaires measuring experiential factors including computer self-efficacy, web experience, and security knowledge, as well as dispositional factors including trust, risk aversion, and suspicion [21]. Each participant was a student in an information systems class and was given a unique code used to access course materials and tests. They completed coursework covering internet security and phishing, and were reminded daily never to divulge their codes to anyone. At the end of the course, a phishing email was sent to each subject requesting their secret code. Approximately one third of the subjects failed to recognize the phishing attack and responded with their codes. The main result from this study was that experiential factors had a much larger impact on phishing success than dispositional ones, indicating the utility training might have improving cybersecurity.

2.2.2 ENDOCRINOLOGY AND TRUST

Studies of the endocrinology of trust have revealed a number of associations with specific hormones and neuropeptides. Oxytocin (OXT) seems to be the most important hormone relating to trust, with higher levels of OXT associated with higher levels of trust [22][23][24][25]. In fact, intranasal administration of OXT has been shown to increase trust behavior relating to social risks in humans without increasing willingness to take other types of risks [26]. Testosterone (TES) tends to make people less trusting [27][28], while estrogen (EST) tends to make people more trusting [29]. The effects due to EST and TES may largely be due to their opposite effects on OXT levels, with TES being associated with decreased OXT and EST being associated with increased OXT [30]. Arginine vasopressin (AVP) has been associated with antisocial behavior and distrust [31][32] and has been characterized as the adversary of OXT [30]. The stress hormone cortisol (COR) has been associated with distrust [33][34]. COR's effects are complicated by its relationships to neuropeptides:

intranasal AVP has been shown to increase COR levels [35][36], while intranasal OXT reduces COR levels [37][33]. Dopamine (DOP), which is commonly associated with reward processing [38], has also been associated with trust behavior, but this relationship is again complicated by relationships between DOP, COR [39], and OXT [40]. Serotonin (SER) has a negative relationship with AVP [41] and thus has been shown to have a positive relationship with trust [42].

The most important findings in endocrinology can be summarized by listing those compounds associated with trust and those associated with distrust. Four compounds have been identified that positively influence trust: EST, OXT, DOP, and SER. Three compounds negatively influence trust: AVP, COR, and TES. The exact relationships between these seven compounds and trust behavior are complicated by the interactions among them. Of the compounds listed, OXT seems to have the most direct influence on trust, and unsurprisingly it also has the most interactions with the other compounds.

2.2.3 NEUROIMAGING AND TRUST

The advent of fMRI has provided an unmatched environment in which to study the relationships between psychological processes or behaviors and specific brain regions. Utilizing this environment, there is a growing number of studies linking particular brain regions to trust behavior.

By far the most popular paradigm used when studying trust in fMRI is the trust game, which mimics asynchronous economic exchanges without contract enforcement [43][44]. Using the trust game paradigm, activity in the caudate nucleus has been associated with benevolence, cooperation, fairness, and the intention to repay [45]. Activation in the caudate has also been interpreted as signaling trust and learning a partner's trustworthiness [46]. Another study showed that the effect of feedback mechanisms in the caudate are modulated by prior moral and social perceptions [47]. Specifically, activity in the caudate differentiated positive from negative feedback only in the case where players were confronted with a morally neutral player and not when confronted with players described as morally good or bad.

A particularly interesting study combined fMRI with intranasal administration of OXT prior to playing the trust game [48]. Compared to placebo, OXT subjects showed decreased sensitivity to having their trust breached. Subjects who received OXT had reduced activity in the amygdala, midbrain, and the dorsal striatum. OXT's effect on trust can then be interpreted as being partly due to modulation of fear processing regions (amygdala and midbrain) and regions associated with behavioral adaptations to feedback (caudate nucleus in the striatum).

There have been a limited number of neuroimaging studies examining trust in the cyber domain. One study used eBay feedback profiles as a stimulus with varying numbers of positive, negative, and neutral comments in order to vary the perceived trustworthiness of different sellers [49]. This study found trust to be linked to activation in regions associated with anticipating rewards (caudate nucleus), predicting the behavior of others (anterior paracingulate cortex), and calculating uncertainty (orbitofrontal cortex). Distrust was linked to regions associated with negative emotions (amygdala) and the fear of loss (insular cortex). Another imaging study used eBay product description texts with varying degrees of trustworthiness as a stimulus [50]. Results were similar to the previous study, specifically that trust was associated with activation of reward processing areas (striatum and thalamus) and mentalizing areas (prefrontal regions and cingulate cortex). Distrust was again linked to regions associated with uncertainty including the insular cortex.

The results of the imaging studies relating to trust can be summarized as implicating three main brain systems. Increased trust is associated with activity in the striatum, which is associated with reward processing, and activity in the frontal cortex, which is associated with deliberate thinking and

mentalization. Activity in the limbic system, which deals with processing of fear, risk, uncertainty, cognitive conflict, and memory, is negatively correlated with trust. Interestingly, dysregulation of the limbic system has also been associated with mood disorders [51][52][53].

2.2.4 INTERVENTIONS FOR CYBER TRUST

Interventions for Cyber Trust can either be targeted toward changing the individual or compensating for the individual's deficiencies. Alert and warning systems help an individual to avoid Cyber Trust mistakes. Training changes the individual to be better at making the correct Cyber Trust decisions. Cyber Trust training typically focuses on either awareness or skill development [54][55]. While awareness can be trained using standard classroom or online lecture formats, skill development requires interactive practice with Cyber Trust decisions. The majority of Cyber Trust interventions focus on awareness training that can be effective in changing some Cyber Trust behaviors. Relatively fewer programs are geared toward skill development through practice and feedback [56]. Simulations are an ideal training format for moving beyond awareness toward developing adaptive expertise in a specific performance domain [57].

Simulations and other synthetic learning environments have the advantages of adding practice environments for developing specific knowledge and skill, monitoring ongoing trainee performance, diagnosing levels of mastery and performance deficiencies, and providing feedback and adapting instruction to needed areas [58]. Simulations can also operate in an instructorless environment and create engaging and realistic learning environments, and allow for learner control [59]. Effective Cyber Trust training seems to require elements that are aligned with simulation capabilities such as feedback, hands-on practice, and meaningfulness [60].

While these theoretical advantages of simulations are strong, there has not been much research into which elements of simulations make them effective for training. This is especially true for the higher level training goals of monitoring, diagnosing, and adapting behaviors that can be developed in a realistic simulation environment [58]. This research builds on what is currently used as best practice in cybersecurity training and combines that with best practices in simulation development to iteratively create an effective framework for developing expert performance in Cyber Trust.

2.3 APPROACH

The fundamental goal of this effort was to lay the foundation for identifying biomarkers for Cyber Trust. The approach adopted by the investigators on the project pursued research objectives along three parallel fronts; 1) the development of a taxonomy for Cyber Trust, 2) a framework for the design of Cyber Trust experiments in neuroimage and behavioral studies, and 3) the construction of a simulation platform suitable for assessing and improving the performance of end-users in making trust decisions online.

2.3.1 TAXONOMY DEVELOPMENT

The work underpinning this project focuses on creating a foundation for designing and implementing Cyber Trust experiments. This required the development of a taxonomy of the types of modalities and cues that end users are required to process. The approach to taxonomy and material development has included the following steps:

1. Collect a large sample of phishing attempts and malware upload attempts from real world examples.

2. Construct a taxonomy categorizing these attempts by modality, decision type, and attack strategy.

3. Construct examples of these attempts for experimentation within the fMRI and eye tracker studies.

4. Construct a system to vary the salience of lures and cues within the examples to adjust the difficulty of identifying the phishing attempts for distinct testing approaches (e.g., fMRI and eye tracker studies).

While most taxonomies in cybersecurity focus on the technical aspects of the subject matter, the objective of developing this taxonomy is to support subject studies that seek to understand the influences of trust elements found in artifacts on user populations. Therefore, the taxonomy as developed should reflect the perceptual aspects of Cyber Trust.

2.3.2 CYBER TRUST EXPERIMENTAL DESIGN

Our approach to developing viable design frameworks for Cyber Trust neuroimage and behavioral studies begins with understanding that social interactions in cyberspace can take place in many domains. Therefore, Cyber Trust decisions involve encounters with a variety of entities including institutions, and known and unknown individuals. Because Cyber Trust involves multiple entity interaction types we propose that Cyber Trust decisions are composed of a conglomeration of trust types and other cognitive processes. Thus, our approach to understanding the neural mechanisms of Cyber Trust is to decompose the functional activation maps derived from functional MRIs of an adapted version of the "Cyber Trust Game" into three main components: interpersonal trust, reward, and theory of mind (Phase 1). Phase 2 would again use fMRI, this time to explore the effect of interventions on neural networks implicated in trust decisions.

The first step toward decomposing the Cyber Trust network into multiple component processes of interpersonal trust, reward, and theory of mind is to perform functional imaging that localizes the neural networks for each of the component processes. Prior to imaging of the "Cyber Trust Game," we use one fMRI scanning session to define the neural networks underlying reward, interpersonal trust, and theory of mind for each subject. While in the scanner, participants will be presented with and perform an adapted version of the Cambridge gambling task [61]. To define the interpersonal trust neural network for each subject, we will have subjects perform a sequential reciprocal-exchange trust game known as the "Trust Game" [47]. Finally, to define the theory of mind neural network, subjects will perform an eye gaze task developed by Baron-Cohen [62].

Following the reward, interpersonal trust, and theory of mind localizer scanning session, participants return for another fMRI scanning session. In this second scanning session participants are scanned while performing a scanner-friendly version of a Cyber Trust Game task. Participants should be naïve to the Cyber Trust Game task, beyond an instructional period immediately prior to scanning. In the scanner version of Cyber Trust Game, participants are to be presented with a predetermined order and number of the simplified versions of simulated emails, web pages, social network contacts, and download requests. For the initial fMRI analysis, each of these cyber modalities were to be combined in to one "grand" modality. The stimuli from each modality contains either a high or low number of risk cues, and the participants are prompted to respond to either "accept" or "reject" the email, web site, social network, or download request. This leaves us with 4 trial types: Low-Risk:Trusted, Low-Risk:Not-Trusted, High-Risk:Trusted, High-Risk:Not-Trusted. The neural networks for each of these four conditions can then be defined and contrasted for developing the cyber trust neural network deliverable, and each of the four networks will be decomposed into similarity to reward, interpersonal trust, and theory of mind using a machine-learning algorithm.

In Phase 2, we planned to recruit and assess a new set of subjects to replicate both the biomarker and neural network model of cyber trust developed in Phase 1. Phase 2 analyses expanded on Phase 1 by investigating the effect of cyber security training and Cyber Trust Game playing experience to quantify changes in behavior and in functional activation of the cyber trust neural network. We hypothesized that post training our neuroimaging biomarker would classify all subjects into one cluster rather than the classification of multiple clusters of subjects prior to training.

Phase 2 could be used to develop a non-neuroimaging biomarker of cyber trust propensity using behavioral and physiological measures. This non-neuroimaging biomarker can be developed in parallel with the neuroimaging biomarker and is designed to provide detection of Cyber Trust propensity without having to perform an fMRI. The non-neuroimaging biomarker could be developed to mimic the classification of participants into multiple Cyber Trust propensity type groups based on neuroimaging and behavioral results.

The approach for development of the non-neuroimaging biomarker is to collect psychiatric (mood scales, family history, etc.), neuropsychological (intelligence testing, reward sensitivity, etc.), and physiological measures (oxytocin, testosterone, estrogen etc.) that have previously been associated with trust, reward sensitivity, or suspicion prior to the initial neuroimaging scanning session. The results of these non-neuroimaging factors will then be explanatory factor inputs into kernel canonical correlation analysis of the neuroimaging classifications.

2.3.3 SIMULATION PLATFORM CONSTRUCTION

The approach to building a simulation platform calls for both a simple form and a free play online game that immerses player-participants in a business environment. The environment confronts players with a variety of operational and transactional decisions, allowing online actions marked by the influence of trust and suspicion. Participants engage simulated versions of email, the web, and social networks to conduct business through relationship building, financial transactions, and company politics.

The business environment is a big box lumber and home store. Two types of simulations are planned. One simulation, the independent game, enforces participant engagement, independent of what is happening in the environment. Thus, limited feedback on trust decisions are provided. The second simulation, the interactive game, allows the participant to engage with the environment more directly and base their gameplay on the results of their trust decisions.

The independent game presents a sequence of independent trust decision events, using an economic environment as a cinematic backdrop incentivizing thoughtful and attentive play. This approach is suitable for evaluating individual trust propensity in a controlled manner since there is limited feedback. The interactive game, which can also be adaptive, permits exploration of a wider spectrum of Cyber Trust issues, including the effects of interventions. Game play ends when all events have been processed and the composite value (and therefore the final score) of each participant is computed as a function of monetary value and inventory. Secondary scoring values can be derived as functions over reputation and productivity.

Participants interact within the game via a dashboard interface that exposes communication, transaction, and scoring functions. Events are pushed to the dashboard and accumulated in badge notifications. Participants process events by selecting the appropriate economy function icon. An event forces an interaction that requires a trust decision, e.g., to commit an asynchronous transaction, respond to an email, extend a business relationship, or install a patch.

In the independent game, subjects are limited to a single means of input, namely a mouse for movement and clicking. The content is similar to what is used inside the magnet for the fMRI studies. These two factors require a minimalist interface that allows the conductors of the experiments to present the subject with a series of trust questions regarding different content samples. Content should be realistic, i.e. framed in the context of the economic scenario, and contain a mix of the relevant trust cues for the respective modality.

The simulation targets four distinct modalities for interaction:

Email – Participants receive emails from other (virtual) participants in the economy. Emails may solicit social network interactions, promote special offers, or announce system upgrades.

Social Network – Relationships in the economy's social network establish conduits for business transactions and influence reputation. Friend requests and messages are the primary events supported by this functional element. Product reviews and corporate ratings also can be found in the economy's social network.

Web – Company websites serve as the primary portal for asynchronous transactions. Company websites also may establish links to supplier and customer websites.

Download – The download function enables participants to manage and enhance the rate and quality of production by installing upgrades and patches. This function may also request information from participants in the execution of installation procedures.

In the simulation, each interaction modality encumbers the subject with a collection of hazards and corresponding set of contextual cues. These broadly represent what confront users and operators of real world systems. Some decision errors may be associated strongly with missed cues that are attentive, while others may stem from social cognitive trust processes. A confluence of factors as determinants for trust decision errors is also likely in many circumstances. For example, in email systems, the decision to open an email attachment is largely influenced by the trustworthiness of the attributed origin. Yet it is also the case that attention paid to the attachment filename can be a factor in a trust decision.

2.4 RESULTS

This section describes the results of the project in each of the three principal domains. The results set the stage for data analysis towards deriving a biomarker for Cyber Trust, and conducting additional Cyber Trust studies to evaluate the efficacy of targeted interventions.

2.4.1 PERCEPTUAL CYBER TRUST TAXONOMY

The Perceptual Cyber Trust Taxonomy [63] focuses on the human-observable elements of phishing attacks. By categorizing perceptual trust elements within a phishing message, we can begin to study the visual tricks used by attackers to deceive end-users. Thus, we approach this work from the perspective of how attackers exploit end-user trust propensity. In general, trust decisions users make online are based on three distinct hazard cue categories: *Content*, *Context*, and *Contract*; (i) Trust decisions based on visual ***content*** presented to the subject, (ii) Trust decisions based on the ***context*** in which content elements are presented to the subject, and (iii) An implied trust ***contract*** is made between the victim and the attacker when a victim is asked to supply sensitive information or to perform a risky action by the attacker.

These hazard cue categories provide a framework with which we can further analyze and classify phishing artifacts across the three aforementioned online modalities. A discussion of the three categories of cues follows. Figure 25 presents an overview of the perceptual Cyber Trust taxonomy.

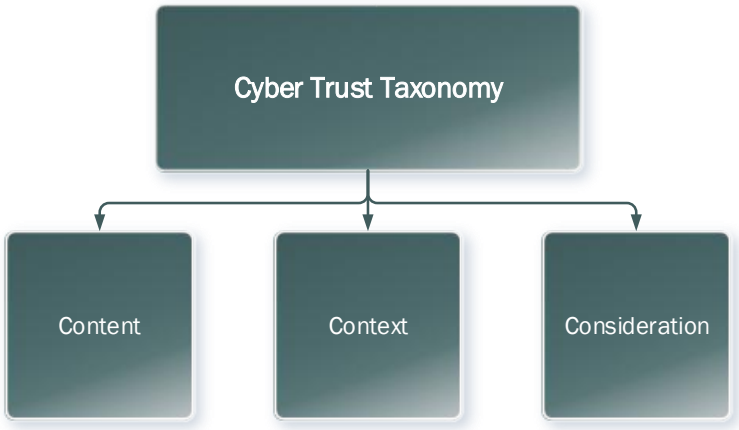


Figure 25: Cyber Trust Taxonomy Hierarchy

2.4.1.1 Content

Visual artifacts that comprise the phishing content with which users are confronted are the front line tools that attackers employ to manipulate user trust. Many users are trained to look for visual security cues, such as HTTPS or padlocks, within a web browser to verify the presence of "security" online. Unfortunately, attackers leverage such reliance on the presence of expected visual indicators, and exploit this to their advantage by embedding such elements in web pages or pop-ups using visual deception. Attackers have also been known to take advantage of the user's lack of knowledge about the syntactical structures of URL domain names, filenames in email attachments, and the reliance on reputable names or logos. Figure 26 illustrates the *Content* branch of our perceptual taxonomy.

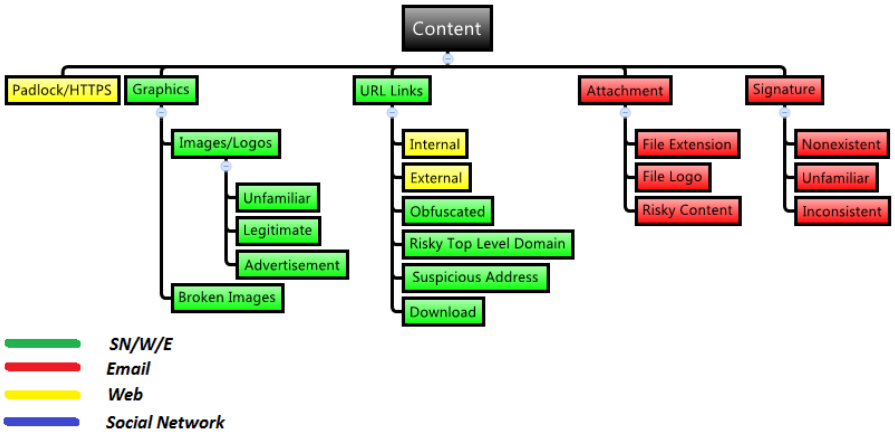


Figure 26: Content Artifacts of Trust

Padlock/HTTPS: These are binary elements within a web browser that indicate two things—a secure connection in the form of an encrypted connection between the user's browser and a web server, and the endorsement of the associated session key on behalf of a certificate authority (CA) [64]. Web browsers, such as Google Chrome, indicate to the user whether or not the CA used for end-to-end encryption is trusted and reputable (Figure 27).

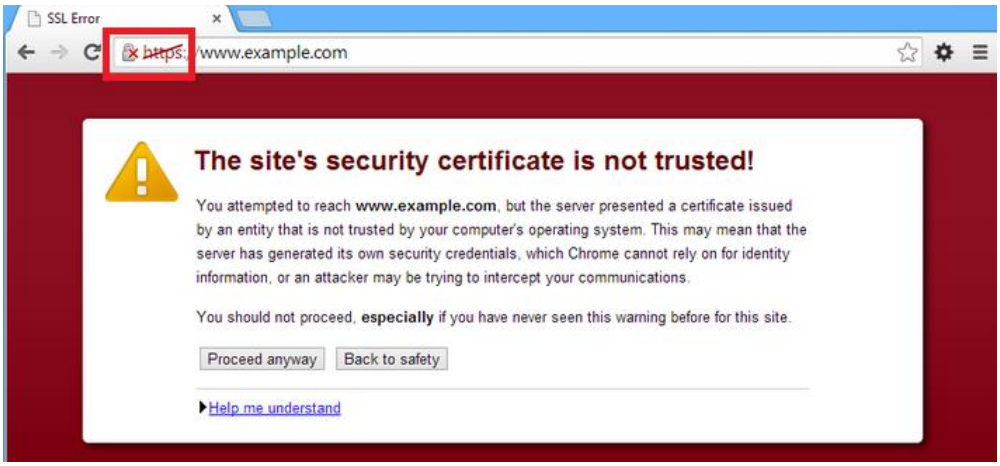


Figure 27: Google Chrome Alerting the User of an Untrusted SSL Certificate

Unfortunately, there is no universally agreed upon method for how these visual indicators should be presented to the user. They are consequently presented in various ways depending on browser vendor and version [65][66]. Attackers can tailor visual padlock deceptions to mimic those of a user's browser [67][64]. Moreover, the variability in cue forms and presentations add to the general confusion of users.

Graphics: The presence or absence of expected images online can influence user trust propensity. Subtle discrepancies in expected logos and images on web pages can have direct influence on user trust propensity. Alternatively, the use of legitimate logos within a phishing site can increase user trust propensity. The perceived production quality or lack thereof in logos and images can also represent cues to users as to the trustworthiness of online content.

URL Links: Universal Resource Locators (URLs) have syntactical cues that can be evaluated. The ability to inspect a URL and readily recognize whether it links to a trusted web site is vital, as URL misdirection is a common tactic for steering users to compromised web pages [68]. URLs that have been obfuscated to hide their real target page are also used to mislead naive users [69]. Legitimate URL aliases, such as TinyURL may confound cue recognition. Evaluation of top level domains (TLDs) at the end of the URL should be carried out before clicking on a link. Domains ending with a risky TLD name such as .RU, .CH, and .KP should be evaluated as extremely risky by the end-user [68]. Finally, a URL can be inspected to gather a general idea as to its action if the user clicks on it. The URL could be a link to immediately download a file. In this case, the file type provides an additional trust/hazard cue.

Attachments: The logos and file extensions of email attachments can be evaluated to identify hazard cues. Any executable file or multi-media file in the form of an email attachment that is processed within a local client application should be considered as risky content. Visual trickery with the way certain email clients render file names can be used to deceive users into thinking they are opening up a PDF file, when in actuality the file is an executable (.exe) file. The file attachment logo can also be used as an observable cue for file type identification.

Email Signature: The mere presence or absence of an email signature block can impact whether recipients trust an e-mail. Subtle discrepancies between signature blocks and other information, such as e-mail header content, can also cause a user to be wary of the legitimacy of an email message.

The presence or absence of visual content trust cues online can provide some level of reassurance to the user about the trustworthiness of an online engagement. These binary indicators

by themselves do not necessarily guarantee a secure state in which an application can be completely trusted, but rather a level of trust that when used within the context of other elements can provide more insight into the trustworthiness of an online engagement.

2.4.1.2 Context

The context in which phishing artifacts are presented can affect user trust propensity as well. The presentation of visual content can influence the focus of users based on designer end goals. Context combines the visual presentations of content with user expectations arising from environmental and historical factors. The origin of an e-mail, URL, or follower can have significant trust implications. The tone of an e-mail or social media message can also affect the user's willingness to trust. This is commonly dependent upon the user's relationship to the other party in question. Grammatical, syntactical, and idiomatic characteristics of sentences, words, and phrases across all three online modalities may affect user trust. Figure 28 presents the *Context* branch of the Cyber Trust taxonomy.

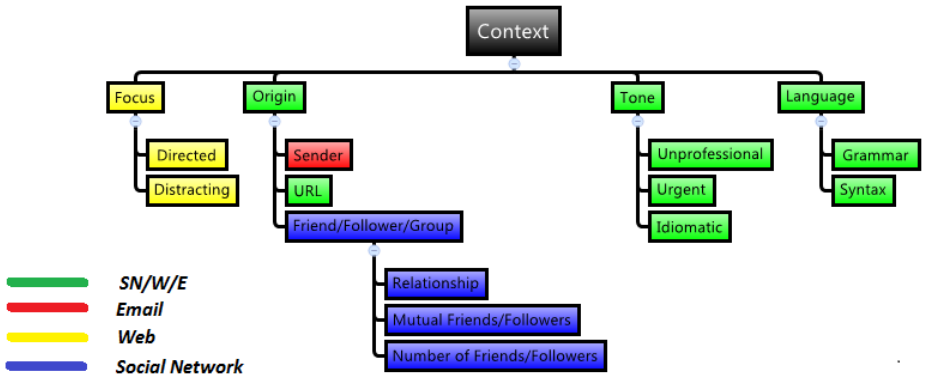


Figure 28: Context Artifacts of Trust

Focus: Positioning of pictures and text on a web page can impact user focus and consequently influence trust propensity. A user trying to accomplish a given task (e.g., downloading a file from a download page) can become easily distracted by web pages containing multiple images and links that read "download" and are cluttered with advertisements. Other web sites may have a more intuitive flow allowing a user to be directed to easily accomplish the task at hand with few, if any, distractions.

Origin: The domain name of the sender address can be used with other trust elements to deduce the trustworthiness of an e-mail message [70]. The same is true of the TLD of a URL. Information related to the relationship to the user, the number of friends/followers, and mutual friends/followers in a social network can be used to derive the legitimacy of an individual or organization via social media networks.

Tone: Tone in the contextual sense refers to the way in which users are addressed via online communication. For example, one might expect a business associate to begin an e-mail with a more formal greeting than a close friend. Conversely, one might not expect a friend to exhibit formality throughout a message. A sense of urgency in a message is often employed by attackers to compel naive users to supply sensitive data or carry out a risky action, such as opening a malicious attachment.

Language: Users expect senders and recipients to exhibit some level of predictability with regard to grammar, vocabulary, and sentence structure. Deviations from anticipated language usage may influence user trust propensity in either a positive or negative way.

2.4.1.3 Contract

This portion of the taxonomy is concerned with the value proposition of an online trust decision. A *Contract* is composed of an offer and consideration. Consideration captures what a user gives in exchange for what they are getting in return (offer). Users are commonly asked to part with sensitive information or execute some action that may place them in some heightened state of vulnerability. The offer ostensibly confers the benefit of complying with the consideration component of the online contract. Figure 29 presents this branch of the taxonomy.

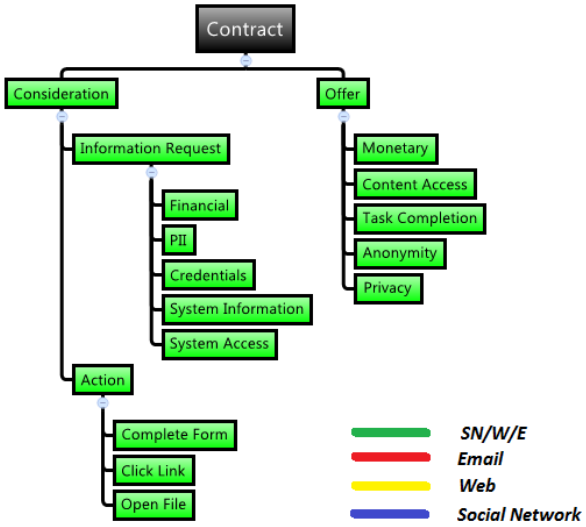


Figure 29: Contract Artifacts of Trust

Consideration: Attackers entice users to provide Personal Identifiable Information (PII), credentials, system access, or general system information in order to further their attacks. Alternatively, an attacker will try to persuade the victim into performing a risky action that can lead to disclosure of sensitive information or system compromise. System compromise can occur when a user is enticed to navigate to a compromised web site, or when a user opens a malicious file.

Offer: This branch is focused on classifying what the attacker is advertising to provide the user in exchange for information. Attackers may entice users with large sums of money, sensitive information, or access to subscription-based content or services.

2.4.1.4 Cyber Trust Taxonomy Applications to Research, Tool Development, and Training

Our perceptual taxonomy of Cyber Trust cues embodies a framework within which to pose new research questions, develop novel tools, and pursue differentiated training for individuals and organizations. Common to these activities is the use of the taxonomy to classify, sort, and otherwise analyze phishing attacks by its constituent and composed elements.

Classification

With respect to research in this field, the taxonomy helps classify phishing attacks pulled from the wild. Against the backdrop of a stable collection of phishing attack categories, several research questions can be posed. *Which attacks are the most effective, and against what targets? What kinds of attack are amenable to mitigation by technical controls? Which are most effectively dealt with by targeted training?*

Classification may be conducted on an *existential* basis, in which artifacts are inspected for the inclusion of specific elements of the taxonomy. Investigators may extend this approach by establishing *types* or values for individual elements for more meaningful comparisons and matching. *Combinational* classification schemes can also be used that, for instance, group attacks sharing an urgent tone with one that asks a user to provide credentials in exchange for the promise of successful completion of a task.

Classifying individual phishing artifacts onto the cyber trust taxonomy provides a basis for clustering similar artifacts. Taxonomical elements in phishing artifacts permit "fingerprinting" – capturing the distinctive aspects of an attack. Clustering artifacts with similar or identical fingerprints may offer some insights as to the origin or development process of an attack.

Classification can also be used to profile the phishing attacks observed by an individual or organization. This yields an added level of intelligence concerning the threat posture of an enterprise that can be used to optimize risk-reducing activities. Such classifications can help a security team develop tailored responses and interventions that target critical facets of the characteristics common to observed attacks. For example, based on the prevalence of phishing attacks targeting PII, procedural or operational controls may be deployed to gather PII for legitimate purposes using a secure/specialized channel, obviating the need to respond to emails requesting PII.

Developing sound and useful classification techniques and schemes for phishing must be a research priority to advance the core science and understanding of the topic. Any such agenda should leverage existing phishing artifact repositories, classifying and analyzing the classified specimens therein. Analysis must be conducted against a core set of measurable attributes or effects of phishing attacks to address the major research questions articulated above. When it comes to understanding the human factors involved, this represents an opportunity to define experiments designed to study the influence of psychology and neurobiology on cyber Trust as directed against distinct taxonomical elements.

Tools

Our taxonomy establishes a blueprint for the development of programmatic tools to facilitate targeted interventions for end-user trust decisions. Assisting in the development of such tools is an XML document type definition (DTD) based on the taxonomy. Figure 30 provides an example showing the content attributes of phishing email that contains an external URL link to a suspicious site embedded within the message. This example also includes the attachment of a risky PDF document.

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE CYBER_TRUST_TAXONOMY SYSTEM "CyberTrust.dtd">
- <CYBER_TRUST_TAXONOMY>
  - <CONTENT>
      <PADLOCK>Yes</PADLOCK>
    - <GRAPHICS>
      - <IMAGES_LOGOS>
          <UNFAMILIAR>No</UNFAMILIAR>
          <LEGITIMATE>Yes</LEGITIMATE>
          <ADVERTISEMENT>Yes</ADVERTISEMENT>
        </IMAGES_LOGOS>
        <BROKEN_IMAGES>No</BROKEN_IMAGES>
      </GRAPHICS>
    - <URL_LINKS>
        <INTERNAL>No</INTERNAL>
        <EXTERNAL>Yes</EXTERNAL>
        <OBFUSCATED>No</OBFUSCATED>
        <RISKY_TLD>Yes</RISKY_TLD>
        <SUSPICIOUS_ADDRESS>Yes</SUSPICIOUS_ADDRESS>
        <DOWNLOAD>Yes</DOWNLOAD>
      </URL_LINKS>
    - <ATTACHMENT>
        <FILE_EXTENSION>Pdf</FILE_EXTENSION>
        <FILE_LOGO>Pdf</FILE_LOGO>
        <RISKY_CONTENT>Yes</RISKY_CONTENT>
      </ATTACHMENT>
  </CONTENT>
  + <CONTEXT>
  + <CONTRACT>
</CYBER_TRUST_TAXONOMY>
```

Figure 30: XML Describing Phishing Email

One such example is a tool to automatically generate and render phishing content across online modalities. This tool produces static content in the form of HTML files, given the quantity of stimuli to produce for each modality (Figure 31). The tool relies on cascading style sheets (.CSS) that define the layout of individual content artifacts that define a web page, e-mail, or social network page (Figure 31). This tool also relies on the presence of a pool of textual messages and images with which to generate and degrade content. These artifacts are stored in a local SQLite3 database. The Cyber Trust content generator produces content with random levels of artifact degradation based on our taxonomy.

Future plans for tool development using the taxonomy should complement the foundational research agenda. Of initial consideration are tools for measuring attributes of phishing (and legitimate) specimens, thus the definition of meaningful metrics becomes an immediate objective. Tools that permit measurement and exploration of such content will facilitate research, and ultimately translate to superior detective controls and technical interventions within browsers, applications, and operating systems.

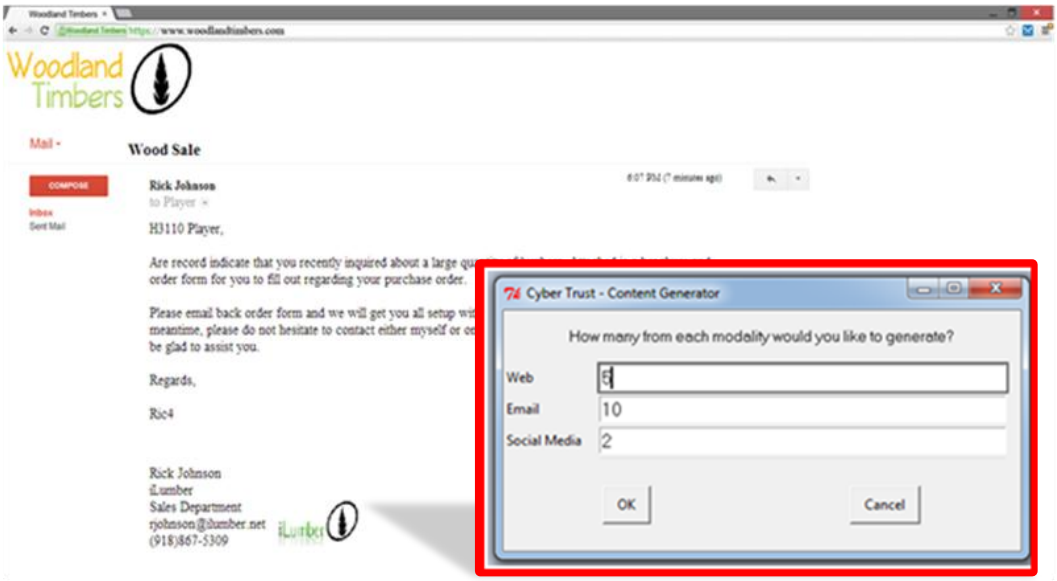


Figure 31: Email Stimulus Produced by Content Generator

Targeted Training

In addition to driving research and promoting tool construction, this taxonomy is useful for developing differentiated training program content designed to strengthen end-user cybersecurity competencies. The goal is to enable users to detect deceitful messages, dodge their deleterious effects, and disseminate information about deceit to stakeholders. How is the goal disaggregated into its trainable facets? The first step involves conducting a learning needs assessment. Training programs are designed to achieve goals that meet instructional needs. It is dangerous and costly to begin any program without a complete assessment of the task, behaviors, and environment [71]. Understanding workforce capabilities is a critical part of identifying areas requiring change. A learning needs assessment involves asking questions that reveal the competencies and development needs of end-users. The taxonomy provides a domain from which to draw these questions. One-size-fits-all training is ill-advised, given the broad continuum of end-user skill-levels and dispositions. Knowledge of end-user strengths and development needs allows organizations to implement differentiated training strategies designed to equip end-users with the requisite know-how to evade cyber-deception. It is a vital step toward creating a frontline defense for otherwise vulnerable organizations.

The taxonomy not only guides the development of end-user learning content, but also provides a blueprint for training program evaluation. The taxonomy establishes criteria for training success. It provides organizations with an organized means of determining whether employees have overcome the learning deficiencies identified by the needs assessment.

To illustrate a training application, consider the taxonomical node URL Links - Obfuscated (Figure 26). To address an organizational concern over phishing, a company administers a taxonomically-oriented test to its employees. The results indicate that employees struggle distinguishing trustworthy URLs from obfuscated URLs. They are vulnerable to typejacking domain name attacks. This type of deception involves surreptitiously altering a legitimate domain name (e.g., www.paypal.com vs. www.paypa1.com).

Having determined that a company is vulnerable to URL obfuscation, the next step is to establish a remedial training objective and module addressing this type of trick, and the adverse consequences of overlooking it. This training module is then applied to those who have been identified as deficient in URL obfuscation detection. Applying a differentiated training strategy is important, because there is no need to train employees who are already competent in detecting URL obfuscation.

To evaluate the effectiveness of URL obfuscation training, one can use a parallel form of the employee test administered in the needs assessment. By sending out periodic decoy messages with obfuscated URLs, an organization may determine whether employees are applying what they have learned. In addition, an organization can evaluate the return on investment by examining the reduction in costs associated with URL obfuscation detection failure.

Beyond the taxonomy itself, it is vital to assess whether the learning process has inspired users to continue their learning. Did that which was learned transfer to the jobs they perform? What observable effects resulted from training (e.g., fewer cyber-attack breaches, fewer cyber-related monetary and customer losses)? Evaluative information from a well-planned training program provides valuable feedback for continuous improvement of training program content, methods, outcomes, and results.

The key to unleashing the potential of the taxonomy in this application domain is building instructional modules tuned to training against specific cues and lures identified within it. As a conjunct with previously stated research objectives, such training modules can be evaluated within the context of human subject studies to evaluate their efficacy in targeted applications. This must be done in coordination with an identified learning model and framework that offers students and users the best opportunity to absorb and retain the content.

2.4.2 CYBER TRUST STUDIES

The principal effort to develop a standard methodology for deriving biomarkers for targeted interventions on Cyber Trust involved the design and execution of a fMRI study. In addition, an eye tracking study was conducted to compare approaches to evaluating the trustworthiness of online content between expert and novice users.

2.4.2.1 fMRI Study

A neuroimage study was designed and executed in an attempt to understand the role of specific neural networks in Cyber Trust. Fifty-two subjects participated in the study at the Laureate Institute for Brain Research (LIBR), which consisted of two visits. The first visit included a neuropsychological assessment followed by a localizer fMRI.

Neuropsychological Assessment

- ANAM (Automated Neuropsychological Assessment Metrics)
- CANTAB (Cambridge Gambling Test)
- WASI (Weschler Abbreviated Scales of Intelligence)
- Furman Test (cyber-security structured interview)

- The Baron-Cohen face version task

Following informed consent, a neuropsychological assessment was conducted on the participant's first LIBR visit. Baseline cognitive behavioral assessment consisted of a computerized neurocognitive test entitled Automated Neuropsychological Assessment Metrics (ANAM); [72]. This assessment battery includes verbal and visual memory tests, simple reaction time measures, processing speed, impulsivity measures, visual spatial processing, sleepiness scale, and code substitution learning. Performance on the full Cambridge Gambling task is assessed using the computerized neuropsychological test automated battery (CANTAB) version of this task. Importantly, the neurocognitive testing battery is designed for serial testing sessions and provides randomized versions of each test for each session. To assess the general intelligence of participants we administered the Weschler Abbreviated Scales of Intelligence (WASI). To quantify participants' computer security knowledge and experience we used portions of the cyber-security structured interview described in [73]. The Baron-Cohen face version of the Theory of Mind task [62] was used to assess social cognition.

Localizer fMRI

To facilitate analysis, three localizer tasks were conducted for each participant. These are used to isolate the effect of Cyber Trust Game decisions on the three neural networks of interest: Risk and Reward, Interpersonal Trust, and Theory of Mind. Within each task are multiple statistical activation maps of general interest. For this study we define specific task contrasts to assess the components of the localizer tasks that will be fit into the multivariate decomposition of Cyber Trust Game decisions.

In the Interpersonal Trust localizer, the a priori contrast of means difference that we will use to define interpersonal trust will be comparing "trust" trials (collapsing across defect and reciprocate trials) versus "non-trust" trials. This activation map will control for stimulus features and should differ only on the decision of interpersonal trust. In the Reward localizer task, results from Roger's et al. [74] suggest that mapping reward circuitry can be accomplished by a priori contrast of means comparing High and Low Gain trials (collapsing across the "likelihood" factor). Finally, for the Theory of Mind localizer task, we will use the a priori contrast of means comparing "Mental State" decisions versus "Gender" decisions. This contrast will, as in each of the previous tasks, control for stimulus features and response characteristics, leaving the participants' attribution of mental state activation map.

All scanning was conducted on a 3T General Electric Discovery-series MR750 scanner using a custom-built Nova Systems 32-channel arrayed head coil. Resting state scans and Diffusion Tensor Imaging was also conducted as part of the first scan session. All anatomical and functional scans utilize parallel imaging techniques with a SENSE factor of 2. The first scan session consisted of:

- fMRI resting state scan
- Mprage (T1-weighted images High Resolution 3D)
- Interpersonal Trust Localizer Task: Trust Game (two runs)
- Reward Localizer Task: Adapted version of A Gambling task (two runs)
- Theory of Mind Localizer Task: Adapted version of the Baron-Cohen Theory of Mind task (two task)

- Diffusion Tensor Imaging (DTI 30 directions)

Of the fifty-two subjects in visit one, forty-eight yielded correct acquisitions for the adapted version of the Baron-Cohen Theory of Mind task. Forty-seven yielded correct acquisitions for the Trust Game, and forty-seven yielded correct acquisitions for the Reward Localizer Task.

The second visit incorporated three primary components: 1) psychiatric ratings, 2) the Cyber Trust Game task fMRI scan, and 3) blood draw.

Psychiatric Ratings

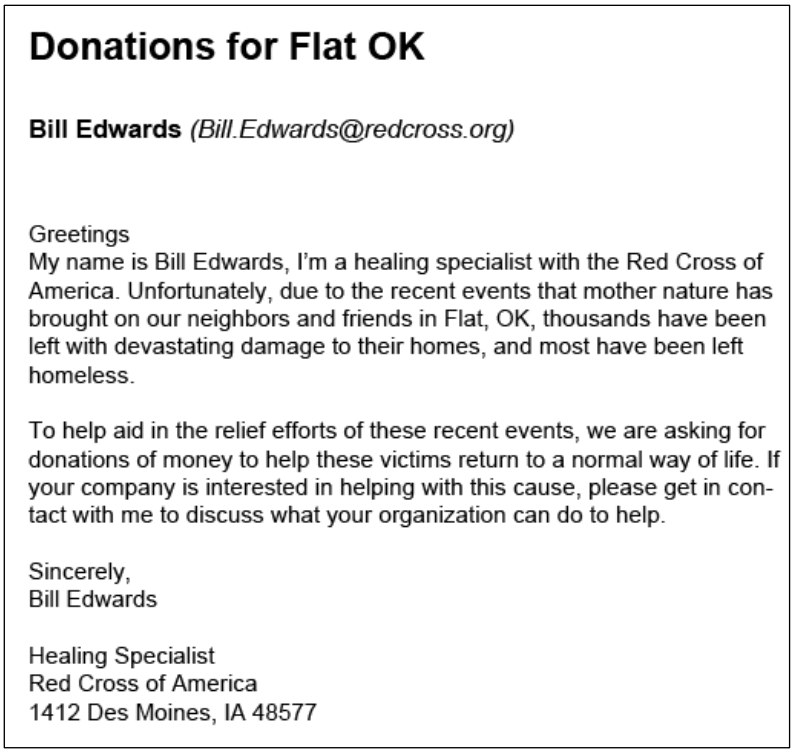
To assess current psychiatric states prior to the fMRI in the second visit, participants had several psychiatric ratings, including the Snaith-Hamilton Anhedonia/Pleasure Scale (SHAPS), the Anhedonia rating scale, and the Hamilton Depression & Anxiety Inventories (HAM-D; HAM-A; respectively) which are 21 item multi-choice questionnaires designed to assess the current level of depression and anxiety. Finally, to assess personality traits, participants were administered a Myers Briggs Test that assesses and rates five factors of personality. Two of these factors, agreeableness and extroversion, have been associated with greater risk taking [75]. Hence, the psychiatric ratings were compiled from:

- HAM-A (Hamilton Anxiety Rating Scale)
- HAM-D (17 item score, 21 item score, 25 item score, 29 item score)
- STAI (State Trait Anxiety Inventory) – (c-1, c-2)
- SHAPS (Snaith-Hamilton Anhedonia/Pleasure Scale)

Cyber Trust Game fMRI Task

Following the reward, interpersonal trust, and theory of mind localizer scanning session in visit one, participants returned for another fMRI scanning session (visit two). In this second scanning session, participants were scanned while performing a scanner-friendly version of our Cyber Trust Game task. Participants were all naïve to the Cyber Trust Game task, beyond an instructional period immediately prior to scanning.

In the scanner version of Cyber Trust Game, participants were presented with a pre-determined order and number of simplified versions of simulated emails, web pages, and social network messages. Figure 32 presents a simplified email artifact. For the initial fMRI analysis, each of these cyber modalities can be combined in to one "grand" modality. The stimuli from each modality contains either a high or low number of risk cues, and the participants were prompted to respond to either "accept" or "reject" the email, web site, social network, or download request. As depicted in Figure 32, this leaves us with four trial types: Low-Risk: Trusted, Low-Risk: Not-Trusted, High-Risk: Trusted, High-Risk: Not-Trusted.



Donations for Flat OK

Bill Edwards *(Bill.Edwards@redcross.org)*

Greetings
My name is Bill Edwards, I'm a healing specialist with the Red Cross of America. Unfortunately, due to the recent events that mother nature has brought on our neighbors and friends in Flat, OK, thousands have been left with devastating damage to their homes, and most have been left homeless.

To help aid in the relief efforts of these recent events, we are asking for donations of money to help these victims return to a normal way of life. If your company is interested in helping with this cause, please get in contact with me to discuss what your organization can do to help.

Sincerely,
Bill Edwards

Healing Specialist
Red Cross of America
1412 Des Moines, IA 48577

Figure 32: Simplified Simulated Email

The neural networks for each of these four conditions can then be defined and contrasted for developing the Cyber Trust neural network deliverable, and each of the four networks will be decomposed into similarity to reward, interpersonal trust, and theory of mind using a machine-learning algorithm. As shown at the bottom of Figure 33, this results in a 12-dimensional parameter space for each subject that will be submitted to an unsupervised multivariate classifier to classify participants based on their Cyber Trust decision decomposition. The resulting classifier will be the neuroimaging biomarker for Cyber Trust decision propensity once the analytical phase resumes.

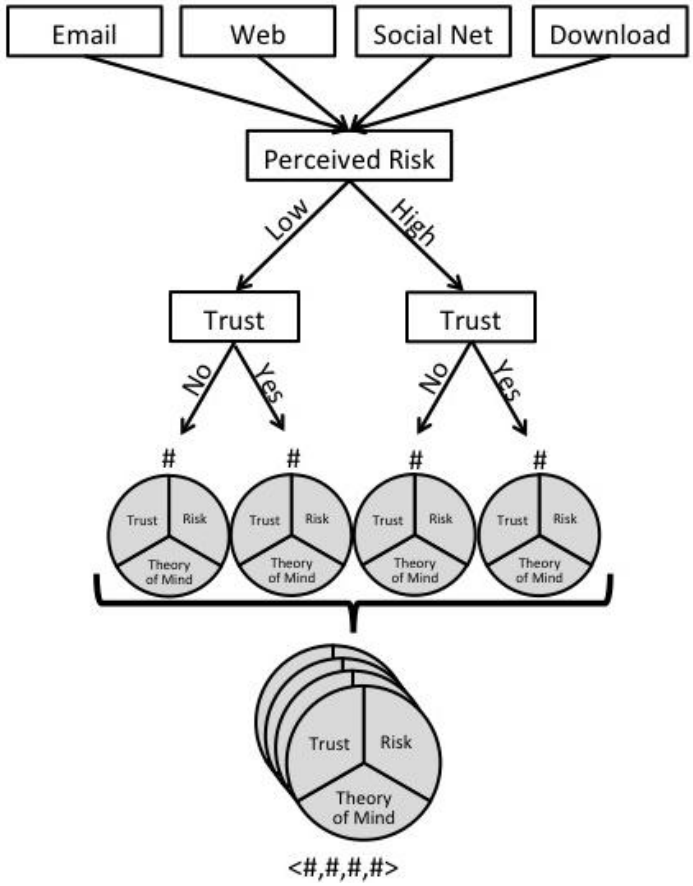


Figure 33: Simple Form Cyber Trust Game and Post Processing Method

- fMRI resting state scan
- Mprage (T1-weigthed images High Resolution 3D)
- Cyber Trust task (6 runs)
- Diffusion Tensor Imaging (DTI 30 directions)

Blood draw

Blood was drawn at the Laureate Institute for Brain Research during visit two. Four vials, approximately 80ml, of blood were drawn and processed at the Saint Francis Hospital laboratory. Blood was tested for levels of glucose, insulin, oxytocin, vasopressin, estrogen, testosterone, cortisol, and C-reactive protein. Glucose and insulin levels can affect attention and processing speed [76][77][78]. C-reactive protein was analyzed for levels of inflammation. High levels of inflammation have been associated with mood dysregulation [79][80] that may affect trust behavior. Finally, oxytocin, vasopressin, estrogen, and testosterone have all been implicated in mediating trust behavior.

A total of fifty-one subjects participated in the second visit of the neuroimage study. There were forty-four correct acquisitions for the Cyber Trust task. As funding to complete the project is

restored, the analysis of the data from the phase 1 fMRI study will resume, and Phase 2 of the fMRI study will commence as planned.

2.4.2.2 Eye Tracking Study

As a companion study, we conducted an experiment comparing expert and novice approaches to determining whether various cyber communications were trustworthy. Ninety-four participants completed the eye tracker experiment making trust decisions on the cyber communications. (There was no overlap in the eye tracker and fMRI study participants.) Participants were asked to evaluate the trustworthiness of various messages while having their eye movements recorded by an eye tracking system. Following the simulation, participants completed a short survey about their experience with computers and their personality.

This resulted in data on individual differences that influenced the accuracy of trust decisions and eye tracking data on the most effective scan patterns and decisions processes for efficient and accurate decision making. Initial data has revealed substantial differences in the eye tracking patterns based on the mindfulness of users. This work will be presented at an upcoming conference [81]. This data is still being analyzed in more detail for journal publications and wider distribution.

2.4.3 EDUCATION, TRAINING AND RELATED MATERIALS

This component of the project primarily involved the development of a simulation to serve as a basis for Cyber Trust Game experiments, as well as for training. In addition, an Internet Security and Phishing presentation slide deck was developed as an intervention strategy to support future Cyber Trust studies. Ancillary surveys and presentations were also developed as tools to assess and raise awareness of Cyber Trust issues.

2.4.3.1 CyberPhishing Game

A CyberPhishing game was developed on a robust simulation platform to assess the ability of individuals to make Cyber Trust decision in the presence of various cues and lures. It also serves as a platform for Cyber Trust training and education. Figure 34 shows a high level architecture of the game as a simulated web environment. On the bottom appear a number of simulation modules included to simulate real world situations in business. The orange modules simulate real world media for delivering suspicious content to the user. The simulation platform provides the capability to simulate emails, tweets, and generic web sites. The middle modules (in green) log user interactions, such as market transactions or Web interactions.

Additional modules (shown in blue), including Market Share, In the News, Stock Ticker, and Performance Feedback, facilitate meta-level user immersion by providing the user with various types of meta-story content. Specifically, the Market Share module provides users with some notion of how their company is performing in its overall market sector, the In the News module simulates news stories that relate to the company and the user's actions, the Stock Ticker denotes a numerical stock price of the company and its competitors, and Performance Feedback acts as a user recommendation engine to train users to recognize suspicious content when they make poor decisions. For suspicion level assessment, only actions affect the blue modules. For training, after the suspicion level baseline is determined, the Adaptive UI Component (blue rectangle) will draw upon information from the blue modules so that trust decisions made in the game affect the meta-story.

The components on the top of Figure 34 show how the game is organized. The User Interface provides the modules for interaction. The Adaptive UI Component is used to change the modules relative to whether the game is being used to baseline the user's suspicion levels or to provide

training based on the performance feedback module and the company's success given the user's decisions. The Economy Engine captures the user interaction in the form of temporally associated events. These events include what decision was made, what other module information was observed prior to making a decision, and how the decision affects the overall meta-story, i.e. the company's performance. The Data provides usage information, such as how long a decision takes, and will incorporate data collected from an event tracker during game play.

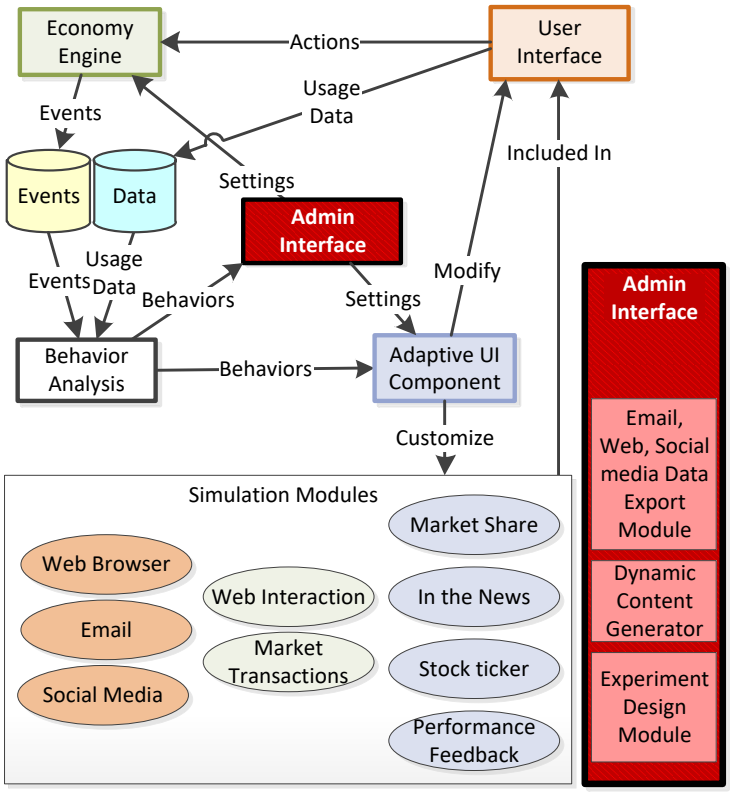


Figure 34: General Game Architecture

The Behavior Analysis component determines a user's suspicion levels by providing automated event and data analysis throughout game play and at the end of the game. This includes

- Sensitivity to certain lures and cues,
- Depth of research and discovery prior to decision making,
- Differences in suspicion levels among the three communication media, and
- Comparisons across users, including demographics.

Sensitivity to lures and cues is determined by how the content is classified within the game, which is discussed in terms of pristineness in the next section. All events that result in a state change to the system are captured. Thus, the depth of research and discovery prior to decision making will be understood given the number and type of events preceding a trust or don't trust decision on content, along with the amount of time the user takes to make the decision. Users may trust or distrust one form of content over another. Such suspicion may stem from earlier training, such as an

email phishing course. To determine if a user is more susceptible to attacks in a particular medium, the content classifications will be normalized across the three communication modalities. Thus, the number of content items in email that are classified at one degradation level will be equal to the number of content items presented to the user in the social network and in the browser, even if their degradations are different. As users complete the game, the system can compare their performance to other users.

Building the Story for Game Immersion

Phishing tactics rely on social engineering and often attempt to convey urgency (e.g., download updates), appeal to a person's sense of greed (e.g., survey coupons), and appeal to a person's sense of compassion (e.g., relief agency requests). The unique story provides the setting for these tactics to be a realistic part of the game, called "CyberPhishing." The user takes on the role of interim president of a company called TimberTrust, a construction and remodeling company. The story provides a backdrop and basis for phishing tactics making them a realistic part of the game. The current president, Gregor Only, is being investigated for fraud but is unable to be questioned, because his boat was captured by pirates while sailing near Cape Horn.

This part of the scenario allows the introduction of many characters to the game, such as those who have spotted Gregor Only, want ransom, know something about the fraud investigation, or are hiding information from investigators. Another major plot point is that a massive tornado has hit nearby and TimberTrust is called into action to help out. Additional characters include those who need help, are part of the insurance industry, want to work with TimberTrust's restoration effort, and want to compete with TimberTrust. The user is involved in daily tasks over a two week period to set the company back on the right track. Performance is measured by stock price increases, building up a legitimate network of partners, media praise, and bonuses. Figure 35 shows some sample tasks the user encounters on Day 1. Each task requires a user to examine a content item and make a decision to trust it or not.

Devising Realistic Content

One critical feature needed for any simulation platform that proposes to increase user awareness and train them to detect suspicious content, such as phishing attempts, is the ability to generate and adjust the simulated content. It is necessary to reduce the manual process. In addition, the content must be specifically tailored to meet experimentation requirements. Thus, content generation processes should follow several guidelines.

1. Limit the technical knowledge needed to create content. A content creator should need minimal HTML or javascript to create simulated web pages, social media posts, or email content.

2. Allow content to be copied and instanced for each user, indicating the need for content templates that allow user attributes (such as name and email) to be plugged in.

3. Provide a preview of the content as it is being generated.

4. Have limited loading times. Web loading times should not be a factor in content generation modules. Specifically, page load time should not hamper or impede the design process.

5. Be immediately classifiable when added to the system so that the experiments can be tailored to cover varying levels of suspiciousness.

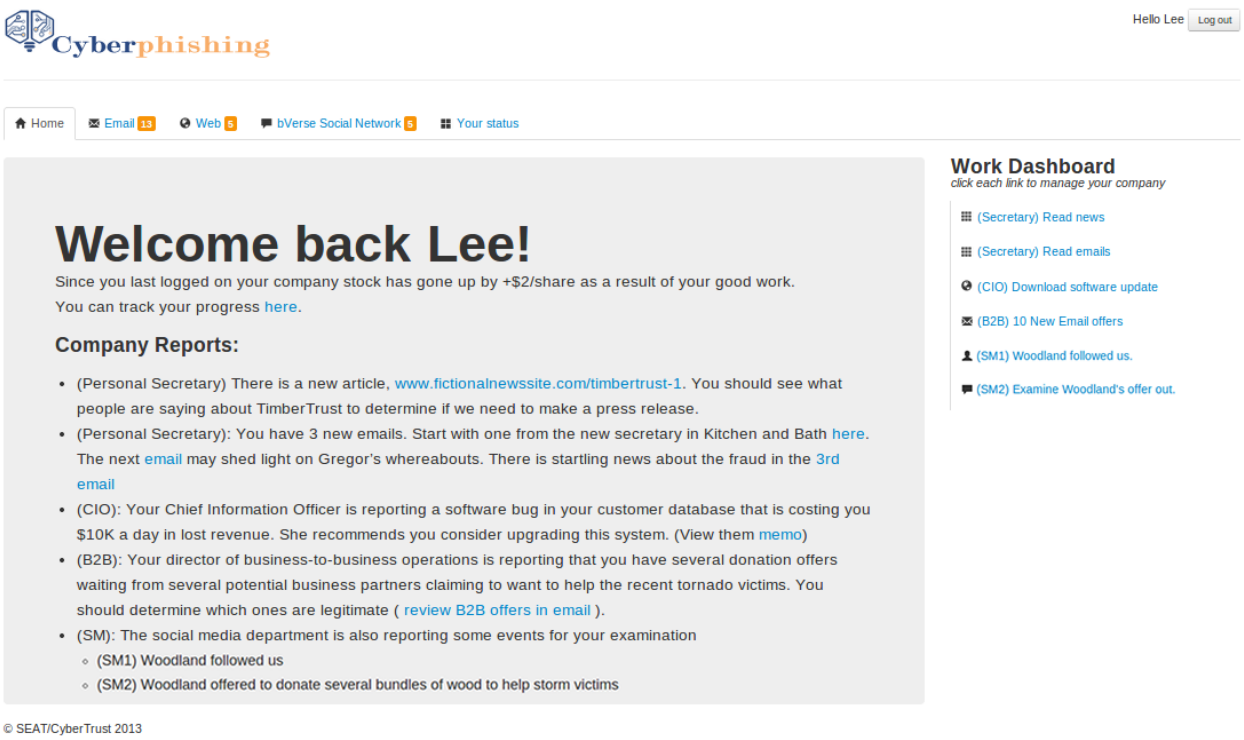


Figure 35: Sample Tasks during Day 1

The simulation and training platform that houses the game recognizes the critical importance of these factors in content generation. Our platform provides a quick, visual, and dynamic content generator where a content designer fills template fields and dynamically previews the results in real time (as shown in Figure 36). The three content generation modules, one for each modality, allow specific content details along with a pristineness or suspiciousness level. The content is instantiated for a user and inserted into visual elements that appear in the simulation user interface, without requiring the designer to interact directly with any code. The highest level of technical know-how required to create content is use of the
 html tag, to create page breaks, and <a> for identifying links. All other code is obfuscated from the designer and filled in automatically by the platform.

One unique feature of the game platform is that it is entirely built as a client-side web application. This means that page interaction is instantaneous and data CRUD (create-read-update-delete) actions do not impede page load times. In other words, a content designer can create phishing content as fast as his or her workflow allows, satisfying principle 4 above. Figure 36 shows a piece of content linked with Day 1 of an experiment that develops a scenario around a potential weather disaster. The template fields at the top of Figure 36 allow content designers to plug information in and see a dynamically rendered preview of the resulting content item (bottom), as a user would see it.

Content is designed for a template user, and instantiated for specific users when a designer selects them to be part of an experiment. While Figure 36 shows the fields that directly apply to email content items, the platform also provides an additional set of general content parameters that affect all types of content. These parameters include valuations for pristineness (discussed below) and other fields that facilitate and track user trust decisions and experiment parameters. The general content parameters that apply to the email being dynamically rendered in Figure 36 are shown in Figure 37.

Pristineness is a value assigned to the content that classifies it according to the number and type of lures and cues. We have surveyed a wide range of literature to determine the most prevalent and

divisive schemes used in phishing attacks and categorized them in association with created content for a role-play game. These are based on known degradations given the content interaction types and include:

- Suspicious URL identifying content sender
- Suspicious links and Pop-ups within content
- Multiple instances of poor and/or faulty grammar
- Use of certain greetings and catchy phrases
- Unnecessary warnings
- People posing as friends
- The use of a full name of receiver rather than a nickname
- References to products not commonly available
- Item prices too good to be true
- Missing security designators, e.g. https and the padlock
- Direct request to input personal data
- Picture-in-picture
- Survey requests with links
- Appeals to an emotion, e.g., urgency or greed

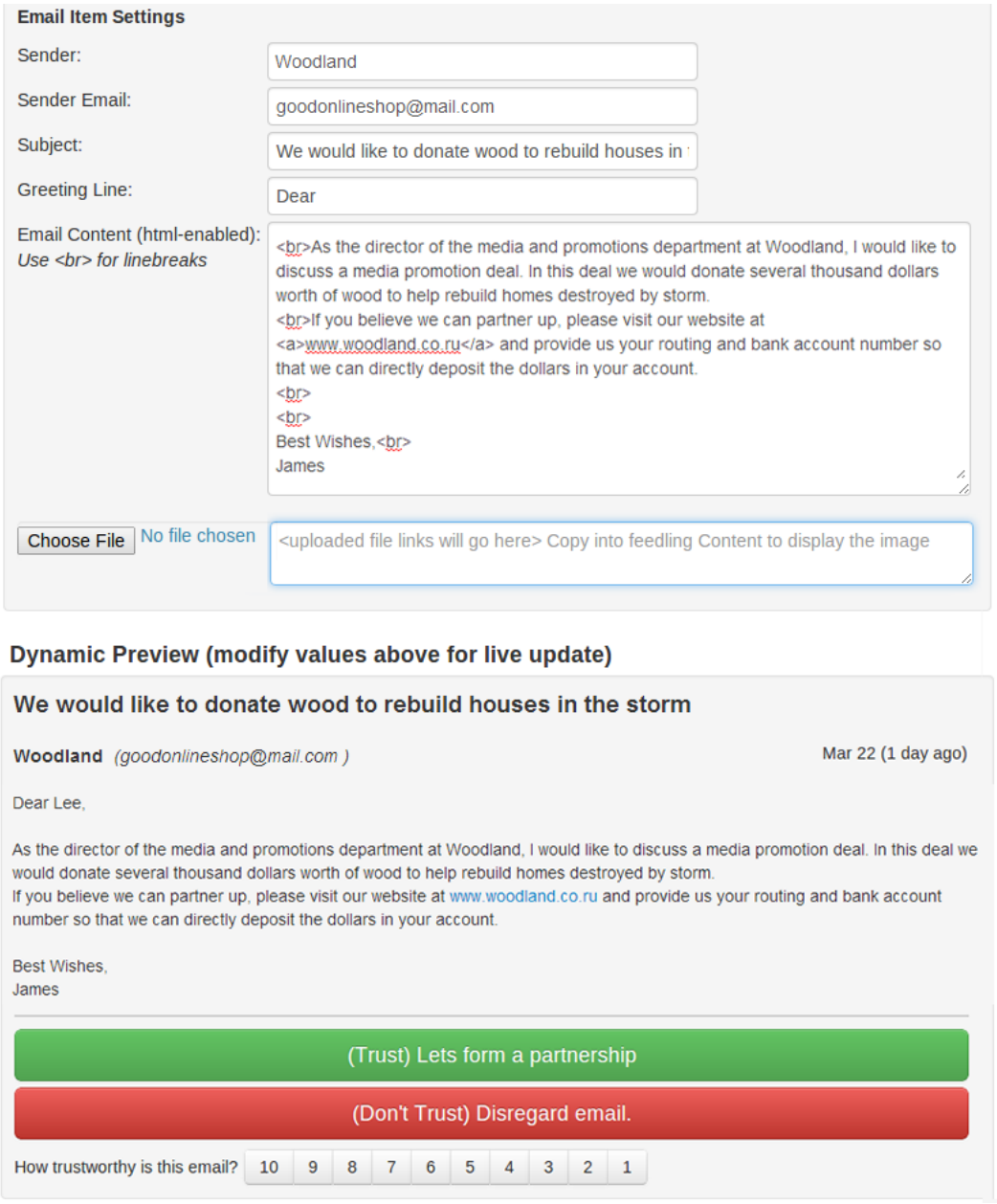


Figure 36: Content being generated (top), Actual content displayed (bottom)

If we let N be the number of identifiable lures and cues used within the game (a number that may increase as phishing content becomes more sophisticated), then the highest pristineness classification is N. Given the list above, in which N = 14, the example content in Figure 36 has pristineness = 14 – 4 = 10, as shown in Figure 37, because it has poor grammar, a suspicious link within the content, an emotional appeal, and a direct request to input personal data. Similarly, the social media post in Figure 38 has a pristineness of 9 because of poor grammar, the presence of a suspicious link within the content, use of a catchy phrase to get you to click the link, the author is not followed, and because the author is posing as a friend.

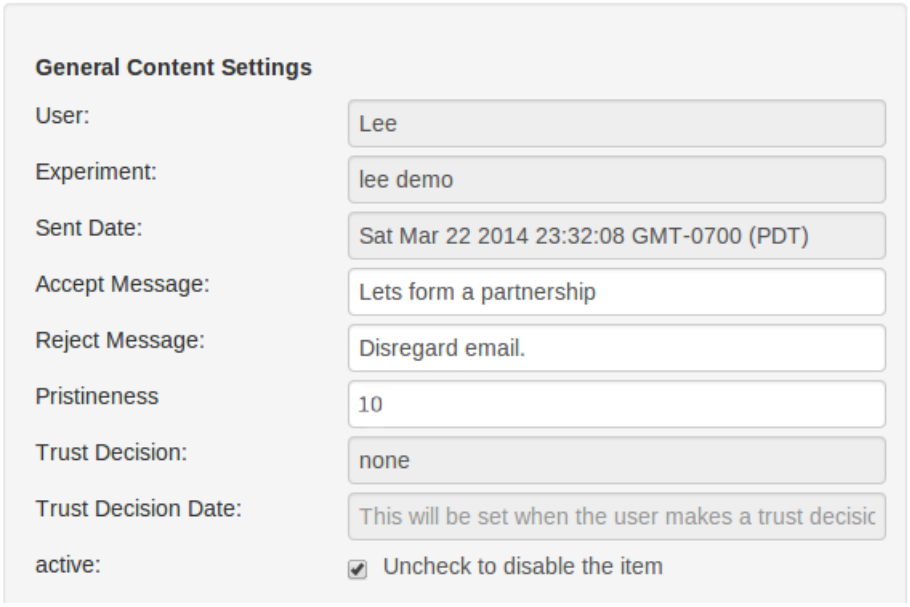


Figure 37: Content setting for the Email shown in Figure 12

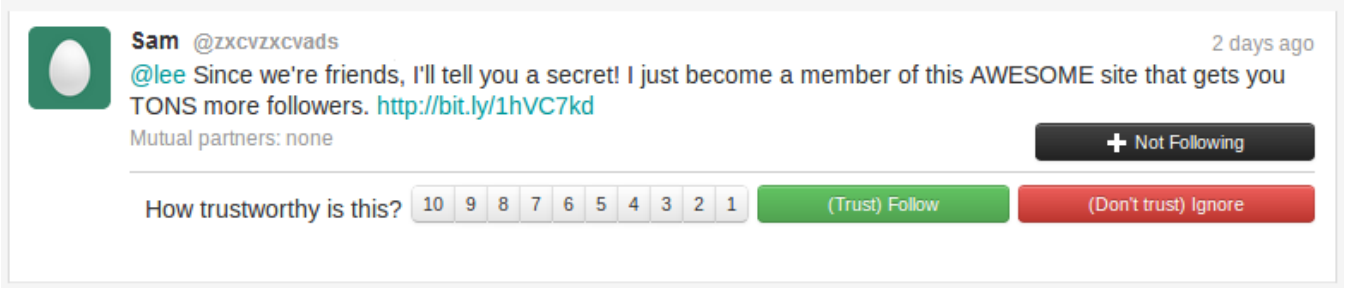


Figure 38: Untrusted Content from Social Media

Figure 39 shows an article that the user is supposed to read as part of Day 1 in the game. Notice that a simulated generic browser is used so that there is no inherent trust bias that might be associated with Internet Explorer, Chrome, Firefox, etc. The content in Figure 39 has pristineness = 1 because there are a number of scam advertisements, a browser error popup and a phishynews.co.ru web address (instead of a website that matches the content, such as metronews.com. This type of content might be seen in situations where a malicious website steals content from a legitimate article or reputable news source (here the fictitious legitimate "metro news") and uses it to scam users, collect private information, or disseminate worms or other viruses. There is also a known spam message with a suspicious link regarding a fake rollercoaster accident. Content showing such characteristics as in Figure 39 are automatically downgraded to the lowest level, where pristineness = 1.

Figure 39: Untrusted Web Browser Content

2.4.3.2 Internet Security and Phishing Awareness Presentation

In addition, an Internet Security and Phishing Awareness training presentation was designed to establish a baseline for internet security and phishing awareness for cyber trust study participants (for both the fMRI and the eye tracking studies). The presentation consists of a slide deck that can be used to train study participants on security awareness to assist in identifying targeted interventions. The phishing awareness content of the presentation links to the Cyber Trust taxonomy and contains targeted phishing lures and cues. This presentation can be delivered through either instructor-led or online formats.

2.4.3.3 Cybersecurity Workplace Training Survey and Education

In conjunction with our experimental work, we engaged in theoretical work on systematically applying our knowledge of cybersecurity to training design in the workplace. We developed a survey to assess a company's current training efforts. We completed two publications to help educate the broader public on how to incorporate cybersecurity training more effectively into the workplace. One is targeted toward a more general audience of stakeholders in cybersecurity [81] and the other is directed toward human resource professionals [82].

2.5 PUBLICATIONS

1. Brummel, B. J., Hale, J., & Mol, M. (In Press). Training cyber security personnel. In S. Zaccaro, R. Dalal, and L. Tetrick (Eds.). *The Psychosocial Dynamics of Cyber Security*. Taylor & Francis.

2. Beyer, R. E., & Brummel, B. J. (2015). Implementing effective cyber security training for end users of computer networks. *SHRM-SIOP Science of HR Series: Promoting Evidence-Based HR*. Society for Human Resource Management and Society for Industrial and Organizational Psychology.

3. Hale, M. & Gamble, R. (2014). Toward Increasing Awareness of Suspicious Content through Game Play, ***IEEE World Congress on Services,*** pp. 113-120.

4. Staggs, J., Beyer, R., Mol, M., Fisher, M., Brummel, B., & Hale, J. (2014). A perceptual taxonomy of contextual cues for cyber trust. *Proceeding of the Colloquium for Information System Security Education (CISSE), 2*, 152-169.

THRUST 3 – CYBER TRUST AND SUSPICION: A HUMAN-CENTRIC APPROACH

The research efforts for Thrust 3 was led by the Texas A&M University with Dr. Hongbin Wang as the principal investigator (PI). The descriptions and results in this chapter were provided by the Thrust 3 PI and Thrust 3 team.

3.1 CYBERSECURITY WITH HUMANS IN THE LOOP

Conventional wisdom has regarded cyberspace security as a pure technology issue – sophisticated information techniques, tools, and policies are a must in order to detect and defeat threats (for defense) and develop and deliver attacks (for offense). At a more foundational level, however, it is now clear that cyberspace security is also, if not more, a human-social phenomenon - how human operators, be they everyday internet users or national intelligence analysts, perceive and make sense of cyber events "closes the loop" and is therefore essential for the ultimate success (or failure) of cyberspace security. Unfortunately, the significant role of human operations in cybersecurity cycles has largely been ignored or less understood thus far.

This is particularly true with regard to cyber trust and suspicion, two fundamental concepts in cyberspace security. The bottom line is that a cyber attack (e.g., worm or sabotage) is more damaging and harmful if it is stealthy and with disguise, and disasters occur when a non-trustworthy source is trusted. One central question in cyberspace security is therefore to understand how cyber trust and suspicion are represented, measured, monitored, and managed. Any security-oriented algorithms and systems must have some form of trust and suspicion management built-in, though often implicitly.

It is critical to realize that trust and suspicion are fundamentally psychological constructs and human traits. Automated trust and suspicion management systems through sophisticated computer algorithms are certainly desirable and have been quite successful. We have to accept, however, that it is humans (not machines) that trust and suspect, and that the computer algorithms have to be based on sound theorization of human trust and suspicion intuition in order to be useful. Such a theorization has the potential to make automated solutions even more powerful, robust, and realistic by inserting key human factors such as motivation, intention, attention, perception, belief, and emotion into the picture. In addition, in cases when computer algorithms are inconclusive, it is human operators' trust and suspicion insights and intuitions that often connect the dots and close the loop.

Trust and suspicion are naturally loaded concepts. In a classic review of the concepts of trust and suspicion, Deutsch defines "trust" as follows: "An individual may be said to have trust in the occurrence of an event if he expects its occurrence and his expectation leads to behavior which he perceives to have greater negative motivational consequences if the expectation is not confirmed than positive motivational consequences if it is confirmed"[83]. And he defines "suspicion" as follows: "An individual may be said to be suspicious of the occurrence of an event if the disconfirmation of the expectation of the event's occurrence is preferred to its confirmation and if the expectation of its occurrence leads to behavior which is intended to reduce its negative motivational consequences"[83]. It is clear from these definitions that trust and suspicion are closely linked to motivations and subsequent decision making. A trust-minded person, compared to a suspicious person, is more willing to take risks in an uncertain environment and therefore is more likely to be caught off-guard if something goes wrong.

It is in this sense that trust and suspicion are relevant and important factors in cyberspace security. Cyberspace fundamentally alters the dynamics of inter-personal and human-machine relationships. The Internet and social media allow never-met-before people to know each other and become "friends." Communications become so fast and cheap that everybody is inundated with information. In these situations, what do "trust" and "suspicion" mean? We can use sophisticated machine learning techniques to mine past data and develop algorithms to tell us the precise likelihood and consequence of such trust in the past, but we may still be uncertain about the intention/implication of the message and the sender, and about the action we should take. Needless to say, all these factors are interwoven and together they form the landscape of today's cyberspace security. A foundational understanding of the underlying dynamics of cyber trust and suspicion is clearly needed for achieving better cybersecurity. It can only be acquired when we close the loop between information systems and human operators, and study their interoperability.

3.2 TOWARDS A MODEL OF HUMAN-CYBER TRUSTWORTHINESS

Although traditionally used for describing human-human social relations, previous research has shown that trust is also an integral aspect of human-machine interactions. In automation, for example, trust facilitates users experience by reducing uncertainty about machine reliability and accountability. However, trust in cyberspace is not well understood. In such contexts, individuals interact with each other via the Internet. The other "parties", be it an individual, a machine, or a software agent, are often invisible, and can only be loosely defined if at all identifiable. Can people develop trust relations in such an environment? If so, how does it work, and how do trust dynamics in cyberspace differ from those in traditional contexts? These questions are especially important as human life is increasingly immersed in cyberspace, and cybersecurity becomes an ever more critical concern. In Thrust 3, we aim to answer these questions by drawing insights from relevant literature on these topics and discussing how three pertinent psychological constructs (trust, suspicion, and uncertainty) might manifest themselves in cyberspace and cybersecurity.

Here we propose a new framework for conceptualizing and modeling cyber trust. As shown in Figure 40, a key feature of this framework is the two dimensions common to trustworthiness scenarios. On the one hand, there is a continuum in terms of the nature of between-entity relations, which can be either thin or thick. Because relations tend to be strategic, we hypothesize that the thickness of a relation is determined by the number of connections between entities. For example, family members have thick relations while strangers have thin relations. The thickness of a relation directly affects trustworthiness and trust dynamics. In cyberspace, many users communicate via the internet and often for specific needs, which leads to thin relations as compared with traditional interpersonal and human-machine interactions.

On the other hand, trustworthiness can also be described in terms of uncertainty.

Whereas people anticipate surprises in highly uncertainty situations, surprises are less likely to occur in situations involving low levels of uncertainty. Therefore, the relative degree of uncertainty requires different modes of trust operations for optimal user performance. Due to the large number of users and systems involved, each with different goals, cyber trust is often highly uncertain.

Taken together, the two dimensions define a space in which different trust dynamics can be represented. It is important to note that dimensions need not be orthogonal, which allows for between-dimensional interactions. For example, whereas interactions among family members or friends are characterized by thick relations and low uncertainty (e.g., a wife asks her husband to pick up their children from school), interactions between people connected in cyberspace are

characterized by thin relations and high uncertainty. This is because in most cyber scenarios the user does not know the other parties with whom he or she is interacting. Moreover, the interactions are indirect (e.g., mediated by computer systems) which contributes additional noise to the situation. Consequently, cyber interactions often lack important qualities that normally guide or facilitate interpersonal relations, such as reputation or commitment.

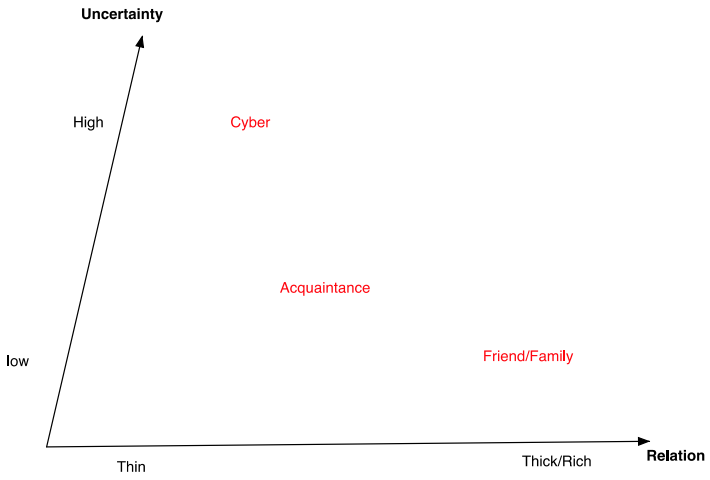


Figure 40: Uncertainty Perception and Human-cyber Trust

In a cyber environment, there can be many reasons for unexpected, and therefore, untrustworthy machine behavior (e.g., hardware or software failures). Cyber-attacks can be one of them. However, cyber-attacks are special due to their inherent stealth and uncertainty. Cyber-attacks are often created with unclear intentions and unspecified targets [84]. They are also difficult to detect and manage. Once they have entered a system, their payload may be immediate or delayed until a specific date, or a set of commands may be executed (e.g. zero-day exploits). Moreover, disruptions to system performance may be localized or widespread. Whereas some malicious code disrupts the behavior of specific operations (e.g., buffer overruns), others compromise the entire system.

In the present study we investigate how uncertainty perception affects human-computer trust in a cyber-attack scenario. In order to examine the effects of uncertainty on human-computer trust and reasoning, we measured participants' adherence to cues about the reliability of a radar display that was susceptible to cyber-attacks. Their main task was to decide whether to fire or hold fire depending on the location of the jet on the display (see Figure 41). Participants were told to fire if the jet appeared in the kill zone (the center of the radar) and to hold fire if the jet appeared in the fly zone (the periphery outside of the kill zone). Occasionally a cyber-attack would cause the jet to appear in a false location on the display (e.g., the jet appeared in the fly zone but it was actually in the kill zone), which made the participants' task challenging – for example, they might accidentally fire on a jet that was wrongly displayed in the kill zone. The frequency of attacks varied systematically so that the display would fluctuate between a specific range of reliable and unreliable states of performance. After participants entered their response, they heard one of two distinct feedback tones: one tone for correct responses and another tone for incorrect responses.

Trust was measured as a function of participants' response accuracy and response time. If participants calibrate their trust of the system appropriately (e.g., based on a relatively accurate estimate of the likelihood of cyber-attacks), then they will adhere to the response feedback tones and

performance will remain relatively constant throughout the task. That is, their responses will be continuous across trials if they distrust the display and trust the response feedback. However, if participants inappropriately calibrate their trust of the system (e.g., they completely distrust the display and response feedback cues), then they will ignore the response feedback tones and performance will decline across the attack conditions with strong performance in reliable trials and weak performance in unreliable trials.

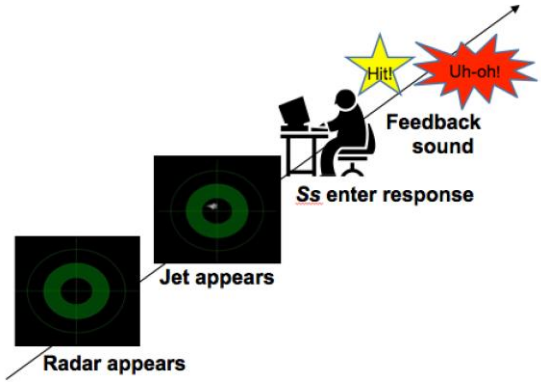


Figure 41: Human Computer Trust Game

Overall, the results of the present study indicate that uncertainty perception plays an important role in human-computer trust. In both experiments, poor decision-making performance was observed in the low reliability conditions, relative to the high reliability conditions. However, there was a non-linear trend in both accuracy and response times, with poorest performance in the least certain conditions, rather than the least reliable conditions. That is, people had the greatest difficulty deciding how to respond when there was an equal (or almost equal) probability of entering a correct response as entering an incorrect response. Although it is unclear how uncertainty perception leads to suspicion (and the resulting effect on performance), it appears that the current findings support the general notion of a trust spectrum with trust on one end, distrust on the other, and uncertainty in the middle. In the current experiments, people calibrated their trust according to the feedback with high trust in the most reliable conditions, suspicion/distrust in the least reliable conditions, and uncertainty when reliability was most variable.

THRUST 4 – USING NON-INVASIVE SENSORS TO PREDICT TRUST AND SUSPICION IN HUMAN OPERATORS

The research efforts for Thrust 4 was led by the Syracuse University with Dr. Leanne Hirshfield as the principal investigator (PI). The descriptions and results in this chapter were provided by the Thrust 4 PI and Thrust 4 team.

With specialties in human-computer interaction (HCI) and machine learning (ML), Syracuse University's (SU's) research explores the use of non-invasive cognitive, physiological, and behavioral sensor measurement to classify user states passively in order to enhance usability testing and adaptive system design. We have a strong record of accomplishment in applying our non-invasive physiological sensor research in the cyber domain. We run a state-of-the-art HCI 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, a desk mounted faceLab eyetracker, two Tobii eyetrackers, several wireless galvanic skin response sensors from Affectiva, faceReader emotion recognition software, Morae usability testing software, key and mouse loggers (courtesy of our Assured Information Security, Inc. (AIS) team members), and several computer workstations for use in experiments.

SU's ongoing research within this effort involves (i) continually updating our capabilities to provide 'ground truth' information about a user's level of trust and suspicion during HCI using the lab's sensors, and (ii) using these capabilities to measure trust and suspicion during a range of experiments stemming from the Cyber Trust and Suspicion (CTS) research initiative.

4.1 PROGRESS YEAR 1

During Year 1 the SU team conducted theoretical research about the definition of suspicion, and we designed and conducted data collection on two experiments. One of the primary roles of SU within this collaborative effort has been to measure trust and suspicion. Although trust has been heavily studied in the theoretical and empirical literature, the construct of suspicion has received very little attention. A theoretical understanding of suspicion is needed in order to properly define, manipulate, and measure suspicion.

Therefore, we ('we' in this case refers to Dr. Hirshfield, along with colleagues Phil Bobko and Alex Barelka) participated in a review of the "suspicion" literature in order to synthesize literature across the social sciences, including management, marketing, communication, human factors, and psychology [85]. Our focus was on state suspicion in information technology (IT) contexts. We found that the pre-existing literature on suspicion was rather sparse, and the majority of researchers who used the term did not define the concept and/or measure it. After reviewing the studies that did define and/or measure suspicion, we noted conceptual commonalities among them. In particular, we noted that the three facets of (i) uncertainty, (ii) malintent, and (iii) cognitive activation were often implicitly or explicitly used in discussions of suspicion. Thus, state suspicion in IT contexts was defined as the simultaneous occurrence of these three facets:

State suspicion in IT contexts is a person's simultaneous state of ***cognitive activity, uncertainty****, and* ***perceived malintent*** *[boldface added for emphasis] about underlying information that is being electronically generated, collated, sent, analyzed, or implemented by an external agent* [86].

In addition to partaking in the theoretical suspicion research, SU also conducted two data collection efforts during the first year of the research project. These data collection efforts are described next.

4.1.1 SURVEY ON THE EFFECTS OF CYBER ATTACKS ON HUMAN OPERATORS

In the Fall of 2013 we collected data (n = 101) by asking respondents about the (self-reported) cognitive and emotional effects that various computer malfunctions would have on them. We developed a survey addressing computer users' experiences with cyber attacks [87]. To do so, we first needed to establish a list of cyber attacks, and associated symptoms, that users may have encountered. We thus reviewed research papers in the cybersecurity domain, and we examined websites and other informational material from cybersecurity companies. We then created a list of 38 items that represented the types of cyber attacks found in our reviews. Here we should acknowledge that the symptoms most users experience could be the primary goal of a cyber attack, or they may be the secondary effects of an attack (e.g., a slow internet connection could be an attack itself or a symptom of malware). For simplicity, from this point on we will refer to both primary and secondary symptoms/effects as 'cyber attacks.'

In order to understand the potential effects of cyber attacks (e.g., capacity to induce suspicion, emotional reactions, increased workload), our survey linked the 38 types of cyber attacks to potential effects on human operators. In the survey, participants were first asked for basic demographic information (age, gender), and they also filled out a computer aptitude survey. Next, participants were asked to estimate their reactions to each of the 38 types of cyber attacks. Possible reactions were rated on 12 scales; each scale was a single item scale (generally ranging from "very little" to "very strongly or extremely").

The first seven scales directly mirrored our interest in suspicion and cognitive load. Those scales, and their accompanying descriptions as listed in the survey, were:

Trusting - having the intention to accept vulnerability based upon positive expectations of the intentions or behavior of another person or agent.

Suspicious - a feeling of uncertainty and perceived malintent about the underlying information or actions of an external agent.

Mental Demand - one of five items taken directly from the NASA-TLX survey, which has been shown to reliably measure perceived workload.

Temporal Demand - one of five items taken directly from the NASA-TLX survey, which has been shown to reliably measure perceived workload.

Effort - one of five items taken directly from the NASA-TLX survey, which has been shown to reliably measure perceived workload.

Performance - one of five items taken directly from the NASA-TLX survey, which has been shown to reliably measure perceived workload.

Frustration Level - one of five items taken directly from the NASA-TLX survey, which has been shown to reliably measure perceived workload.

To create a measure of "workload," we summed responses to items 3-7. We also used five scales that describe a range of emotional or mental states. The scales were taken from the positive and negative affective schedule (PANAS) short form survey.

Alert - quick to perceive and act.

Upset -unhappy, disappointed, or worried.

Nervous - anxious or apprehensive.

Afraid - feeling fear or anxiety; frightened.

Aggravated - To be provoked, roused to exasperation or anger. (Note: the PANAS item was labeled *hostile,* but we used "aggravated" as a label, as it seemed more relevant to our context; aggravation/anger is considered similar to, though not exactly the same, as *hostile* in emotion research.

The survey was administered online, and respondents (N = 101) were comprised of students, faculty, and staff from a university in the Northeast who considered themselves to be proficient computer users. Of the respondents, 64 were female. About half of the respondents (*N* = 55) were between the ages of 18 and 21, while 27 respondents were between the ages of 22 and 30, and 31 respondents were over the age of thirty. Participants were asked to rate each of the above twelve dimensions once for each of the cyber symptoms. In total, participants completed 456 items in the survey (12 ratings × 38 cyber attacks = 456).

With our hypotheses in mind [87], we used the survey data to empirically determine cognitive and emotional correlates of suspicion. We then used the data to create a latent structure for categorizing cyber attacks based on their effect on human operators. An initial step was to conduct a principal components, exploratory factor analysis of the 38 types of cyber attacks. As input, we used respondents' reactions to the "suspicion" scale (i.e., how suspicious would they become if a particular type of attack occurred). We used a visual scree test and eigenvalues greater than 1.0 to determine the number of factors (five). We then rotated the components, using varimax rotation, to aid in factor interpretability. The results of the factor analysis are presented in Table 9. To aid in interpretability, the column labeled "items" provides the highest loading items on that factor (loadings over .55). The column labeled "factor" provides our qualitative labeling of the subset of items based on the content of the items.

We categorized the items in Factor 1 as 'well-known and overt attacks.' We posit that many computer users' mental models are associated with these most commonly publicized and destructive cyber attacks. Computer viruses, destruction of files, and dramatic increases in pop-ups are all represented in this factor.

We labeled the second factor as 'low and slow attacks, as well as errors that might be commonly made by humans.' As indicated by their title, 'low and slow' attacks are a subset of cyber attacks that can often go undetected by keeping the strength and/or duration of attacks at a suitable low level. For example, the StuxNet worm was able to operate undetected for months, and if it were not for human operator error, it may have never been detected. The StuxNet attack infiltrated the computer network of Iran's nuclear enrichment program with the intent to covertly disrupt and degrade its nuclear program over an extended period of time. This mode of attack was reportedly never detected by the Iranians themselves, but rather discovered three months after its initial infiltration, once it had migrated beyond the bounds of its original target. This type of attack may often fly under the radar of human operators as they attribute these malfunctions to 'normal' slow networks and computer systems. We also suggest that according to most users' mental models, these occurrences are

common, and they are often related to slow and/or 'old' computers or slow internet connections. Errors that might commonly be caused by humans are also found in this factor, such as typing the wrong key, pressing the wrong mouse button, etc.

Table 9: Factor Analysis Results on Suspicion Likert Item

Factor	Items
Factor 1: Well-known and overt attacks	My computer suddenly crashes. I receive a warning that my computer is about to crash. The program I am working on suddenly crashes. Program or data files corrupted. Computer denies requests for information. System performance degrades no fault. I receive a warning that my computer has a virus. My monitor flickers more than normal or becomes unusually blurry. The number of pop-ups I receive is much greater than normal.
Factor 2: Low and slow attacks, as well as errors that might be commonly made by humans	The keyboard response is slower. The mouse is very unresponsive or sluggish. Websites give me inappropriate error messages. The internet is unusually slow. The program I am using gives me erroneous error messages. The program I am using slows down considerably. I click on a key and the wrong keystroke appears on my screen. I click on my mouse and the wrong side of the mouse is triggered.
Factor 3: Common errors that are not perceived to be the fault of the user	My print requests are routed to the wrong printer. While using two monitors, the desktops on each screen swap. For no good reason I receive an error message when I print. I receive local weather alerts or event warnings that are incorrect or misleading.
Factor 4: Perceptual/memory errors	Posts are put on my social networking sites without my permission. It seems objects out of my view randomly disappear. Objects on the screen randomly disappear.
Factor 5: Evidence of tampering or remote control of the computer	Additional text is inserted into my emails or posts. I hear strange music or voices coming from my computer speakers. When I opened a file a colleague sent me it seemed a bit "off".

The third factor is labeled 'common errors that are not perceived to be the fault of the user.' Our respondents were affiliated with a large university, and we presume that most of these items relate to systems that would likely be set-up by a third party; i.e., having access to multiple printers, or multiple desktop monitors, would be common. In these scenarios, it may be usual to assume that one's computer system has been improperly configured. Incorrect weather forecasts and alerts were also in this category, perhaps because people often comment (and even joke) about the inaccuracy of weather forecasters.

It was somewhat difficult to distinguish between Factors Four and Five, and we believe more research is needed to further explore users' reactions to these two types of attacks. Also, two of the three items in Factor Four include the word 'disappear', which may have contributed to the cohesiveness of those items. Nonetheless, we labeled the fourth category 'perceptual/memory errors.' These appear to be errors that might be indicative of a cyber attack, but which might also be attributed to mistakes of perception (disappearance of display objects) or memory (unremembered posts) on the part of the operator. We labeled the fifth factor as 'evidence of tampering or remote control of the computer.' These items seem to map onto strange or unexpected events that could

indicate an outside entity has remote access to the computer and may be actively altering content or processes.

We suggest that these factor-analytic results are unique – and helpful to researchers and practitioners - in that they empirically suggest an underlying, latent structure regarding how different types of cyber attacks are interpreted and/or experienced by computer users. In our paper [87], we show how each of the five factors correlates to respondents' self-reported emotional state, level of trust, and cognitive workload. We also describe several suggestions for future cyber training and security measures based on the results from our survey.

4.1.2 RESCHU EXPERIMENT

During this experiment we recorded participants' mouse movements and their brain data (via functional near-infrared spectroscopy) while they worked with the Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU), a testbed where participants control unmanned vehicles. This experiment was done in concert with the AIS team, as one of our goals has been to determine 'ground truth' information about trust and suspicion using fNIRS and to use that information to find patterns in remotely gathered metrics (key, mouse movements and webcam) that are indicative of trust and suspicion as well. We do this because there is a need to measure users' changing levels of trust and suspicion remotely, where data is only recorded via monitoring of an individual's HCI.

Throughout the experiment we introduced manipulations into the testbed to cause it to function improperly. The goal of the study was to measure the response to these malfunctions in both the brain data and the mouse data, to determine if a significant difference was noted. RESCHU is an online experimental test bed that allows operators to control a team of different types of Unmanned Vehicles (UVs). A screen shot of the RESCHU environment is shown in Figure 42. All vehicles are engaged in surveillance tasks, with the ultimate mission of locating specific objects of interest in urban coastal and inland settings. Users can control their UVs by clicking on the UV and assigning it to one of several 'targets' available on the screen. At that point, the assigned UV will indicate that it has been set to approach that target by showing a line between the UV and its assigned target.

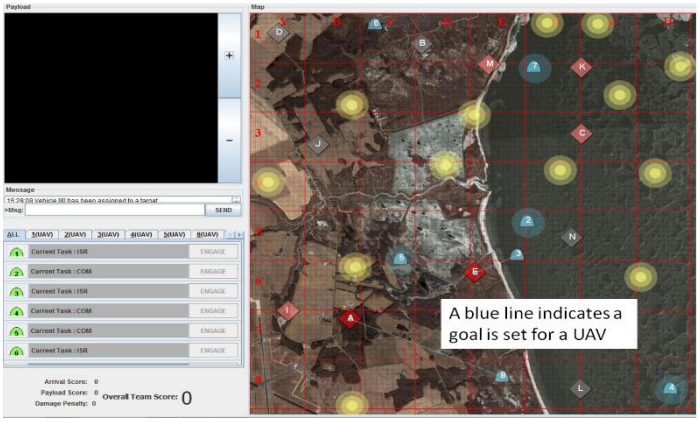


Figure 42: A Screen Shot of the RESCHU Environment

4.1.2.1 Conditions

In this experiment, the control condition included subjects simply working with the RESCHU environment under normal working conditions. In addition to the control, there were two experimental conditions.

1. *Remove UV Target Within User's Locus of Attention:* The first condition will remove a target (which also removes the line that is drawn between the UV and its assigned target) while that UV is within the user's locus of attention. Henceforth, this is referred to as an overt manipulation.

2. *Remove UV Target Outside User's Locus of Attention:* The second condition also removed an assigned target from a UV, but it did so when the UV was not within that user's current locus of attention. Henceforth, this is referred to as a covert manipulation.

These manipulations were selected because there was support in the physiology and attention domains suggesting that we could reasonably expect to see a difference in the physiological response from computer users' when the goal was removed within, rather than outside of, the users' locus of attention. We also expected the goals to affect users' levels of trust and suspicion, enabling us to tie the results from our physiological sensors to our D5 framework.

Ten participants completed this experiment in the Fall of 2013. Participants' mouse movements and cognitive data (via fNIRS) were measured throughout the experiments. Unfortunately, as detailed in our Tech Report on the matter, we were unable to locate significant differences in the fNIRS data between the conditions. AIS reported similar lackluster results on their analyses with the mouse tracking data from the studies. In our exit interviews with participants, over half indicated that they did not notice the manipulations, and that the difficulty of working with RESCHU itself had them completely immersed in the task, without being able to notice the disruptions. Building on these findings, our subsequent research has focused on selecting tasks and manipulations that are better suited to reliably manipulate trust and suspicion in participants.

4.2 PROGRESS YEAR 2

During Year 2 we designed and implemented three experiments relating to the CTS project.

4.2.1 MEASURING TRUST AND SUSPICION VIA KEYLOGGING AND SENTIMENT ANALYSIS

With the integral role that trust plays during successful interactions, there is a need to objectively measure trust, and the related construct of suspicion, during computer-mediated-communication (CMC). Furthermore, these measures should be made unobtrusively, so as not to disrupt computer users. The primary goal of this research was to use data from a keylogger to build machine learning models capable of predicting a user's trust and suspicion during CMC [88]. This research was done in collaboration with the AIS team, and they provided us with their custom keyloggers. We designed a set of experiments that were theoretically grounded on prior trust and suspicion research. During the experiments, participants interacted in dyads on a variation of the desert survival teamwork scenario, using an instant messenger (IM) application while manipulations were introduced to affect their trust and suspicion toward their partners. After a five-minute IM session, each participant filled out a survey indicating their level of trust and suspicion toward their partner during the prior session (Figure 43). These values were included as 'ground truth' for later machine learning analysis.

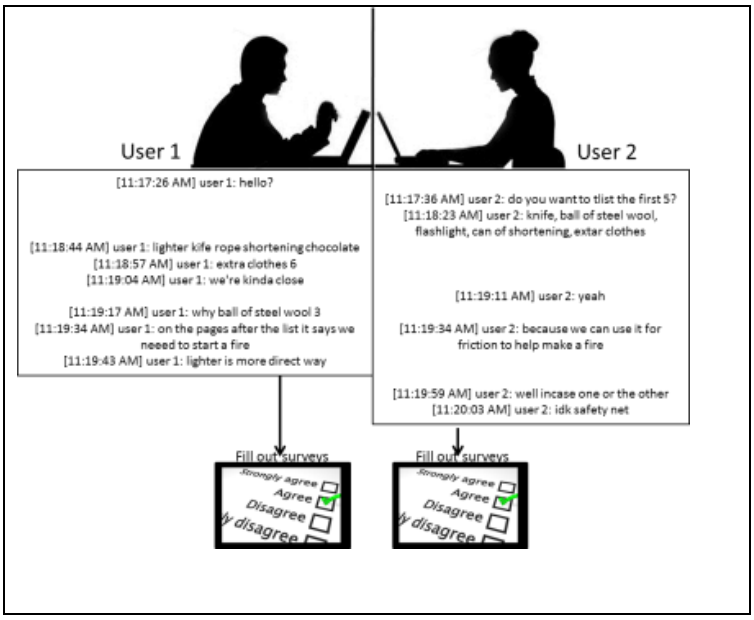


Figure 43: Users are placed in dyads where they chat during five minute sessions. Surveys are filled out after each session. The process is repeated for all dyad combinations

Data was gathered throughout the experiments with a keylogger in order to build models that classify a user's trust and suspicion toward his or her partner with accuracy well above baseline.

Features were generated that represented keystroke timings as well as textual descriptions such as content of words, grammar, and phrases. AIS generated five features for keystroke timing, as illustrated in the AIS report. The features were key hold time for the first key (KHT1), key hold time for the second key (KHT2), key interval time (KIT), key press latency (KPL), and key release latency (KRL). KHT represents the duration a key was pressed, while KIT represents the time between the release of the first key and the press of the second key. KPL is the time between the press of the first and the press of the second key, while KRL represents the time between the release of the first and the release of the second key. It is worth noting that there is collinearity among these features, which is common in real-world time series datasets. Text collected from the participants' IM sessions was also subjected to linguistic and sentiment analysis by the SU team to uncover the uses of different features. Each individual's text from a five-minute session was analyzed using a Python program to generate the features described below. The program conducted basic feature selection from each participant's textual data during each five-minute IM session. The 19 features are available in Table 10. IM sentiment (positive or negative) was also recorded using a Naïve Bayes classifier trained on a text corpus of 2,000 previously classified movie reviews. This was completed using the Natural Language Toolkit (NLTK) platform for Python.

Table 10 The NLP Features generated from IM data

Feature Category	Features
Spelling and Grammar	first_letter_cap - the number of messages that begin with a capital letter. quote -The number of properly formed quotations in the messages. ie., "My name is Rachel" end_punctuation – the number of messages that ended with proper punctuation. (number of messages with punctuation as the last character). comma - number of IM messages that a "," appeared in during the 5 minute session misspelledWords - number of misspelled words in 5 minute session missing_apostrophe - number of messages containing a badly formed contraction: doesnt, wont, etc.
Quantity of writing	totalWordCount - total words in 5 minute session num_posts - the number of messages a user sent in a particular session.
References to self	lowercase_i - the number of messages that a lowercase i was used in. uses_i - number of times lower or uppercase _I is used throughout a session (i.e.," I think the compass would be the most useful.")
Sentiment	exclamation - number of IM messages that a "!" appeared during the 5 minute session negative - number of messages classified as negative by the Naive Bayes classifier positive - number of messages classified as positive by the Naive Bayes classifier question_mark - number of IM messages that a "?" appeared in during the 5 minute session total_likeable_abbreviations - number of IM messages that contained 'lol', 'lols', 'lolz', 'ha', 'haha', 'lmao' or happy emoticons ☺.
Other slang and words of interest	swear - number of IM messages that contained a swear word during the 5 minute session idk - number of IM messages that a 'idk' appeared in during the 5 minute session saboteur - an overall count of the number of individual messages containing at least one mention to saboteur. Other abbreviations – this included 'omg', 'wtf', 'wth', 'smh', 'fml', 'rofl', 'nbd'

After building a supervised classification model based on a neural network, we were able to predict a user's level of trust in their partner with 91% accuracy, and their suspicion toward their partner with 68% accuracy [88].

4.2.2 PREDICTING PERSONALITY, PROPENSITY TO TRUST, AND NEED FOR COGNITION WITH USERS' SOCIAL MEDIA POSTS:

Individuals have trait specific individual characteristics that may affect their state trust and suspicion in a given scenario. This research project attempted to predict Facebook users' personality type, need for cognition, and propensity to trust using just their social media posts [89]. Personality type, need for cognition, and propensity to trust were all individual characteristics posited to have a direct effect on state suspicion, as noted in our prior theoretical paper [85]. We ran an experiment that combines natural language processing and machine learning methods, in order to predict people's scores on the Big Five personality inventory, Mayer's propensity to trust survey, and the need for cognition psychometric tests. Our machine learning predictive model showed promising results in personality prediction and prediction on people's need for cognition scores. However, we were not able to achieve accuracy greater than random at predicting user's propensity to trust from their Facebook data [89].

4.2.3 MEASURING EMOTIONAL STATE CHANGES WITH A WEBCAM:

As noted previously, one of our goals has been to determine 'ground truth' information about trust and suspicion using fNIRS, and to use that information to find patterns in remotely gathered

metrics (keyboard, mouse movements, and webcam) that are indicative of trust and suspicion as well. We do this because there is a need to measure users' changing levels of trust and suspicion remotely, where data is only recorded via monitoring of an individual's HCI. In our prior theoretical research on suspicion [85], we noted that suspicion causes an increase in negative affect. In this experiment we aimed to use data gathered from a webcam to gauge emotional states [90], which are a key component of suspicion. Although currently there are commercial software packages that predict emotional states using a standard webcam (such as Noldus FaceReader) these software packages lack the capability to update their predictive models with new data, and they tend to be trained on 'extreme' facial expressions (i.e., a large smile indicating happiness), and they do a poor job at distinguishing emotion in the absence of these extreme expressions. Thus, we designed our own feature extraction method to take data from a webcam during HCI, extract features from that data, and predict the emotional state of that person in an attempt to measure negative affect via the webcam. In particular, we focus on features that include pupil size measurements, facial expression detection, and eye gaze information in order to detect subtle, naturally occurring emotional states.

Emotions have different levels, and the most common ones to be studied in the field of HCI are the valence and arousal dimensions. The valence dimension varies from negative to positive reactions, whereas the arousal one varies from calm to very excited reactions. In this experiment, webcam data from 20 participants were collected while they listened to positive, neutral, and negative auditory stimuli chosen from the IADS database. Participants were asked to remain as steady as possible to maintain a constant distance from the camera. They sat within a controlled lighting environment while listening to the stimuli. These stimuli were chosen to elicit "compatible" positive and negative emotions (i.e. at the same intensity level). The neutral stimuli were chosen to be used as a frame of reference. We explored various image processing techniques, including the use of wavelets in the extraction of pupil size, gaze patterns, and parameters to analyze facial expressions. Changes in the shapes of the eyes and mouths were the parameters of interest. Using a suitable classification process, we present promising classification results [90] in differentiating between reactions to positive, negative, and neutral stimuli.

4.3 FINAL REPORT SUMMARY

In summary, the SU portion of this effort focused on (i) developing our capabilities to measure and predict changing trust and suspicion using different combinations of the labs' sensors, and (ii) running human subject experiments focused on trust and suspicion that integrate research from other team members. This research has resulted in five publications and one Tech Report, which will be published in the near future.

1. 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.
2. 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. *Accepted in* the Human Factors and Engineering Society Journal (HFES 2014).
3. Solinger, C. **Hirshfield, L**. Hirshfield, S., Friedman, R., Lepre, C. Beyond Facebook Personality Prediction: A Multidisciplinary Approach in Predicting Social Media Users' Personality. *Accepted in* Proc. of the International Conference of Human Computer Interaction, Crete, Greece. (2014).
4. Sommer, N., **Hirshfield, L.,** Velipasalar, S., Our Emotions as Seen Through a Webcam. *Accepted in* Proc. of the International Conference of Human Computer Interaction, Crete,

Greece. (HCII 2014).
5. **Hirshfield, L.M.,** Dora, R.Weber, C., Bobko, P. Using Keylogger Data to Predict Trust, Deception, and Suspicion During Online Interactions. Tech Report. 2014
6. **Hirshfield, L.M.,** Bobko, P., Barelka, A., Costa, M., Finomore, V., Funke, G., Knott, B., Mancuso, V., The Role of Human Operators' Suspicion in the Detection of Cyber Attacks. International Journal of Cyberwarfare and Terrorism (2015).

4.3.1 MODEL OF THE PHYSIOLOGICAL CORRELATES OF TRUST, DISTRUST, AND SUSPICION

One valuable outcome from the research described above has been a model describing the physiological correlates of trust, distrust, and suspicion (Figure 44), which is being prepared (along with supporting experiment results) for publication. The model is grounded in the theoretical suspicion and trust work done by Bobko, Barelka and Hirshfield [85], and it combines the findings from Dr. Hirshfield's prior AFOSR physio- sensor experiments from this effort, as well as from prior AFOSR efforts [87][90][88][91][92][93]. Although the model has implications for interpersonal communications, the focus of the model is on the physiological correlates of trust, distrust, and suspicion within the IT domain. The rest of this section describes Figure 44 in detail.

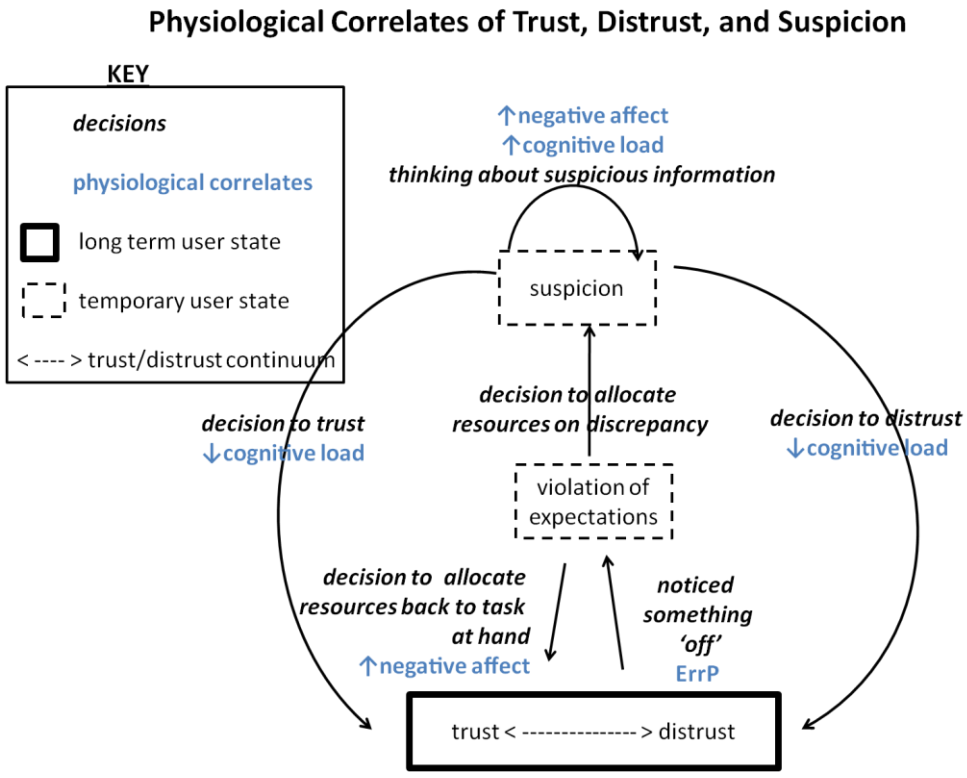


Figure 44: Hirshfield's hypothesized physiological correlates of trust, distrust, and suspicion

At a high level, the model in Figure 44 depicts the mental states that a computer operator transitions between when encountering problems (likely caused by computer manipulation

techniques) with their computer system. These computer operators spend most of their time in long term user states of trust and distrust (bottom of Figure 44). While the literature does not fully agree on the relationship between trust and distrust, this model views trust and distrust as uni-dimensional or, at the very least, highly overlapping multi-dimensional, constructs. In practical terms, computer operators are continually updating their levels of trust based on interactions with their environment.

Many operators are predisposed to have a high level of trust (or distrust) that usually serves as a lens by which they view their interactions in their environment (called propensity to trust). Most computer operators tend to be cognitive misers; they prefer to spend their time focusing their cognitive resources on the task at hand, rather than spending effort interpreting interactions from the computer system or an external agent. For example, operators with a high propensity to trust view interactions with others as honest and straightforward, enabling them to adapt a truth heuristic that can limit workload during interactions with their environment. Viewing the environment through a predisposed 'high trust' lens is one piece of information built into people's mental model (i.e., one's internal model of how the world around them works) that helps them to interact with their environment with minimal cognitive load.

When something occurs that violates an operator's mental model (such as a program crash, or an obviously deceitful email allegedly from a coworker) that operator's expectations are violated. In that case, something called an error potential (ErrP) occurs in the computer user's brain. This can, and has been, detected with EEG, even when the violation of expectations is extremely subtle [93]. Once a violation of expectations has occurred, the user can either 1) attend to that item and become suspicious or 2) return to the task at hand.

If a computer user becomes suspicious, he experiences a "feeling of uncertainty and perceived malintent about the underlying information or actions of an external agent" [85]. This involves an increase in negative affect as well as cognitive load, as he generates hypotheses about the source and meaning of the discrepancy. The suspicious state ends when the user decides either to re-enter a trustful state (go back to the task at hand) or to become distrustful (which likely involves an action such as making a call to the IT department, or running antivirus software).

If the user does not become suspicious (and many computer users experience violations in expectations time and time again without ever becoming suspicious), then that user will remain in his long term state residing somewhere within the trust/distrust continuum. It is worth noting that the Bobko, Barelka, Hirshfield (2014) [85] theory states that distrust increases likelihood to become suspicious while trust buffers suspicion. This means that a person who is predisposed to view the world with high distrust will be quicker to become suspicious than will his highly-trusting peer. In the case where the user experiences a violation in expectations without becoming suspicious, there will likely be an increase in negative affect (though not always) if the user believes the discrepancy is a result of his or her own human error. Alternatively, the user may attribute the discrepancy to problems with the computer system (i.e., "this program is so slow!") and trust may be lowered.

As stated previously, the model in Figure 44 represents the culmination of all of Dr. Hirshfield's prior theoretical and experimental work in this domain. The research conducted during this two year effort was invaluable in helping to develop the model. The importance of this model is that it is not only grounded in theoretical work from research on trust, distrust, suspicion, cognitive load, and cognitive heuristics, but it shows how transitions between states of trust and suspicion can be measured in experimental settings with physiological sensors.

THRUST 5 – ASSESSING, ATTRIBUTING, AND MANIPULATING OPERATOR SUSPICION

The research efforts for Thrust 5 was led by the Assured Information Security, Inc. (AIS) with Dr. John S. Bay, initially as the PI, followed by Dr. Barry McKinney. The descriptions and results in this chapter were provided by the Thrust 5 PI and Thrust 5 team.

5.1 SUMMARY

Under the Cyber Trust and Suspicion (CTS) effort, Assured Information Security, Inc. (AIS) sought to explore the utility of cyber sensors in the realm of trust and suspicion. For the purpose of this effort, AIS examined keystroke data collected in experiments conducted at Syracuse University (SU) by Dr. Leanne Hirshfield.

AIS analyzed data from several experiments conducted using cyber sensors to identify statistically significant correlations between suspicion and sensor data. The analysis proved inconclusive for mouse data, but the keystroke data yielded positive results. There is a statistically significant negative correlation between key interval time (KIT) – the time between the release of the first key and the press a second key – and suspicion. An experiment was devised to collect data on suspicion attribution, but the contract was cancelled before the experiment could be conducted.

5.2 INTRODUCTION

The Cyber Trust and Suspicion (CTS) Thrust 5 project consisted of three separate research sub-thrusts spread across three years, with the first thrust (5a) focusing on detecting suspicion, the second (5b) on attributing the perceived and actual sources of suspicion, and the third (5c) on manipulating suspicion. Each thrust served as a feasibility study that produced prototype sensors to collect data during experiments performed at Syracuse University (SU).

The effort produced four distinct cyber sensors to be used in conjunction with the Assured Information Security, Inc. (AIS) IntroVirt platform[5]: a keylogger, a mouse logger, a context logger, and a gaze tracker. These sensors were designed to collect data that could be used to detect and attribute suspicion. Suspicion detection can be performed based on keystroke data, while all sensors (including additional data collection points from the IntroVirt platform) may be useful in the attribution of suspicion.

To investigate the utility of cyber sensors towards suspicion detection, AIS developed a keylogger and mouse logger for use in two experiments conducted at SU – one focused on collecting mouse data while subjects used an unmanned system simulator and another that allowed subjects to collaborate through an instant messenger to solve the Winter Survival Problem [94]. AIS then examined keystroke and mouse data collected during the experiments.

In the second thrust, AIS developed several more cyber sensors to determine if the source of suspicion, perceived and actual, could be determined based upon cyber data. AIS and SU developed a third experiment to collect data on suspicion attribution, with the sources of suspicion scoped to only include deny, disrupt, deceive, degrade, and destroy (D5) effects. Unfortunately, the effort was cut short after the second year, before the attribution experiment was performed. As a result, the suspicion attribution analysis and the suspicion manipulation, scheduled for the third thrust, were not performed.

[5] Assured Information Security Inc., "IntroVirt Final Technical Report," 2013.

5.2.1 OBJECTIVES

The objective of CTS was to assess the feasibility of non-invasive cyber sensors for detecting, attributing, and manipulating user suspicion. This effort was carried out in three thrusts, with suspicion detection research performed during the first thrust, the attribution research performed during the second, and manipulation exercises scheduled for the third.

5.2.1.1 Thrust 5a

Thrust 5a sought to determine the feasibility of using cyber sensors to detect suspicion. Five key tasks were created to achieve this goal:

(1) Develop a keylogger.

(2) Develop a mouse logger.

(3) Support experiments at SU.

(4) Analyze the experimental data to find correlations between suspicion and sensor data.

(5) Document the feasibility of the sensors for detecting suspicion.

5.2.1.2 Thrust 5b

Thrust 5b sought to determine the feasibility of using cyber sensors to attribute the actual and perceived sources of suspicion. Eight key tasks were created to achieve this goal:

(1) Develop an application logger.

(2) Develop a context logger to monitor user behavior within applications.

(3) Develop a gaze detection sensor.

(4) Adapt IntroVirt to create an experimentation platform.

(5) Integrate the sensors into IntroVirt.

(6) Support an attribution experiment at SU.

(7) Analyze the experimental data to find correlations between suspicion source and sensor data.

(8) Document the feasibility of the sensors for attributing suspicion.

5.2.1.3 Thrust 5c

Thrust 5c would seek to determine the feasibility of using cyber sensors to manipulate suspicion. Five key tasks were created to achieve this goal:

(1) Improve and enhance detection and attribution sensors.

(2) Map D5 effects to cognitive states and anticipated responses (*Operator Mapping*).

(3) Support experiment at SU.

(4) Analyze the experimental data to evaluate and optimize the operator mapping.

(5) Document the feasibility of the sensors for predicting operator response to suspicion sources.

5.2.2 BACKGROUND

CTS was a multi-disciplinary effort that relied on established research in psychology, keystroke dynamics, mouse dynamics, and computer network operations (CNO). In particular, the effort required reliable methods for detecting, inducing, and measuring suspicion in experimental trials.

5.2.2.1 Trust and Suspicion

As both trust and suspicion are major research areas, the definition of both are widely debated [95]. For the purpose of this effort, the following definitions apply:

- **Trust** is the confidence in the benevolence of another individual
- **Distrust** is the confidence in the animosity of another individual.
- **Suspicion** is a lack of confidence in the benevolence of an individual, as it must exist between trust and distrust.

The principal area of focus was on suspicion in the realm of computer-mediated communication (CMC). Trust in CMC has been widely studied and current research demonstrates that users can achieve similar levels of trust and suspicion through CMC as they would in face-to-face interactions [96]. The methods used to detect, induce, and measure suspicion in this effort relied on the works of Bobko, Barelka, and Hirshfield [86] and, in particular, Bobko and Odle-Dousseau's [97] state-suspicion survey instrument.

In addition, researchers have found that (neuro-) physiological sensors can be used to measure trust. To date, functional magnetic resonance imaging (fMRI) [49], functional near-infrared spectroscopy (fNIRS) [98], electroencephalogram (EEG) [99], and galvanic skin response (GSR) [98] have all been used to measure trust, working memory, and other cognitive states. As fMRIs limit the movement of subjects, fNIRS, EEG, and GSR were selected for ground-truth in some experimental trials.

5.2.2.2 Keystroke Dynamics

Monitoring keystroke dynamics has proven useful as a biometric authentication and verification technique [100], and has even shown value in identifying demographic features [101]. As such, this effort sought to determine if keystrokes could also provide information about mental state, with regard to suspicion in particular.

Keystrokes dynamics typically uses dyads, – pairs of keys – triads, or greater linkages for analysis. This effort employs keystroke dyads and four critical features derived from keystroke press and release timings: key hold time (KHT) for each key, key interval time (KIT), key press latency (KPL), and key release latency (KRL). These features can all be derived from the key press and key release timings, as demonstrated in Figure 45, below.

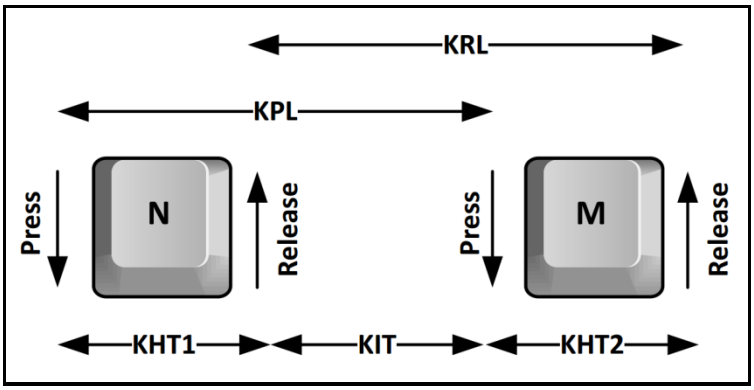


Figure 45: Keystroke Timing Features

5.2.2.3 Mouse Dynamics

As with keystroke dynamics, measuring mouse dynamics has been used for user identification and identity verification. For identification using implicit methods, researchers have examined a number of features of mouse dynamics. Pusara and Brodley [102] built models of user mouse movements, collecting and categorizing data in the hierarchy shown in Figure 46, where *Non-Client Movement* refers to movement within an applications title and menu bars. Data points were collected every 100 milliseconds, and distance, angle, and speed were calculated for selected pairs of points within temporal windows of data. After a training period in which models of user behavioral patterns were built, anomaly detection was used to detect users different than the originally authenticated user.

Schulz [103] examined features of curves within mouse movement such as curve length, number of points within curves, curvature area, and inflection points, and, for each user, computed a histogram of the user's typical mouse movement curves. Ahmed and Traore [104] used a similar histogram-based technique, but based on a different set of four characteristics: mouse movement; drag and drop; point and click; and silence (non-movement). From these characteristics, they calculated the traveled distance (ratio of distance traveled for different types of action), action type (relative frequency of different action types), movement direction (ratio of actions performed in given directions), average movement speed, movement speed versus travelled distance, and time elapsed during movement.

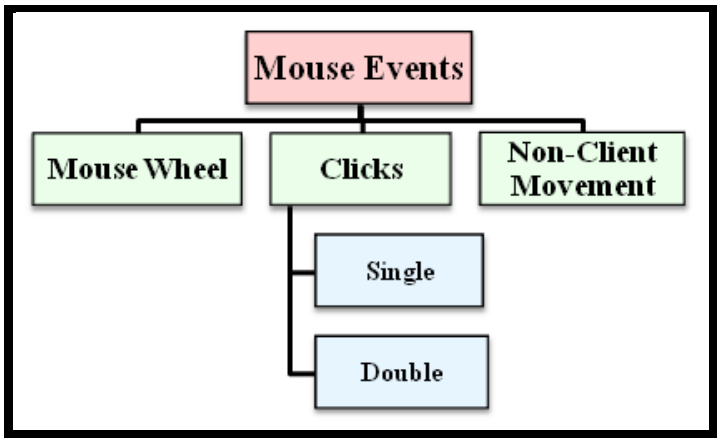


Figure 46: Pusara and Brodley (2004) mouse event hierarchy

Feher et al. [105] used a non-histogram based approach that focuses instead on individual mouse actions. Their hierarchy of mouse action recognizes a number of primitive (level-0) mouse

movement (m), and button-up and button-down events for the left and right buttons (lu, ld, ru, and rd). Higher-level (level-1) events are recognized as sequences of these lower-level events (when the lower-level events occur within a certain timeframe, which may differ for each type of higher level event). In turn, level-2 events are recognized as sequences of level-1 events, and level-3 events as sequences of level-2 events. Characteristics of mouse events, such as trajectory center of mass and third or fourth moments, can be extracted, and it is for these features that classifiers are finally constructed. These kinds of features are typical of those commonly used for analysis using mouse dynamics. As such, these existing features will be used rather than attempting to generate new mouse features.

5.2.2.4 Computer Network Operations

AIS has developed numerous D5 effects on various efforts, including Deny and Disrupt[6]. More information is available at a higher classification level, upon request. These effects target computer systems and operators alike. As such, many of these effects will be used to induce suspicion in operators.

5.2.2.4.1 IntroVirt

IntroVirt is an AIS internal research and development effort that generated an Introspective Hypervisor and Toolstack that provides real-time, event-based monitoring and analysis of virtual machines. IntroVirt provides operators with the ability to monitor and manipulate the critical resources of a virtual machine, including, but not limited to, file, network, registry, driver, and memory[7]. AIS planned to use IntroVirt as a platform to deliver D5 effects and capture subject responses in the proposed experimental design for the Thrust 5b experiment.

5.2.3 TECHNICAL APPROACH

CTS focused primarily on basic research, developing sensors to assist in the collection of data. Each thrust consists of three basic phases: (1) sensor development, (2) experimentation, and (3) analysis of experimental data to validate or invalidate hypotheses. The first thrust focused on suspicion detection, while the second sought to attribute suspicion, and the third sought to manipulate suspicion. Further details on such technology, the experimental procedures, and the analysis process are available in the following section.

5.3 METHODS, ASSUMPTIONS, AND PROCEDURES

This section describes the research and development methods, necessary assumptions, and engineering procedures that were used throughout the effort. Each thrust consisted of research, development, and validation. Data collection sensors were developed for the experiments within the thrust, and data processors were used in the analysis of the experimental data to validate or invalidate the hypotheses. All of the thrusts used a shared database for data collection and processing.

5.3.1 ARCHITECTURE

Four major components were developed under this effort: cyber sensors, database, data processors, and analysis tools. The database is the storage mechanism for the entire system, and the intermediary between the sensors, data processors, and analysis tools. The components interact as depicted in Figure 47. The sensors log information about user activity on the system and store it in the database. The data processors read the data stored by the sensors and extract features. The

[6] Assured Information Security Inc., "Deny and Disrupt Final Technical Report," 2010.
[7] Assured Information Security Inc., "IntroVirt Final Technical Report," 2013.

extracted features are written back to the database and later read by the analysis tools, which perform statistical analysis and generate plots.

Each component plays an important role in monitoring suspicion. The effort produced four cyber sensors capable of collecting information about the user's activity on the system, as shown in Table 11. While each sensor can provide useful information for detecting suspicion, the combined results provide a more accurate picture. Each sensor is independent of the others, with the exception of the gaze tracking system, which is dependent upon the mouse logger. As the sensors are independent, sensors can be easily added and removed.

The database is, likewise, designed to be adaptable. As illustrated in

Figure *48*, each sensor has a dedicated table managed by a master table, which provides information about the events recorded by all of the sensors. Additionally, the database stores the features extracted from the sensor data by the data processors. Each data processor also has a table dedicated to its output. These tables are not linked by the master table; any tool requiring this data accesses it directly.

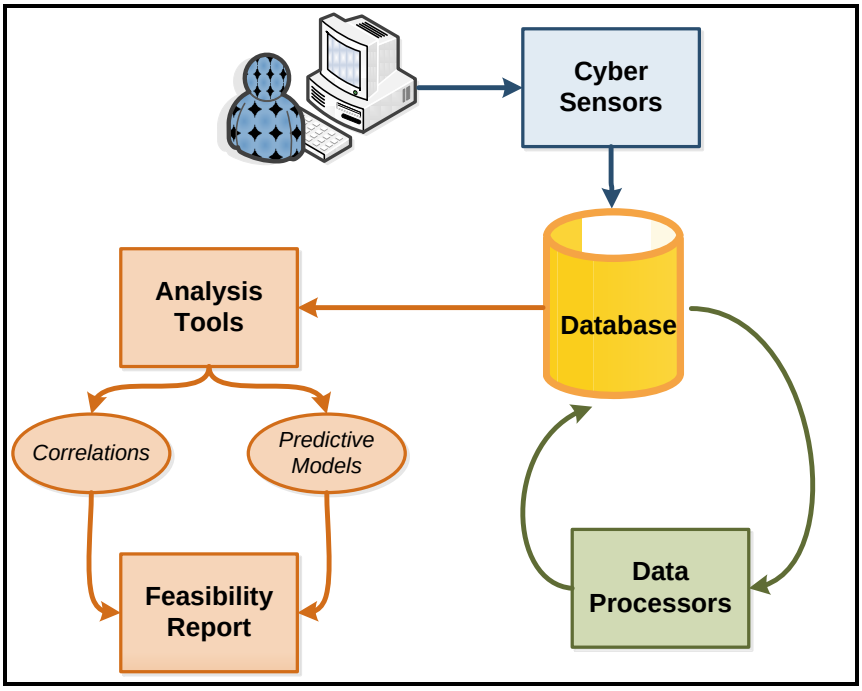


Figure 47: Cyber Trust and Suspicion Components

Table 11: Suspicion Detection Sensors

Sensor	Description
Keylogger	Logs user keystrokes.
Mouse Logger	Logs user mouse movements, clicks, and scroll wheel actions.
Context Logger	Logs the active window, application launches, application closes, and user actions within supported applications. Actions include menu options clicked and the window that the cursor is over.
Gaze Tracker	Logs the on-screen coordinates of where the user is looking.

There is a single data processor for each sensor. These data processors are dedicated to processing the unique data that is output by each sensor and extracting particular features from that data. For example, the keylogger's data processor outputs information about the duration of a key press and the time between presses, among other features. Each of these outputs is a feature that is stored in the database after the data processor has finished running. Analysis is then performed on the processed data using a variety of analysis tools to determine which features are significantly different during the intervals in which suspicion occurred.

5.3.2 DATABASE

The database is responsible for storing all of the information collected and processed. It is designed to be adaptable, accommodating any number of sensors simultaneously. Each sensor has a set of tables, which link to the master table. The master table contains a record of all of the *events* (e.g., key press, mouse movement) that occurred across all sensors and a timestamp for each of these events. Likewise, each data processor has a set of tables which it uses to store its data. For example, Figure 49 demonstrates how the mouse event tables link to the master table.

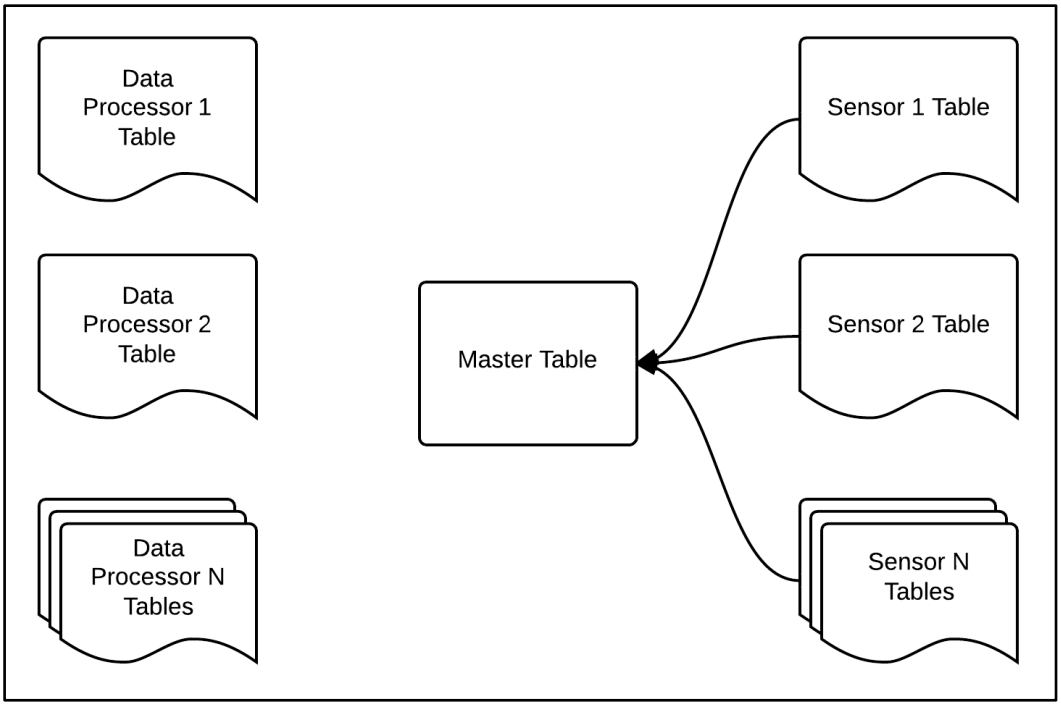


Figure 48: Database Structure

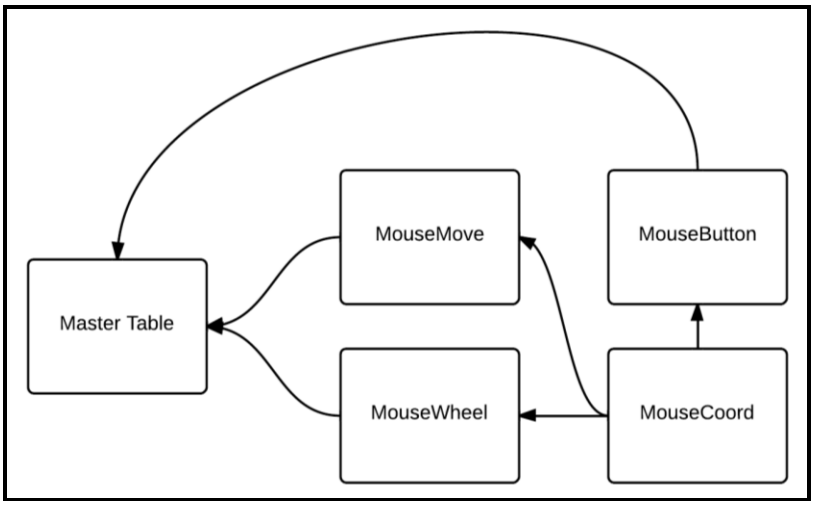


Figure 49: An Example Sensor-Database Relationship

5.3.3 THRUST 5A: SUSPICION DETECTION

The first thrust sought to determine if cyber sensors were viable methods for detecting suspicion. As such, two sensors were developed and each was tested independently via experiments conducted at SU. The resulting data was processed to extract features on which analysis was performed to produce a feasibility report on the use of cyber sensors for detecting suspicion.

5.3.3.1 Sensors

AIS developed two sensors, a keylogger and a mouse logger, in Thrust 5a to investigate their utility in detecting suspicion. These sensors were chosen because of their prevalence in the realm of behavioral biometrics for authentication and verification [100].

5.3.3.1.1 Keylogger

The keylogger logs the key presses and releases from any keyboard attached to the system and records the time at which the key event occurred. It is built for Windows XP and newer Windows operating systems (OSs), and utilizes the raw input scheme supplied by the Windows OS to receive notifications directly from the keyboard when its state is modified (Figure 50).

The keylogger is an invisible message-only window built in Microsoft Visual C++ 2008. When it is run, it is undetectable to the normal human operator; however, it will appear in the task list. The keylogger utilizes MySQL Connector/C++ to interface with the MySQL database[8] and store the data collected. When key codes are received by the keylogger, they are translated to their corresponding key character and case. After translation, the keylogger stores the keystroke data in the database.

5.3.3.1.2 Mouse Logger

The mouse logger records user interaction with any mouse device connected to the system. It is capable of recording user actions for up to eight different mouse buttons, as well as mouse movement and scroll wheel actions. For each of these actions, the mouse logger records a timestamp. The mouse logger is built for Windows platforms version Windows 7 using Microsoft Visual C++ 2008 and MySQL Connector/C++. It runs as an invisible message-only window and utilizes the

[8] http://www.mysql.com/

Windows raw input scheme (Figure 50) to capture mouse data whenever the state of the mouse is modified. Modifications of mouse state are shown in Table 12.

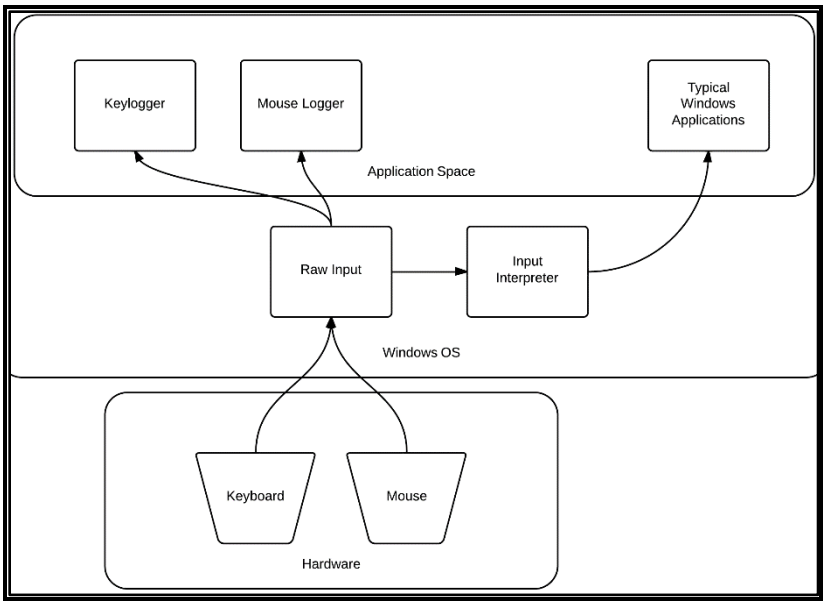


Figure 50: Windows Raw Input Scheme

Table 12: Mouse State Modifiers

State Change	Modification Condition
Mouse Button Up	The user releases a mouse button
Mouse Button Down	The user presses a mouse button
Scroll Down	The user scrolls the mouse wheel toward them
Scroll Up	The user scrolls the mouse wheel away from them
Mouse Movement	The user moves the mouse

The mouse logger records two different sets of mouse coordinates: relative coordinates and physical coordinates. Relative coordinates are relative to the physical position of the mouse on the desk and are much more precise, as they can change even when the mouse does not move on the screen. Relative coordinates also do not factor in mouse pointer acceleration, a mechanism that Windows uses in order to make pointer movements across the screen appear more fluid and natural to a user. The physical coordinates recorded by the mouse logger are the coordinates of the mouse pointer on the screen, in pixels, where the origin is in the upper left of the main monitor. The X direction increases in value as the mouse pointer moves right, and the Y direction increases in value as the mouse pointer moves down. In a multi-monitor system, if the main monitor is to the right of another monitor, it is possible to have negative coordinates. The physical coordinates factor in pointer acceleration, unlike the relative coordinates.

5.3.3.2 Experimentation

Thrust 5a consisted of two experiments, the Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles (RESCHU) Study to collect mouse data, and the Winter Survival Study (WSS) to collect keystroke data. Both studies were designed by and conducted at SU with support from AIS.

5.3.3.2.1 RESCHU Study

The RESCHU Study focused primarily on mouse movement and, as such, serves as the data source for the mouse dynamics analysis. The study consisted of eight users. The experiment attempted to induce suspicion by removing key parts of the interface during pseudo-timed intervals, both while the subject was and was not looking at the specific areas on the interface. However, users rarely reported being suspicious during these intervals, such that no conclusive analysis could be performed on the relationship between mouse features and suspicion. As a result, the analysis performed for the RESCHU Study focused on determining if there were changes when manipulations occurred.

5.3.3.2.2 Winter Survival Study

The WSS included 20 users. As subjects only interacted via Skype Instant Messenger[9], this experiment was selected as the primary data source for collecting and analyzing keystroke data. During the WSS, subjects were asked to prioritize a list of items based on their utility in surviving in a winter environment after conversing with another subject through Skype. After a control trial, some subjects were asked to sabotage the lists of others during these conversations, while subjects who were not selected to be saboteurs were informed that they may be interacting with saboteurs.

5.3.3.3 Data Processors

The data processors are responsible for extracting features from the data provided by the sensors. These features are later used in analysis to determine which features provide the most information for suspicion detection.

5.3.3.3.1 Preprocessing Scripts

In some cases, sensor data must be normalized and outliers must be removed prior to processing. These scripts handle translating the timestamps gathered by the sensors to Unix time, as well as adding the suspicion information from the ground-truth measure.

5.3.3.3.2 Keyboard Data Processor

The keyboard data processor is responsible for taking the data output by the keylogger and extracting features. The keys are paired to form dyads, after which the four features are extracted (Table 13): KHT, KIT, KPL, and KRL. The keyboard data processor is built using Microsoft Visual C++ 2008 and MySQL Connector/C++. It is compatible with Windows XP, Vista, and 7.

Table 13: Keyboard Data Features

Feature	Description
KHT	Time between a press and release. This is calculated for each key in the dyad.
KIT	Time between the release of the first key and the press of the second key.
KPL	Time between the presses for each key in the dyad.
KRL	Time between the releases for each key in the dyad.

5.3.3.3.3 Mouse Data Processor

The mouse data processor is responsible for extracting features from the data output by the mouse logger. It is compatible with Windows XP and newer OSs, and is built using Microsoft Visual C++ and MySQL Connector/C++. The features that the mouse data processor extracts are adaptations of features used by user verification experiments using the mouse. These features are shown in Table 14.

[9] http://www.skype.com/en/

5.3.3.4 Analysis

AIS performed various types of analysis on the keystroke and mouse data collected to determine their feasibility for suspicion detection. The analysis was performed on the processed data from each experiment. AIS performed five key statistical tests on each dataset, as shown in Table 15: Principal Component Analysis (PCA), Logistic Regression, Rank Sum, Weighted Feature Analysis, and Feature Plotting. In addition to these tests, Dr. Hirshfield used the keystroke data to perform further machine learning at SU.

Table 14: Mouse Data Features

Feature	Description
Silence	A period of inactivity lasting at least 300 milliseconds
Movement Curve	Any number of consecutive movement coordinates not separated by a silence interval
Left Click	A left mouse button down and left mouse button up event occurring within 500 milliseconds of one another
Right Click	A right mouse button down and right mouse button up event occurring within 500 milliseconds of one another
Double Click	Two left clicks or two right clicks occurring within 500 milliseconds of each other
Drag and Drop	A left or right mouse button down, a movement, and left or right mouse button up (respective to the mouse button down event) occurring more than 500 milliseconds apart
Mouse Move + Drag and Drop	A movement curve preceding a drag and drop, without silence between them
Mouse Move + Click	A movement curve preceding either a right or left click, without silence between them
Mouse Move + Double Click	A movement curve preceding a double click, without silence between them

PCA provides the foundation by revealing which features have the most variance within the dataset. This may drive model reduction for future tests to reduce statistical noise. After PCA, AIS performed linear modeling with logistic regression to determine how a simple model would perform. Logistic regression is useful for datasets where the predictor variable is categorical, not continuous. From the simple models in logistic regression, AIS performed hypothesis testing to determine if there was a statistically significant difference between one population or another, in this case between a user's actions while suspicious vs. non-suspicious. Rank sum hypothesis testing was used because it can be used on both normal and non-normal datasets. Weighted feature analysis was performed only for the mouse data to assess its complexity and to determine if this complexity could be correlated with suspicion. Finally, feature plotting was used to visualize trends and identify areas for further analysis.

Table 15: Analysis Tests

Analysis	Function
Principal Component Analysis	Determines the combination of features providing the most variance to the data set.
Logistic Regression	A machine learning algorithm used to predict suspicion state from a set of features.
Rank Sum	Statistical hypothesis testing used to determine correlation between two features.
Weighted Feature Analysis	Creates plots of specified features, giving each feature a specified "weight." These weights are summed for each record, and the resultant sums are graphed.
Feature Plotting	Plots a single feature set vs its timestamps (e.g., plotting clicks vs time).

5.3.4 THRUST 5B: SUSPICION ATTRIBUTION

The second thrust sought to determine if the source of suspicion, both as perceived by the user and the actual cause of suspicion, could be determine using cyber sensors. As such, three sensors were developed and an experiment was designed to collect data with the sensors at SU. Due to the early termination of the effort, the experiment and analysis was not performed.

5.3.4.1 Sensors

AIS developed two sensors in Thrust 5b, a context logger and a gaze tracker, to investigate their utility in attributing suspicion. These sensors were selected because they provide additional information on user behavior when suspicious.

5.3.4.1.1 Context Logger

The context logger began as an application logger that monitors the system for a change in the foreground window. If the current foreground window is replaced by a different window, the new window is logged with a timestamp of the time the change occurred. The application logger accomplishes this logging using global hooks and DLL injection to monitor every window on the system. If a hooked window receives a message telling it that it is the new foreground window, the hook procedure sends information about this window to the application logger, which logs this information in the database. Figure 51 depicts this process.

Advanced functions were then added to the application logger to form the context logger. In addition to performing all of the application logging functions, it monitors user actions within each running program and reports a variety of actions, listed in Table 16. This monitoring is achieved using global hooks and DLL injection to intercept the window messages sent to each window on the system. When a window message of interest is received, information about the action indicated by the window message, as well as the state of the window, is sent to the application activity logger, which writes this information to the database.

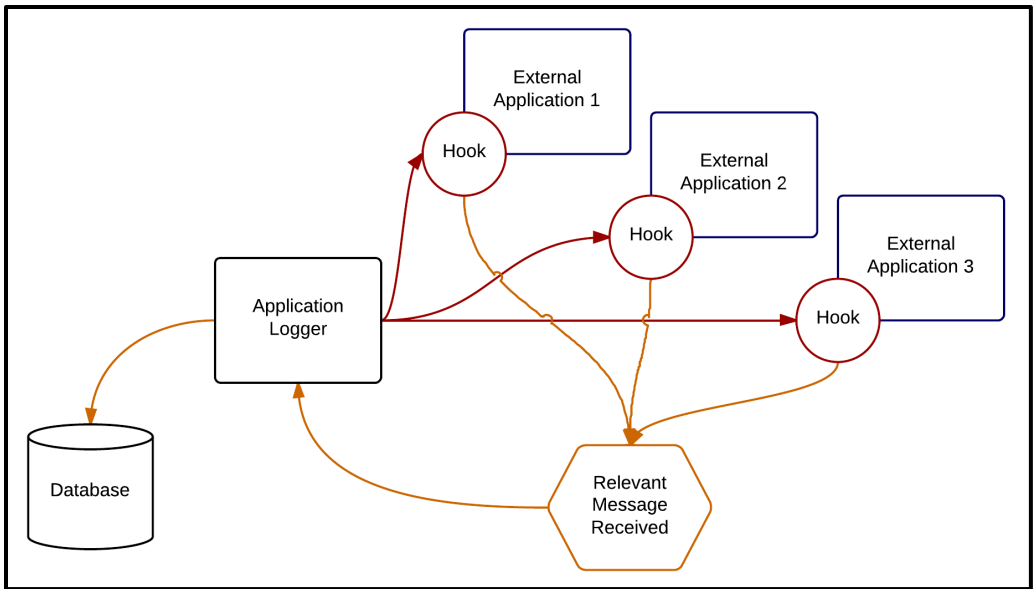


Figure 51: Application Logger Program Flow

5.3.4.1.2 Gaze Tracker

Using any typical webcam supported by Windows XP and newer OSs, the gaze tracker is able to track the eye movement and on-screen gaze coordinates of a user. The gaze tracker is built using Microsoft Visual C# 2008 and MySQL Connector/C#. It was adapted from the open-source project, IT University of Copenhagen Gaze Tracker[10].

Table 16: Logged Actions

Action	Information Collected
Application Launched	The name application that was opened, its class, executable path, window message, and process ID.
Application Closed	The name application that was closed, its class, executable path, window message, and process ID.
Menu item highlighted	The text of the highlighted menu item.
Cursor in menu bar	The part of the menu bar the mouse is over (e.g., the minimize button, the close button, blank space, etc.)
Cursor over window	The name of the widow that the cursor is over (this window does not necessarily have the mouse focus).
Window is activated	The name of the window being activated.
Window is minimized	The name of the window being minimized.
Window is maximized	The name of the window being maximized.
Window is restored	The name of the window being restored.
Window is resized	The name of the window being resized.
Window is moved	The name of the window being moved.

The gaze tracker is a powerful sensor that is augmented by the addition of other sensors. Alone, the gaze tracker is able to track fixations and saccades in a user's eye movements. Paired with the mouse logger, it is able to estimate the on-screen coordinates of a user's gaze. When paired with the mouse logger and the context logger, the gaze tracker can identify the specific application that a user is currently viewing.

When paired with the mouse logger, the gaze tracker is able to calibrate a user's gaze using the fixations that occur when a user clicks the mouse button. To support this, the mouse logger reports the position of the mouse click and the gaze tracker creates a calibration target for that position. Once enough targets are created, the gaze tracker calibrates and begins reporting the gaze coordinates of the user. The gaze tracker continues calibrating after the initial calibration, adding new calibration targets for each subsequent mouse click to become more accurate the longer that it is run.

5.3.4.2 Experimentation

AIS developed an experimental design, in collaboration with SU, for the attribution experiment, as follows:

[10] http://www.gazegroup.org/

The subjects will be split into three groups; one group will perform a typing task, another will perform a mouse task, and the final group will perform a combined mouse and typing task. Each group will have four trials, two of which are control and two of which are experimental during which D5 effects (e.g., drop keystroke, invert mouse movements) are launched against them. AIS will capture the subject's behavioral data using the detection and attribution sensors within IntroVirt. For this experiment, the sensors will capture: (1) keystroke timings, (2) mouse movements, (3) application launches, exits, and focus changes, (4) menu choices, (5) eye movement, and (6) system calls. SU will provide ground-truth data for the experiment using physiological sensors, such as fNIRS, EEG, GSR, and video capture (Figure 52).

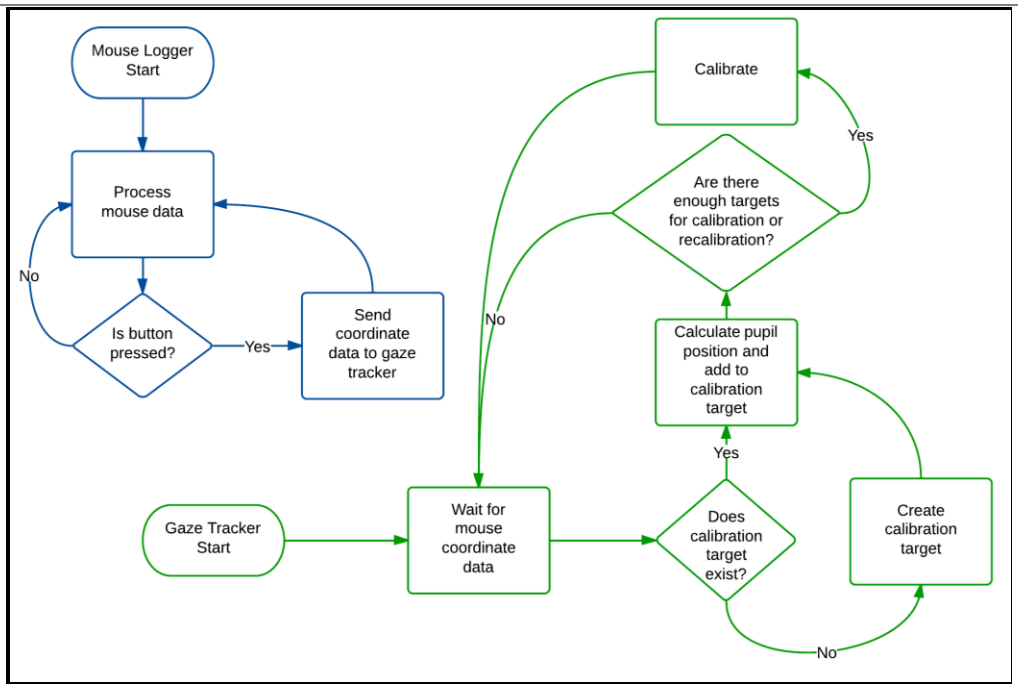


Figure 52: Gaze Tracker Calibration Process

5.3.4.2.1 Experimental Overview

Hypothesis: Cyber sensors can be used to determine the reason that a subject becomes suspicious (perceived or actual).
Independent Variable: D5 Effects (to be launched against subjects)
Dependent Variable: Suspicion
Scope: This experiment will use a small subset of D5 effects (5) from three of the D5 categories (deny, disrupt, and deceive) to induce suspicion. This experiment will seek to determine if the sensor data can be used to differentiate between the effect or class of effect based on the cyber and physiological sensors.
Sensors: Keylogger, Mouse Logger, Application Logger, Context Logger, Graze Tracker, System Call Monitor, EEG, fNIRS, GSR, Surveys (Post-Experiment)
Analysis: Find all, if any, statistically significant correlations between the sensor data and the class of D5 effect or specific D5 effect.
Number of Subjects: $\geq$ 24 (at least eight per group for three groups).
Restrictions/Limitations:

- No subjects can participate in multiple groups, as they may identify a pattern relating to when the D5 effects occurred, potentially biasing the results.
- No graduate students that have worked on the CTS effort (analysis, design, or experimental setup) can participate, as knowing or suspecting that we are trying to induce suspicion could bias results.

- Subjects cannot be informed that we are trying to induce suspicion until after all trials for the experiment to avoid biasing the results.
- No surveys shall be given until all four trials have been completed, so as to not bias users that we are looking to induce suspicion. The (neuro-) physiological sensors will provide ground-truth.
- There must be a break between Trials 2 and 3. This break may span 10 minutes or the session may be split across two days. Subjects will have D5 effects launched against them in Trial 2. The experimenter will tell all subjects the following:
    - *"Some of you experienced some problems with the experimentation environment during the last trial. We're bringing in the support staff to take a look at the machines and should have them running properly shortly."*
    - When subjects return for Trial 3, they will be told: *"There was a problem with the experimentation environment. They were able to fix it, but let us know if you experience any more issues."*
- The clocks must be synched for all physiological sensors and all computers used during the experiment. The IntroVirt platform should handle this for all computers.

5.3.4.2.2 Attribution Experiment

The set of experiments for this effort will find three groups of subjects participating in different tasks, where group one will be performing a basic Internet searching task (combined task), group two will be given a transcription task (typing task), and group three will be asked to fill out a digital survey (mouse task). All groups will perform a practice trial to establish a baseline, after which subjects will have D5 attacks launched against them during Trials 2 and 4. Each task is expected to take between 10 and 20 minutes. Each group will complete four trials. Subjects will be asked to fill out different surveys, transcribe different texts, and search for different items in different trials to avoid becoming too repetitive and losing subject focus. A tentative schedule is displayed in Table 17.

Table 17: Tentative Experiment Schedule

Trial	Experimental
1	Practice
2	D5
Break	
3	Normal
4	D5

Participants will be incentivized to complete their task as quickly as possible with additional bonuses dependent on the speed with which their task is completed. During each session, the participant will be placed in the small experiment room. The administrator will leave the subjects unattended for the duration of the task and return upon the conclusion of the task. The administrator will inform the subjects of his or her location should a major problem arise. Any subject who requests help or reports a problem with the computer will be informed that technical support will examine and resolve the issue prior to the next session (Table 18).

Table 18: Tentative D5 Effects

Type	Effect
Deceive	Fake File Save
	Application Error Message
	Redirect Host
Deny	Lock Screen

Disrupt	Drop Keystrokes
	Invert Mouse Movements

Group 1: Internet Search Task

- **Trial 1:** Subjects will be asked to find places to purchase a list of items and note the best price for each item.
- **Trial 2:** Subjects will be asked to find places to purchase a list of items and note the best price for each item. Some of the items will exist only on pages created specifically for this task (e.g., fake brand name). One or more of these pages will be setup with a redirect sending the user back to Google or another common website.
- **Trial 3:** Subjects will be given a large list of items (broken into two groups) and asked to find places to purchase the list of items and note the best price for each item. Subjects will be asked to save this information in two separate text files (one for each group) and then aggregate the information at the end of the trial.
- **Trial 4:** Subjects will be given a large list of items and asked to find places to purchase the list of items. Subjects will be asked to save this information in two separate text files (one for each group) and then aggregate the information at the end of the trial. Throughout the trial, the Fake File Save effect will prevent a user from saving a file, but tell the user that the file save was successful.

D5 Effects: Fake a file save, Redirect Host

Group 2: Typing Task

- **Trial 1:** Subjects will be asked to transcribe a written document that is several pages long. They will be asked to proofread each page (upon completion of each page) to ensure accuracy.
- **Trial 2:** Subjects will be asked to transcribe a different written document. They will be asked to proof-read each page (upon completion of each page) to ensure accuracy. Some keystrokes will be dropped (every 12th keystroke) for the duration of the task.
- **Trial 3:** Subjects will be asked to transcribe an online article. They will be asked to proof-read each page (upon page completion) to ensure accuracy. During this trial, the clipboard will be disabled, so that the subjects cannot copy and paste the article.
- **Trial 4:** Subjects will be asked to transcribe an online article. They will be asked to proof-read each page (upon page completion) to ensure accuracy. During this trial, the clipboard will be disabled, so that the subjects cannot copy and paste the article. The subject's will experience an error message that will pop up three times throughout the trial.

D5 Effects: Drop Keystrokes, Application Error Message

Group 3: Survey Task

- **Trial 1:** Subjects will be asked to complete a survey on their life experiences.
- **Trial 2:** Subjects will be asked to complete a series of surveys on their personality features. At some point midway through the task, the subject's mouse movements will be inverted for 75 seconds. This will occur at two separate points throughout the second half of the trial.
- **Trial 3:** Subjects will be asked to complete a survey on their educational experiences.

- **Trial 4:** Subjects will be asked to complete a survey on their experiences during the study. The subject's screen will be locked up to three times, starting approximately halfway through the trial (10 minutes in).

Mouse-based D5 Effects: Invert Mouse Movements, Lock Screen

5.3.4.3 DATA PROCESSORS

The data processors are responsible for extracting features from the data provided by the sensors. These features would later be used in analysis to determine which features provide the most information for suspicion attribution. The data processors were not fully developed due to the early termination of the contract.

5.3.4.3.1 Preprocessing Scripts

In some cases, sensor data must be normalized and outliers must be removed prior to processing. These scripts handle translating the timestamps gathered by the sensors to Unix time, as well as adding the suspicion information from the ground truth measure.

5.3.4.3.2 Context Data Processor

The context data processor is responsible for taking the data output by the context logger and extracting features. Potential features include the application class, both as listed by the Windows OS and as determined manually (e.g., Skype is an Instant Messenger, whereas Microsoft Word is a word processor), action taken (e.g., application opened, closed, moved, resized), window coordinates, and menu actions (e.g., menu opened, menu item clicked).

5.3.4.3.3 Gaze Data Processor

The gaze data processor is responsible for taking the data output by the gaze detection sensor and extracting features. Potential features include gaze coordinates and eye movement type (i.e., fixation, saccade).

5.3.4.3.4 System Call Data Processor

The system call data processor is responsible for taking the data output by the IntroVirt system call logger and extracting features. The primary objective of this data processor is to reduce the number of system calls collected to a manageable and useful dataset. For example, one application launch can trigger numerous system calls; if, on average, there were 10 system calls per second that would total 12,000 system calls in a 20-minute session. The selected features would then be grouped by function or class, similar to applications to indicate their utility.

5.3.4.4 ANALYSIS

The analysis performed in Thrust 5b would have mirrored much of the analysis performed in Thrust 5a for the new sensors and the suspicion detection sensors, but would focus primarily on PCA, logistic or linear regression (for categorical and continuous variables, respectively), hypothesis testing, and feature plotting. The experiment and subsequent analysis were not performed due to the early termination of the contract.

5.3.5 THRUST 5C: SUSPICION MANIPULATION

The third thrust sought to determine if the cyber tools could be used to manipulate suspicion and other mental states. This thrust was designed to provide the groundwork for suspicion manipulation research, but not consist of substantial development or experimentation due to the

scope of the effort. No work was performed on the Thrust 5c tasking due to the early termination of the contract.

5.3.5.1 Sensors

No sensors were slated for development in Thrust 5c, though Thrust 5a and Thrust 5b sensors were scheduled for optimization based on the analysis performed during the thrusts.

5.3.5.2 Operator Mapping

The primary goal of Thrust 5c was to develop an *Operator Mapping* that provided a link between D5 effects, cognitive states, and anticipated operator behavior. Initially, each effect (e.g., adjust flicker rate/resolution, random mouse movements, close/minimize windows, turn off networking, redirect websites) would be associated with a set of anticipated psychological states (e.g., suspicious, frustrated, surprised/confused) and behavioral responses (e.g. examine network connections, attempt to restart adapter, check display settings, reboot, run virus scan). This mapping would then be tested and optimized via an experiment at SU involving cyber D5 effects. A nominal set of effects, anticipated cognitive states, and anticipated user responses are displayed in Table 19.

5.3.5.3 Experimentation

The fourth CTS experiment to be conducted at SU was designed to use all of the sensors developed throughout the effort to collect additional data on additional D5 effects to measure subject responses and cognitive states (via physiological sensors and surveys). From this data, the Operator Mapping was to be optimized.

5.3.5.4 Analysis

The analysis scheduled for the third thrust would lift many of the procedures from the first two thrusts, but focus on identifying correlations between: subject behavior and cognitive states (including more states than just suspicion), cognitive states and sensor data, and subject behavior and sensor data. This analysis would be used to update the Operator Mapping.

Table 19: Nominal Anticipated Actions by Effects

Type	Effect	Cognitive State	Actions
Deceive	Change System Time	No Change	Operate based on incorrect time
		Confusion	Attempt to fix system time
		Fear	Run anti-virus/spyware program
	Change IP Address	No Change	Ignore
		Surprise → Trust	Release and renew IP
			Statically set IP
			Repair network adapter
		Fear and/or Suspicion	Disable network
			Run anti-virus/spyware program
Deny	Box Mouse Movements	Trust	Check USB/PS2 connection
			Switch USB ports
			Try alternate mouse
			Reboot terminal
		Frustration	Use keyboard
	Change User Privileges	No Change	Ignore (task dependent)
		Trust	Attempt to change privileges
			Circumvent privileges
			Reboot terminal
		Fear and/or Suspicion	Run anti-virus/spyware program
	Website Redirects	No Change	Ignore
		Suspicion	Clear cache/temporary files
		Fear and/or Suspicion	Run anti-virus/spyware program
Disrupt	Insert/Drop Keystrokes	No Change	Ignore
		Trust	Try alternate keyboard
			Reboot terminal
		Suspicion	Examine active processes
		Fear and/or Suspicion	Run anti-virus/spyware program
	Flicker Screen	No Change	Ignore
		Trust	Check display settings
			Reboot terminal
		Fear and/or Suspicion	Run anti-virus/spyware program

5.4 RESULTS AND DISCUSSION

This effort has successfully identified a method for detecting suspicion using keystroke dynamics. In addition, AIS has developed several other sensors and an experimental design that can be used to evaluate these sensors for use in suspicion attribution. The results of this effort should spawn future research and development efforts to expand upon and leverage these findings to provide tools and situational awareness to CNO operators and mission planners.

5.4.1 LIMITATIONS

Due to the nature of the effort, certain assumptions were made in order to simplify the problem space. The following limitations arose from those assumptions, the scope of the effort, and other unforeseen changes:

- Cyber sensors were developed for Windows OS. Some are for Windows Vista/7/8, whereas others also support Windows XP; these sensors will not work on Linux OSs or other versions of Windows.
- Due to other obligations under the Basic Research Initiative (BRI), SU was unable to perform the experiments and provide all of the experimental data at the originally scheduled times, pushing some of the analysis to much later in the effort.
- The analysis of the data collected by the mouse logger proved inconclusive. AIS planned to perform additional analysis on the mouse data during the attribution experiment to determine if the mouse features could provide information on suspicion detection.
- Due to the scope of the effort, the number of subjects who participated in the experiments was sub-optimal. The analysis still provided statistically significant results, but larger studies are recommended for future research.
- The program was terminated early due to the departure of the primary principal investigator from the prime organization. As a result, Thrust 5b was not completed and Thrust 5c was not started.

5.4.2 RESULTS

During the first thrust, AIS found a statistically significant correlation between KIT and suspicion, indicating that keystroke data is a viable method for measuring suspicion. The detailed analysis results are provided in this section. Two studies were conducted for the first thrust, the RESCHU Study and the WSS, while each other thrust had a single experiment scheduled.

5.4.2.1 RESCHU Study (Thrust 5a)

The RESCHU Study focused primarily on mouse movement and, as such, serves as the data source for the mouse dynamics analysis. The study consisted of eight users. The experiment attempted to induce suspicion by removing key parts of the interface during pseudo-timed intervals, both while the subject was and was not looking at the specific areas on the interface. However, users rarely reported being suspicious during these intervals, such that no conclusive analysis could be performed on the relationship between mouse features and suspicion. As a result, the analysis performed for the RESCHU Study focused on determining if there were changes when manipulations occurred.

5.4.2.1.1 Feature Derivation

The mouse logger collected the following data points: mouse movements, button clicks, and scroll wheel actions. AIS used results from various other mouse dynamics studies to select the features to be analyzed using the mouse data processor (Section 5.3.3.3). When extracted, these features were kept separate according to which study they correspond to, in order to validate the feature sets as a whole. Some analyses therefore ran tests on feature sets, whereas others focused on other subsets of the features. Some features have specialized metrics associated with them, these metrics are detailed in Table 20 [105].

5.4.2.1.2 Suspicion Assignment

For the RESCHU Study, the self-report of suspicion conducted at the conclusion of the experiment was used for ground-truth. Unfortunately, few subjects reported that they had become suspicious. Additionally, self-report data was collected at the conclusion of the trial, and did not provide the exact time at which trust began to vary. As such, AIS refocused the analysis to look for differences in the data between when a subject was expected to be suspicious (after a manipulation occurred) and when a subject was expected to be trusting (during control trials). Additional analyses were also performed on the survey results, to rigorously test the survey's efficacy as a ground-truth. To accomplish this, AIS divided data into trust and suspicion populations by labeling all data in control trials and all data in overt and covert trials prior to a manipulation as trust. All other data was labeled with suspicion.

Table 20: Feature Metrics

Metric	Description	Feature
Trajectory Center of Mass (TCM)	A measure of the average time it takes for the mouse pointer to travel a distance	Movement Curve
Scattering Coefficient (SC)	A measure of how much the mouse movement deviates from the TCM	Movement Curve
Third Moment (M3)	A permutation of the duration of the movement curve and the distance traveled	Movement Curve
Fourth Moment (M4)	A permutation of the duration of the movement curve and the distance traveled	Movement Curve
Trajectory Curvature (TCrv)	A measure of the curvature of the movement's trajectory graph	Movement Curve
Velocity Curvature (VCrv)	A measure of the curvature of the movement's velocity graph	Movement Curve
Click Time (CT)	The time between the mouse down and mouse up of a click	Click, Move + Click
Traveled Distance during Click (TDC)	The distance traveled between the mouse down and mouse up of a click	Click, Move + Click
First Click Time (FCT)	The CT of the first click	Double Click, Move + Double Click
First Click Distance (FCD)	The TDC of the first click	Double Click, Move + Double Click
Interval Time (IT)	The time between the two click actions	Double Click, Move + Double Click
Interval Distance (ID)	The distance traveled by the cursor between the two click actions	Double Click
Second Click Time (SCT)	The CT of the second click	Double Click, Move + Double Click
Second Click Distance (SCD)	The TDC of the second click	Double Click, Move + Double Click
Distance to Click (DC)	The distance between the end of the movement curve preceding a click and the click itself	Move + Double Click, Drag and Drop
Time to Click (TC)	The time between the mouse movement preceding a click and the mouse down of a click	Drag and Drop

5.4.2.1.3 Data Pre-Processing

There were few outliers identified for the RESCHU Study. As a result, any outliers (results that were wildly outside the normal) were noted, but included in the analysis. Surveys were conducted between sessions and all data collected while the subjects filled out the survey was removed whenever possible. Due to problems with clock synchronization, there are likely some remaining data points, but they should be insignificant when compared with the sample size of the remaining data.

5.4.2.1.4 Data Analysis

Many statistical and observational algorithms were applied to subsets of the data in order to determine whether the data collected could be used to predict when a user became suspicious. The analysis of mouse features for the RESCHU Study consisted of four key components: PCA, Hypothesis Testing (Rank Sum, Direct Analysis, and Feature Plots), "Byte Analysis", and Logistic Regression

Principal Component Analysis

In order to determine which of the derived features or feature combinations varied the most within the dataset, the derived features were processed by a PCA algorithm. PCA is used to determine which variables within a data set contain the most variance by breaking the data set up into components. Each component can be composed of one or more variables. This is useful because features which contain most of the variance are generally the most explanatory variables. The following tables show the results of running PCA on each derived feature table. The proportion of variance *row in each* Component Statistics *table shows how much variance is captured by the component. Each* Component Composition *table (Table 21 -*

Table 32) shows what variables each component is composed of.

Table 21: Movement Features Component Statistics

	Component 1	Component 2	Component 3	Component 4	Component 5
Std. Dev.	529.2	18.59	$5.01e^{-1}$	$9.60e^{-2}$	$1.71e^{-2}$
Proportion of Variance	0.998	0.001	$8.95e^{-8}$	$1.044e^{-9}$	$2.33e^{-11}$
Cumulative Proportion	0.998	0.999	1.00	1.00	1.00

Table 22: Click Component Statistics

	Component 1	Component 2
Standard Deviation	101.59	$7.60e^{-2}$
Proportion of Variance	0.999	$5.60e^{-7}$
Cumulative Proportion	0.999	1

Table 23: Movement Features Component Composition

	Component 1	Component 2	Component 3	Component 4	Component 5
TCM			– 0.509	0.622	0.170
SC			– 0.477	0.262	– 0.664
M3			– 0.489	– 0.124	0.697
M4			– 0.524	– 0.728	– 0.212
TCrv		– 1.00			
VCrv	– 1.00				

Table 24: Click Component Composition

	Component 1	Component 2
CT		– 1
TDC	1	

Table 25: Double Click Component Statistics

	Component 1	Component 2	Component 3	Component 4	Component 5
Std. Dev.	102.45	58.71	35.01	$1.644e^{-1}$	5.5124
Proportion of Variance	0.691	0.227	0.080	$1.783e^{-6}$	$2.003e^{-7}$
Cumulative Proportion	0.691	0.919	0.999	$9.999e^{-1}$	1

Table 26: Double Click Component Composition

	Component 1	Component 2	Component 3	Component 4	Component 5
FCT					– 0.833
FCD			– 0.998		
IT				– 0.999	
ID	– 0.990	0.128			
SCT					– 0.833
SCD	– 0.128	– 0.992			

Table 27: Mouse Move and Click Component Statistics

	Component 1	Component 2	Component 3	Component 4
Std. Deviation	73.695	41.044	$2.283e^{-1}$	$4.768e^{-2}$
Proportion of Variance	0.763	0.236	$7.326e^{-6}$	$3.195e^{-7}$
Cumulative Proportion	0.763	0.999	$9.999e^{-1}$	1

Table 28: Mouse Move and Click Component Composition

	Component 1	Component 2	Component 3	Component 4
TC			0.998	
DC	– 0.222	0.975		
CT				– 0.998
TDC	– 0.975	– 0.222		

Table 29: Mouse Move and Double Click Component Statistics

	Component 1	Component 2	Component 3	Component 4	Component 5
Std. Dev.	58.432	35.082	14.644	$1.815e^{-1}$	$5.474e^{-2}$
Proportion of Variance	0.702	0.253	0.044	$6.786e^{-6}$	$6.166e^{-7}$
Cumulative Proportion	0.702	0.955	0.999	$9.999e^{-1}$	1

Table 30: Mouse Move and Double Click Component Composition

	Component 1	Component 2	Component 3	Component 4	Component 5
DC		0.116	0.993		– 0.599
FCT					
FCD		0.993	– 0.116		
IT				– 1.000	
SCT					– 0.801
SCD	– 1.000				

Table 31: Mouse Move and Drag and Drop Component Statistics

	Component 1	Component 2
Standard Deviation	50.560	$5.038e^{-1}$
Proportion of Variance	0.999	$9.931e^{-5}$
Cumulative Proportion	0.999	1

Table 32: Mouse Move and Drag and Drop Component Composition

	Component 1	Component 2
TC		-1
DC	1	

Byte Analysis

Further analysis was performed by turning each data vector into a byte. Each bit in the byte corresponds to a single feature, detailed in Table 33. If the feature is "on," the bit has a value of one; otherwise, it is given a value of zero.

Table 33: Bit-Feature Pairings

Bit	Feature
1	X
2	Y
3	Left Mouse Button Down
4	Right Mouse Button Down
5	Scroll Up
6	Scroll Down
7	Middle Button Down

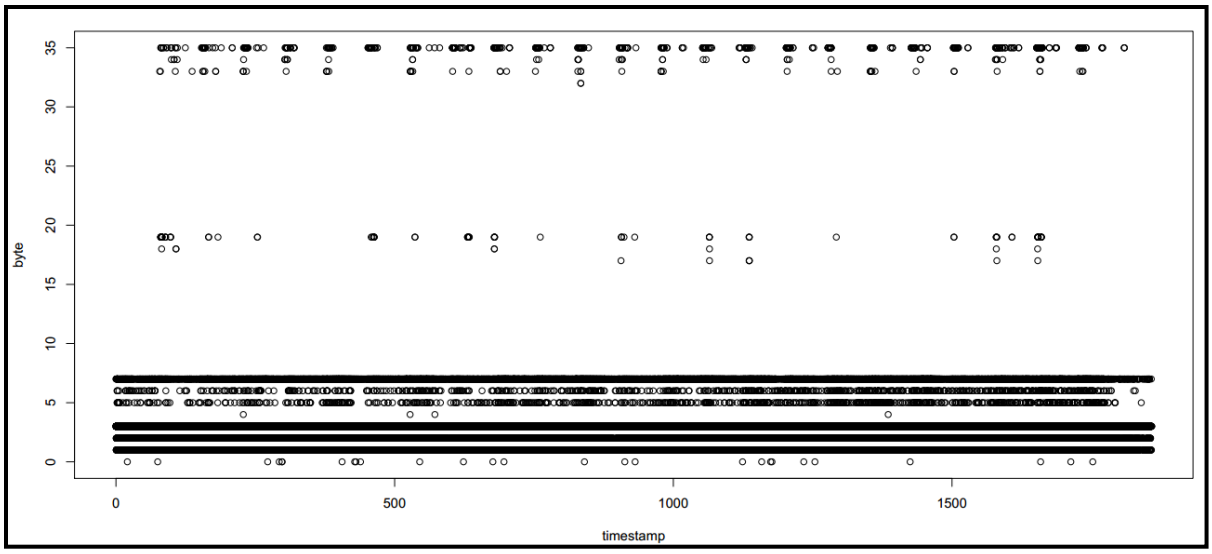


Figure 53: Plot of Bytes from Trusting Intervals

Interpreting these bytes as numeric values allows the observation of unique combinations of features within each data vector. When plotted against time, these "byte" plots revealed that users perform more complicated mouse actions (i.e., combinations of more features) when they are in a trusting state (Figure 53) versus a suspicious state (Figure 54). However, users in the RESCHU Study were in a trusting state more often than they were in a suspicious state, which increases the probability of encountering different types of mouse actions. This inequality is simply a result of the experimental design; there are an equal number of control trials and overt/covert trials, but users were not expected to become suspicious until a manipulation was launched during overt or covert trials. These manipulations were never launched at the beginning of the trial, and therefore some trust data was always produced by a trial that also produced suspicion data.

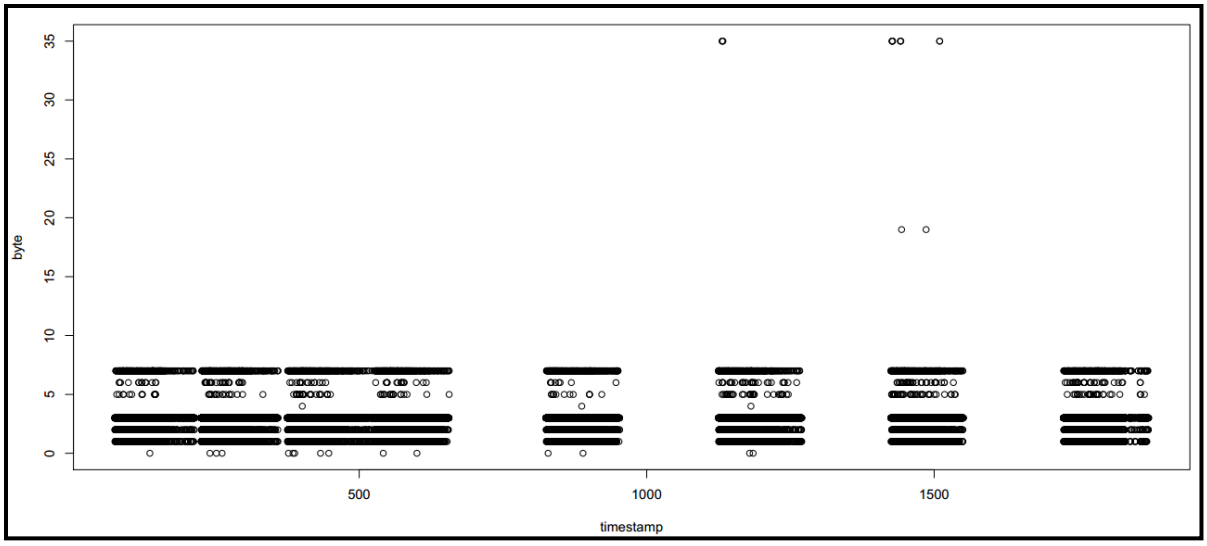


Figure 54 Plot of Bytes from Suspicious Intervals

There were also very few data points that represented complicated mouse actions. Complex mouse actions are represented by higher byte values in the charts. It is possible that these data points

were simply outliers that did not occur during the suspicious intervals. Some of the trusting intervals may contain data from the time that the user spent filling out surveys between trials. These surveys would require more complicated mouse actions (e.g., scrolling the wheel while moving the mouse and left clicking) than the RESCHU testbed. Due to this uncertainty, the next step in this analysis sought to determine if this phenomenon was statistically significant. Table 34, Table 35 and Table 36 provide a more detailed look at the byte plot data, allowing us a view of its composition.

The complicated mouse actions (byte values 17 through 35) make up less than 1% of the data in either the suspicious or trusting intervals. As the data is composed of so few of these byte values, their presence appears insignificant. In order to verify this observation, a two-tailed Rank Sum hypothesis test was performed on the suspicious and trusting datasets of byte values to determine if the two populations differed in a statistically significant way. A p-value of 0.01 was selected for the significance threshold. The results of the test yielded a p-value of 0.48, which is above the threshold; thus, AIS concluded that the two populations do not differ significantly.

Table 34: All Bytes

Byte Value	Number of Occurrences	Percentage of Data
1	25854	11.3
2	22226	9.7
3	164402	71.6
4	7	0.003
5	1038	0.5
6	1024	0.4
7	14907	6.5
17	7	0.003
18	0	0
19	87	0.04
32	2	0.0009
33	57	0.02
34	48	0.02
35	638	0.28

Table 35: Bytes during Suspicion Intervals

Byte Value	Number of Occurrences	Percentage of Data
1	5195	10
2	4401	8.8
3	36778	73.5
4	3	0.006
5	208	0.4
6	181	0.3
7	3230	6.5
17	0	0
18	0	0
19	2	0.004
32	0	0
33	0	0
34	0	0
35	12	0.02

Table 36: Bytes during Trusting Intervals

Byte Value	Number of Occurrences	Percentage of Data
1	20643	11.5
2	17734	9.9
3	127654	71.1
4	4	0.002
5	836	0.46
6	842	0.47
7	11749	6.5
17	7	0.004
18	8	0.004
19	85	0.05
32	2	0.001
33	57	0.03
34	46	0.03
35	628	0.35

Logistic Regression

Logistic regression was used on the raw mouse data to determine if there was a particular mouse state that indicated trust or suspicion. A mouse state is a combination of active features (e.g., right mouse button down, cursor at X = 50, Y = 105, left button up, and no scroll wheel action). The results of the logistic regression (Table 37) show that most features are not predictive, but some have the potential to be. In particular, and in support of the observations of the byte analysis, the results from logistic regression show that the occurrence of scroll up and scroll down actions are negatively correlated with suspicion. Also in accordance with the byte analysis, the standard error for these two features is rather high, mirroring the high p-value found from hypothesis testing on the byte data.

Table 37: Logistic Regression Results (RESCHU)

Feature	Estimate	Standard Error	Z Value	Pr(>|z|)
Right Button Down	$3.799e^{-1}$	$9.813e^{-2}$	3.872	$< 2e^{-16}$
Left Button Down	$-2.858e^{-2}$	$5.212e^{-2}$	– 0.548	0.000108
Scroll Down	$-1.305e^{1}$	$8.770e^{1}$	– 0.149	0.583399
Scroll Up	$-1.304e^{1}$	$3.601e^{2}$	– 0.036	0.881711
X	$5.892e^{-5}$	$4.685e^{-5}$	1.258	0.971112
Y	$-1.006e^{-3}$	$6.333e^{-5}$	– 15.892	0.208551

Hypothesis Testing

Due to the results of logistic regression, the observations made while plotting data, and analyzing the surveys, the first step of this test sought to determine if the experiment had induced suspicion. Hypothesis testing was performed to see if a user's reported suspicion and trust scores from the survey correlated with the intervals during which the experiment attempted to induce suspicion (i.e., after a manipulation occurred).

1. Rank Sum: Rank sum is a hypothesis test used to determine if two distributions differ significantly. This test was applied to the survey data for trust scores and suspicion scores. Each of these metrics was separated into a trust population corresponding to the intervals during which no manipulation occurred and a suspicion population corresponding to the intervals during which manipulations did occur. For this analysis, a positive result would have a small p-value because the distributions need to differ to provide value (i.e., there should be a significant difference in the survey responses depending on whether a manipulation occurred or not). A p-value less than 0.5 would be necessary to indicate that the distributions differed at all, with an ideal p-value being less than 0.3. The hypothesis test yielded a p-value of 0.69 for the trust populations and 0.9617 for the suspicion populations. Both of these values are above the threshold, indicating that the populations are do not differ significantly. Interestingly, it appears that the trust metric is more suited to reporting a user's mental state in the trust/suspicion domain. This could be due to a user's interpretation of suspicion in the context of the experiment.

2. Direct Analysis: To explore the direct relationship between one interval's expected and reported suspicion, a direct analysis on the pairing of these values was performed. For this exploration, we assume suspicion is a boolean variable; they are either suspicious or they are not. However, the survey data represents suspicion as a range between 1 and 7. As such, the survey data was adjusted to represent suspicion as a boolean variable. The suspicion data was normally a discrete value between one and seven (inclusive). For this analysis, only values of one were labeled as trusting and all other data was labeled as suspicious. This generous labeling was used because the suspicion scores did not vary greatly. While likely not the ideal method for separating the data, it gives the hypothesis the best chance for success; i.e., failing this test indicates that the suspicion survey data does not provide value for this study. The test indicated that there was no significant difference between the populations, therefore the focus of the analysis could be shifted towards identifying a link between mouse features and trust and mouse features and the manipulations. Further metrics found during this test are shown in Table 38.

Table 38: Direct Analysis Raw Metrics (RESCHU)

State	Count
Correct Positive	28
False Positive	49
Correct Negative	79
False Negative	36

These metrics show that there was only a 6% increase in the number of subjects reporting suspicion during an intended suspicious trial. While this indicates some correlation, it provides no information as to where suspicious data exists in the dataset because some subjects reported suspicion during a trusting interval as well as during a suspicious interval. Furthermore, because of how low the suspicion scores on the subjects' surveys were (the highest suspicion score was four out of seven, with a median value of one), it is not clear if suspicion was ever induced. As there is so much uncertainty in the data, it is unclear whether meaningful results can be drawn from this data by relying on the survey's measure of suspicion.

3. Feature Plots: Another key aspect of direct analysis was the creation of feature plots. These plots provide better visualization of the data. In particular, the suspicious and trusting datasets were compared on various features. This analysis revealed the following noteworthy observations about particular feature metrics:

- The DC metric for a Double Click feature is always zero in the trusting dataset, however it varies between 0 milliseconds and 20 milliseconds for two users in the suspicious data.
- The SCD metric for a Mouse Move + Double Click feature has far more data points above 1,000 milliseconds than the trusting data does.
- The TDC metric for a Mouse Move + Click feature has more values above 400 milliseconds in the suspicious data than the trusting data.
- Most of the values for the SCD metric for a Double Click feature are zero milliseconds and 50 milliseconds in the trusting data. The suspicious data appears to either be very close to zero, or above 100 milliseconds.
- The TDC metric for a Click feature has many more values near 800 milliseconds in the suspicious data than the trusting data, even though the trusting data has a higher maximum value.

5.4.2.2 Winter Survival Study (Thrust 5a)

The WSS included 20 users. As subjects only interacted via instant messengers, this experiment was selected as the primary data source for collecting and analyzing keystroke data. During the WSS, subjects were asked to prioritize a list of items based on their utility in surviving in a winter environment after conversing with another subject through Skype instant messenger. After a control trial, some subjects were asked to sabotage the lists of others, while subjects who were not selected to be saboteurs were informed that they may be interacting with saboteurs.

5.4.2.2.1 Feature Derivation

Key presses, key releases, and the timestamps of these actions were collected on each subject's machine. The keystrokes were paired together to form dyads, from which a set of five features was extracted: KHT1, KHT2, KIT, KPL, and KRL.Table 39, describes the features, where KHT is included twice (once per key). These features were selected because of their prevalence in keystroke dynamics literature and their utility for keystroke biometric verification and identification[11].

Table 39: Feature Descriptions (WSS)

Feature	Description
KHT	Time between the press and release of a single key
KIT	Time between the release of one key and the press of the next consecutive key
KPL	Time between the press of one key and the press of the next consecutive key
KRL	Time between the release of one key and the release of the next consecutive key

5.4.2.2.2 Suspicion Assignment

The currently accepted experimental ground-truth for suspicion is self-reporting. As a result, the experimental design included a question that asked users if they suspected that the person they were conversing with was a saboteur and, if so, when they became suspicious. Many subjects had scored high on the state suspicion survey [97], but indicated that they did not think they were conversing with a saboteur. AIS hypothesizes that there are two possible reasons for this: (1) the subject was suspicious of the other subject for other reasons (e.g., the subject doubted the accuracy of the other subject's list) or, (2) as suspicion was not defined in the surveys, users may have confused suspicion (i.e., lack of confidence in the benevolence of an individual) and distrust (i.e., confidence in the

[11] Assured Information Security Inc., "Remote Suspect Identification Final Technical Report," 2011.

animosity of another individual). As a result, the point at which a subject became suspicious could not be accurately determined.

As analysis progressed, it became clear that subjects in this study became suspicious far sooner than indicated. According to their suspicion scores from the survey and from their instant messenger conversations, it appeared that they became suspicious as soon as the saboteurs were introduced into the study. To mitigate this problem, AIS ignored the suspicion scores and labeled all data collected during the control trial as trusting and labeled all data collected in subsequent trials as suspicious. The assumption was made that subjects will be suspicious after the subjects have been informed that saboteurs exist for most of the trial, until they decide whether or not to trust or distrust the corresponding subject.

5.4.2.2.3 Data Pre-Processing

Keystroke dynamics focuses on the natural flow of keystrokes. As a result, all pauses and breaks in typing were removed. All data within a feature set where KIT is greater than 2,000 milliseconds or less than – 2,000 milliseconds was excluded. Feature sets in which KHT is greater than 2,500 milliseconds were dropped, as this is above the threshold at which the Windows OS begins to artificially repeat key strokes. As the focus of the effort is on actual user behavior and not keystrokes generated by the OS, simulated keystrokes (i.e., keys created by the OS when a key is held down) were dropped. Data produced by users given the role of saboteur was also dropped, as their mental state is subject to more variables. For example, saboteurs may be less likely to become suspicious or deliberate deception could produce an increase in mental workload that impacts keystroke timings.

5.4.2.2.4 Data Analysis

The data analysis performed for the WSS consisted of three key steps: PCA, logistic regression, and hypothesis testing using Rank Sum. This section describes the procedures for each stage in the analysis and includes a summary of the analysis and plots to better visualize the results. All analysis was performed on the pre-processed data to determine which, if any, features correlate with suspicion.

Principal Component Analysis

PCA was performed on the entire dataset to determine which features contained most of the variance within the dataset. The results from PCA (Table 40 and Table 41) suggested that KRL, KPL, and KIT were the most prominent features. As KRL and KPL both incorporate KIT (KPL = KHT1 + KIT, KRL = KIT + KHT2), KIT appeared to be the most promising feature.

Table 40: Component Statistics (WSS)

	Component 1	Component 2	Component 3	Component 4	Component 5
Std. Dev.	383.49	158.05	90.52	$2.256e^{-5}$	0
Proportion of Variance	0.816	0.138	0.045	$2.824e^{-15}$	0
Cumulative Proportion	0.816	0.954	1.000	1.000	1

Table 41: Component Composition (WSS)

	Component 1	Component 2	Component 3	Component 4	Component 5
KRL	0.652	– 0.401	0.196	– 0.428	0.438
KPL	0.517	0.268	– 0.535	– 0.270	– 0.549
KHT	0.110	– 0.781		0.428	– 0.438
KIT	0.543	0.380	0.248	0.698	0.111

Logistic Regression

As PCA showed many significant features, an ideal model for this data set would include all features. However, when performing linear modeling with logistic regression, the variables included in the model must be linearly independent to produce a valid model. Therefore, the number of variables used in this model must be reduced to a linearly independent subset. The logical choice for this reduction was to include KIT, KHT1, and KHT2, because these variables are the building blocks of KPL and KRL. The results of logistic regression on this variable set are shown below (Table 42).

Table 42: Logistic Regression Results (WSS)

	Estimate	Std. Error	Z Value	Pr(>|t|)
Intercept	1.725	$3.484e^{-3}$	495.241	$< 2e^{-16}$
KIT	$-7.601e^{-5}$	$8.042e^{-6}$	–9.451	$< 2e^{-16}$
KHT1	$-7.309e^{-5}$	$2.453e^{-5}$	–3.001	0.00269
KHT2	$3.827e^{-5}$	$1.316e^{-5}$	2.909	0.00363

The resulting p-values from this analysis indicate a close fit. However, the R^2 value for this model was 0.001221, indicating that the model accounts for only a small amount of the variance. The next natural step was to proceed to train the model and assess its accuracy at predicting whether a user is suspicious or not. The model was trained on half of the data set and predicted on the second half of the data set. The model was tested this way five times with different random seeds for each run. The results are in Table 43.

Table 43: Linear Model Predict Scores

Trial	AUC
1	0.5227
2	0.5275
3	0.5284
4	0.5277
5	0.5261

The results show that the model is only slightly better than random at predicting whether a user is suspicious or not suspicious. This is in line with the small R^2 value produced by the model. In an effort to increase the accuracy of the model, logistic regression was run on a data set with the same features, except with the features averaged across segments of the data. These segments are each 1/35 of the data of one pairing of users. It was theorized that this approach might help boost accuracy because it would lower the variance in the data, thereby causing the R^2 value for the model to rise. The results of this test are presented in Table 44 and

Table 45.

Table 44: Logistic Regression on Intervals Results

	Estimate	Std. Error	Z Value	Pr(>|t|)
Intercept	$9.195e^{-1}$	$2.982e^{-2}$	30.831	$< 2e^{-16}$
KIT	$-4.345e^{-4}$	$9.383e^{-5}$	–4.631	$3.78e^{-6}$
KHT1	$-7.551e^{-4}$	$3.375e^{-4}$	–2.237	0.0253
KHT2	$9.110e^{-5}$	$2.448e^{-4}$	0.372	0.7098

Table 45: Linear Model on Intervals Predict Scores

Trial	AUC
1	0.5527
2	0.5698
3	0.5574
4	0.5695
5	0.5588

The p-values from the model are much worse than the model that does not use the intervals. However, as predicted, there was an increase in the R^2 value. This model performed slightly better than the previous model, with very consistent results. Unfortunately, the model is not sufficiently predictive to be considered a good model. As such, AIS performed hypothesis testing and sought to develop more accurate models with other algorithms.

Hypothesis Testing

A two-sided Rank Sum test was used as the hypothesis test during this analysis. Rank sum aims to determine which, if any, of the features has a statistically significant location shift between its suspicious and trusting data. In order to be statistically significant, a threshold p-value of 0.01 was chosen. Any p-value equal to or less than this value was considered significant. The results (Table 46) show that all four features have significant location shifts, however KIT, KPL, and KRL have the most significant shifts. As KPL and KRL are partially composed of KIT, KIT was selected as the focus of further analysis.

Data Summary: *After demonstrating that the difference in distribution between suspicious and trusting datasets was statistically significant, some general statistics on the data were calculated. These statistics quantitatively highlight the differences between the trusting (*

Table 47) and suspicious (Table 48) data. All features are measured in milliseconds.

Table 46: Two-tailed Rank Sum Results (WSS)

Feature	P-value
KPL	$< 2.2e^{-16}$
KRL	$< 2.2e^{-16}$
KIT	$2.999e^{-16}$
KHT	0.0007619

Table 47: Summary of Trusting Dataset (WSS)

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
KRL	– 1646	100	149	218.3	224	4034
KPL	0	109	146	212.6	216	2230
KIT	– 1766	– 40	58	94.25	109	1997
KHT	5	78	109	124.1	141	3862

Table 48: Summary of Suspicious Dataset (WSS)

	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
KRL	– 1825	94	140	200.5	203	5828
KPL	0	100	140	193.8	194	3856
KIT	– 1941	– 47	50	76.43	94	1990
KHT	0	72	108	124	140	4868

This data shows a large difference in means and a moderate difference in medians for KIT between suspicious and trusting datasets. The median values are better indicator of each population's trend because the populations cannot be assumed normal. Indeed, as indicated by several plots, the populations are not normal – they tend to take on a bimodal distribution.

Feature Plots: The following density plots (Figure 55) depict the trusting (green) and suspicious (red) datasets overlaid on the same graph. The X-axis represents a keystroke feature in milliseconds. The distributions for each feature tend to follow the same general shape; however, the intensity at particular points differs between trusting and suspicious datasets. For example, in Figure 56, the suspicious data set has higher peaks and lower lows in this bimodal distribution.

In Figure 55, the suspicious data shows a slight tri-modal shape and a leftward shift compared to the more normally distributed and rightward shifted trusting data set. Figure 57 also shows the same trend.

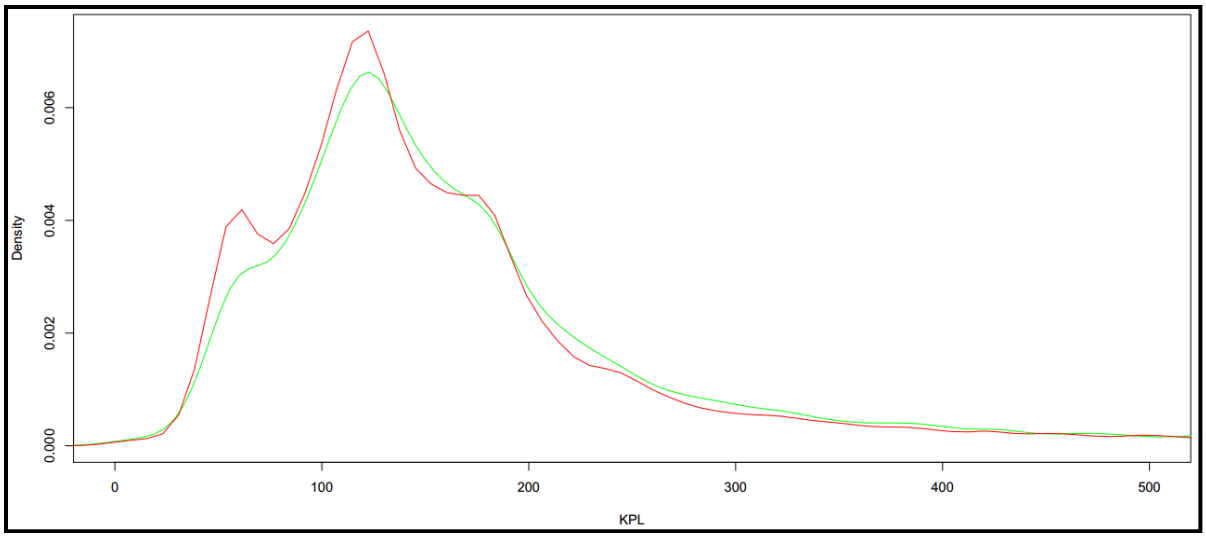


Figure 55: KPL Density Plot

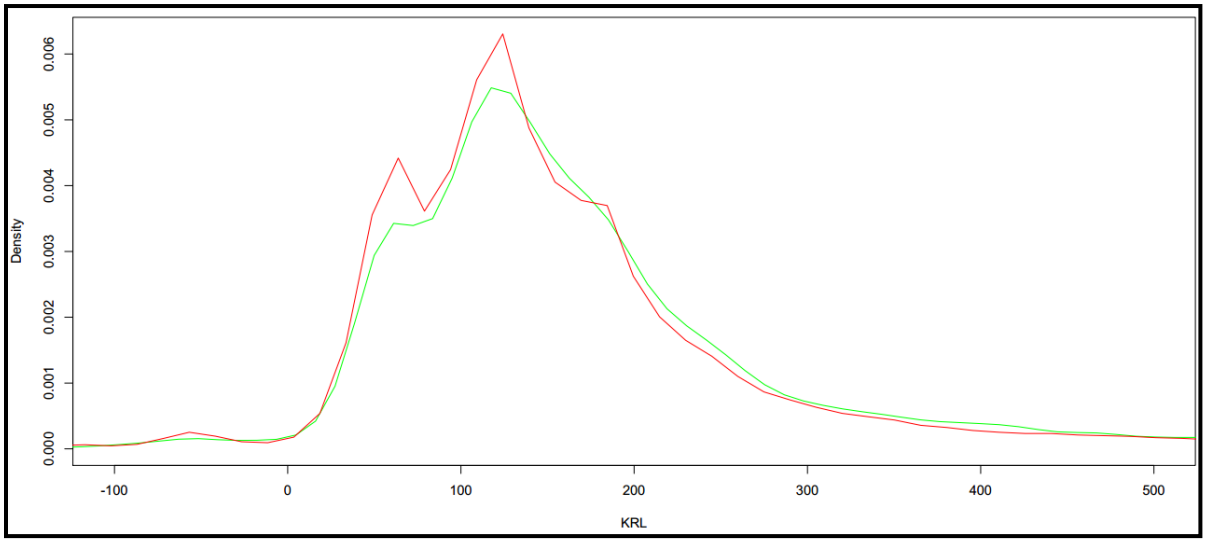


Figure 56: KRL Density Plot

The suspicious data in Figure 57 appears more normally distributed with a slight bimodal tendency in the trusting data and a slight tri-modal trend in the suspicious data. Both the trusting and suspicious data are more closely aligned for this plot than the others.

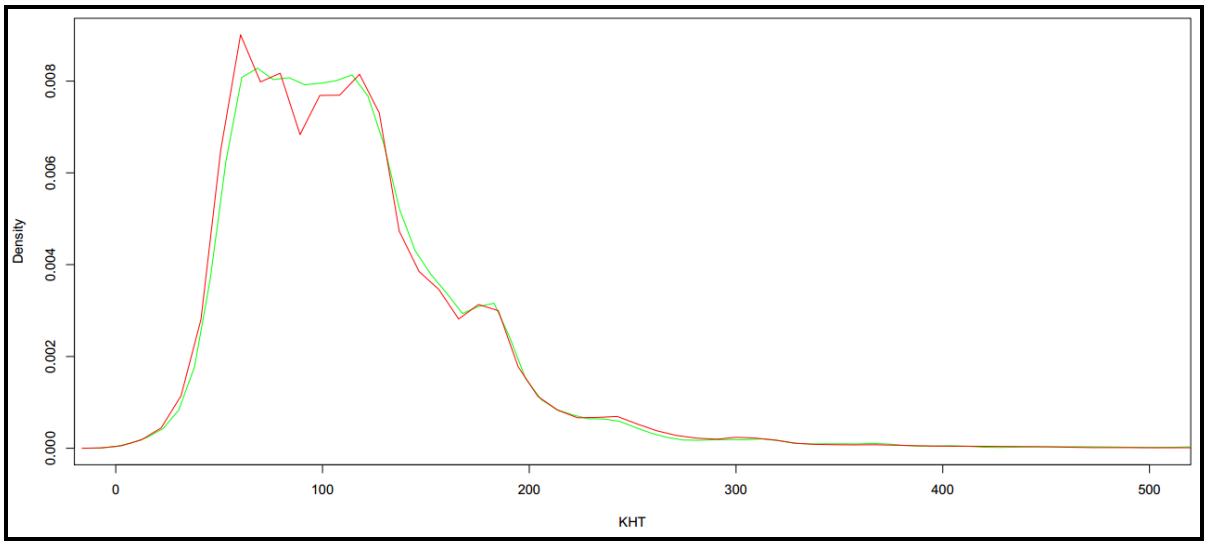


Figure 57: KHT Density Plot

5.4.2.3 Attribution Experiment Analysis (Thrust 5b)

The experiment and subsequent analysis for this effort was not performed due to the early termination of the contract.

5.4.2.4 Manipulation Experiment Analysis (Thrust 5c)

The experiment and subsequent analysis for this effort was not performed due to the early termination of the contract.

5.5 CONCLUSION

This effort has demonstrated that cyber sensors provide some information towards a user's cognitive state (suspicion, specifically). However, further research is needed to determine the degree to which the keystroke sensor can be used to detect or measure suspicion. In addition, many other sensors have been developed that may add value in suspicion detection. These sensors were designed to assist with suspicion attribution, but the experiment to collect data to assess their viability has not been performed. Finally, once suspicion can successfully be measured and attributed, further research is needed to determine if suspicion can also be manipulated remotely.

While further research is certainly necessary, the results indicate a promising new field of research that could be of great value to both the military and commercial worlds. For example, blue-force operators could be monitored to identify changes in cognitive states or red-force operators could be remotely measured to identify vulnerabilities. From the commercial side, the sensors could be adapted to address insider threats.

5.6 RECOMMENDATIONS

The research performed under this effort should be expanded in future research efforts to provide critical tools to cyber operators. First and foremost, second and third thrusts should be completed in the anticipated follow-on effort to establish the feasibility of using cyber sensors and tools to attribute and manipulate suspicion, respectively.

Perhaps the best and most useful expansion of the work performed under this effort would be to determine if the sensors developed under this effort can detect other cognitive states. If possible, it is reasonable to assume that cyber tools may be useful for manipulating these states. Such research could:

- Spawn entirely new D5 effects.
- Improve existing D5 effects.
- Reduce the chance of D5 effect detection (e.g., can the configurable parameters be tuned to avoid detection).
- Develop methods to perform battle damage assessment to measure the effectiveness of a D5 effect (e.g., was it detected by the operator).
- Perform risk management for operators (e.g., guiding a friendly operator to the correct suspicion attribution to avoid mission compromise).
- Devise methods for improving operator vigilance.
- Potentially detect incoming D5 attacks.

There are many other methods that can be used to expand upon the research performed under this effort and many ways in which it could be applied to existing problems. The following recommendations are but a subset that should provide value to the CNO realm. The effort could benefit from incorporating other cyber sensors and expanding to other platforms (e.g., mobile devices) to gather more user data that could be indicative of suspicion or assist in attribution. The current framework uses a database as an intermediary between data acquisition and processing; however, a real-time system could be developed using the algorithms that this research has unearthed. This system would be capable of detecting, attributing, and manipulating suspicion as it is occurring to achieve the aforementioned research goals.

CONCLUSION

In this report, we provide the major findings of the Cyber Trust and Suspicion research project, which focused on modeling and understanding trust and suspicion in cyber organizations, especially on how they affect insider threats. Taking into account the breadth of this research, the project had five main thrusts, led by individual teams. The thrusts were:

- THRUST 1: A Social, Cultural, And Emotional Basis For Trust And Suspicion: Manipulating Insider Threat In Cyber Intelligence & Operations
- THRUST 2: Targeted Interventions Derived From Biomarkers Of Cyber Trust
- THRUST 3: Cyber Trust And Suspicion: A Human-Centric Approach
- THRUST 4: Using Non-Invasive Sensors To Predict Trust And Suspicion In Human Operators
- THRUST 5: Assessing, Attributing, And Manipulating Operator Suspicion

The report provides details on the research objectives, methodology, results and findings in each of the thrusts during the truncated project performance period from 09/30/2012 to 12/31/2014. As future work, our focus will be on integrating the models and findings from the individual thrusts to formulate methodology(ies) which are overarching in its understanding of the influence of cyber trust and suspicion in modeling and analyzing the critical issue of insider threats in cyber organizations.

REFERENCES

[1] A. Zinck and A. Newen, "Classifying emotion: A developmental account," *Synthese*, vol. 161, no. 1, pp. 1–25, 2008.

[2] R. Plutchik, "The Multifactor-Analytic Theory of Emotion," *The Journal of Psychology*, vol. 50.

pp. 153–171, 1960.

[3] L. R. Goldberg, "The structure of phenotypic personality traits.," *Am. Psychol.*, vol. 48, no. 1, pp. 26–34, 1993.

[4] H. E. P. Cattell and A. D. Mead, "The Sixteen Personality Factor Questionnaire (16PF)," *SAGE Handb. Personal. Theory Assess.*, pp. 135–159, 2003.

[5] S. E. Krug and E. F. Johns, "A large scale cross-validation of second-order personality structure defined by the 16PF," *Psychol. Rep.*, vol. 59, no. 2, pp. 683–693, 1986.

[6] D. W. Gerbing and M. R. Tuley, "The 16PF related to the five-factor model of personality: Multiple-indicator measurement versus the a priori scales," *Multivariate Behav. Res.*, vol. 26, no. 2, pp. 271–289, 1991.

[7] *US v. Lloyd BT - F. 3d*, vol. 269, no. No. 00–2409. Court of Appeals, 3rd Circuit, 2001, p. 228.

[8] S. Gaudin, "Case Study of Insider Sabotage: The Tim Lloyd / Omega Case," *Comput. Secur. J.*, vol. XVI, no. 3, pp. 1–8, 2000.

[9] A. Harris and M. Rohde, "Wisconsin Researcher Accused of Economic Spying for China - Bloomberg Business," *Bloomberg Business*, 2013. .

[10] E. Santos Jr. and E. S. Santos, "A Framework for Building Knowledge-Bases Under Uncertainty," *J. Exp. Theor. Artif. Intell.*, vol. 11, pp. 265–286, 1999.

[11] E. Santos Jr., J. T. Wilkinson, and E. E. Santos, "Bayesian Knowledge Fusion," in *Proc. 2nd International FLAIRS Conference*, 2009, pp. 559–564.

[12] J. Heron, "Catharsis in human development," *Hum. Potential Res. Proj.*, 1998.

[13] T. Weiner, D. Johnston, and N. A. Lewis, *Betrayal: The Story of Aldrich Ames, an American Spy*. New York: Random House.

[14] F. G. Moeller, E. S. Barratt, D. M. Dougherty, J. M. Schmitz, and A. C. Swann, "Psychiatric Aspects of Impulsivity," *Am. J. Psychiatry*, vol. 158, no. 11, pp. 1783–1793, Nov. 2001.

[15] S. P. Whiteside and D. R. Lynam, "The five factor model and impulsivity: Using a structural model of personality to understand impulsivity," *Pers. Individ. Dif.*, vol. 30, no. 4, pp. 669–689, 2001.

[16] E. D. Shaw and L. F. Fischer, "Ten tales of betrayal: The threat to corporate infrastructure by information technology insiders analysis and observations," DTIC Document, 2005.

[17] E. D. Shaw, J. M. Post, and K. G. Ruby, "Inside the Mind of the Insider," *Secur. Manag.*, vol. 43, no. 12, pp. 34–42, 1999.

[18] M. McBride, L. Carter, and M. Warkentin, "The Role of Situational Factors and Personality on Cybersecurity Policy Violation," 2012.

[19] R. J. Larsen and T. Ketelaar, "Personality and susceptibility to positive and negative emotional states.," *J. Pers. Soc. Psychol.*, vol. 61, no. 1, p. 132, 1991.

[20] F. L. Greitzer, L. J. Kangas, C. F. Noonan, A. C. Dalton, and R. E. Hohimer, "Identifying at-risk employees: Modeling psychosocial precursors of potential insider threats," in *45th Hawaii International Conference on System Science (HICSS 2012)*, 2012, pp. 2392–2401.

[21] R. T. Wright and K. Marett, "The Influence of Experiential and Dispositional Factors in Phishing: An Empirical Investigation of the Deceived," *J. Manag. Inf. Syst.*, vol. 27, no. 1, pp. 273–303, 2010.

[22] C. K. W. De Dreu, L. L. Greer, M. J. J. Handgraaf, S. Shalvi, G. A. Van Kleef, M. Baas, F. S. Ten Velden, E. Van Dijk, and S. W. W. Feith, "The neuropeptide oxytocin regulates parochial altruism in intergroup conflict among humans.," *Science*, vol. 328, no. 5984, pp. 1408–1411, 2010.

[23] M. Mikolajczak, J. J. Gross, A. Lane, O. Corneille, P. de Timary, and O. Luminet, "Oxytocin makes people trusting, not gullible.," *Psychol. Sci.*, vol. 21, no. 8, pp. 1072–1074, 2010.

[24] H. Tost, B. Kolachana, S. Hakimi, H. Lemaitre, B. A. Verchinski, V. S. Mattay, D. R. Weinberger, and A. Meyer-Lindenberg, "A common allele in the oxytocin receptor gene (OXTR) impacts prosocial temperament and human hypothalamic-limbic structure and function.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 31, pp. 13936–13941, 2010.

[25] P. J. Zak, R. Kurzban, and W. T. Matzner, "Oxytocin is associated with human trustworthiness.," *Horm. Behav.*, vol. 48, no. 5, pp. 522–527, 2005.

[26] M. Kosfeld, M. Heinrichs, P. J. Zak, U. Fischbacher, and E. Fehr, "Oxytocin increases trust in humans.," *Nature*, vol. 435, no. 7042, pp. 673–676, 2005.

[27] P. A. Bos, D. Terburg, and J. van Honk, "Testosterone decreases trust in socially naive humans.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 22, pp. 9991–9995, 2010.

[28] R. T. Johnson and S. M. Breedlove, "Human trust: testosterone raises suspicion.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 107, no. 25, pp. 11149–11150, 2010.

[29] P. J. Zak and A. Fakhar, "Neuroactive hormones and interpersonal trust: international evidence.," *Econ. Hum. Biol.*, vol. 4, no. 3, pp. 412–429, 2006.

[30] R. Riedl and A. Javor, "The biology of trust: integrating evidence from genetics, endocrinology, and functional brain imaging.," *J. Neurosci. Psychol. Econ.*, vol. 5, pp. 63–91, 2011.

[31] Z. R. Donaldson and L. J. Young, "Oxytocin, vasopressin, and the neurogenetics of sociality.," *Science*, vol. 322, no. 5903, pp. 900–904.

[32] R. Thompson, S. Gupta, K. Miller, S. Mills, and S. Orr, "The effects of vasopressin on human facial responses related to social communication.," *Psychoneuroendocrinology*, vol. 4530, no. 02, pp. 35–48, 2004.

[33] M. Heinrichs, T. Baumgartner, C. Kirschbaum, and U. Ehlert, "Social support and oxytocin interact to suppress cortisol and subjective responses to psychosocial stress.," *Biol. Psychiatry*, vol. 54, no. 12, pp. 1389–1398, 2003.

[34] T. Takahashi, K. Ikeda, M. Ishikawa, N. Kitamura, T. Tsukasaki, D. Nakama, and T. Kameda, "Interpersonal trust and social stress-induced cortisol elevation.," *Neuroreport*, vol. 16, no. 2, pp. 197–199, 2005.

[35] R. F. P. de Winter, A. M. van Hemert, R. H. DeRijk, K. H. Zwinderman, A. C. Frankhuijzen-Sierevogel, V. M. Wiegant, and J. G. Goekoop, "Anxious-retarded depression: relation with plasma vasopressin and cortisol.," *Neuropsychopharmacology*, vol. 28, no. 1, pp. 140–147.

[36] R. P. Ebstein, S. Israel, E. Lerer, F. Uzefovsky, I. Shalev, I. Gritsenko, M. Riebold, S. Salomon, and N. Yirmiya, "Arginine vasopressin and oxytocin modulate human social behavior.," *Ann. N. Y. Acad. Sci.*, vol. 1167, pp. 87–102, 2009.

[37] B. Ditzen, M. Schaer, B. Gabriel, G. Bodenmann, U. Ehlert, and M. Heinrichs, "Intranasal oxytocin increases positive communication and reduces cortisol levels during couple conflict.," *Biol. Psychiatry*, vol. 65, no. 9, pp. 728–731, 2009.

[38] W. Schultz, "Getting formal with dopamine and reward.," *Neuron*, vol. 36, no. 2, pp. 241–263, 2002.

[39] M. Barrot, M. Marinelli, D. N. Abrous, F. Rougé-Pont, M. Le Moal, and P. V Piazza, "The dopaminergic hyper-responsiveness of the shell of the nucleus accumbens is hormone-dependent.," *Eur. J. Neurosci.*, vol. 12, no. 3, pp. 973–979, 2000.

[40] T. A. Baskerville and A. J. Douglas, "Dopamine and oxytocin interactions underlying behaviors: potential contributions to behavioral disorders.," *CNS Neurosci. Ther.*, vol. 16, no. 3, pp. e92–123, 2010.

[41] C. F. Ferris and Y. Delville, "Vasopressin and serotonin interactions in the control of agonistic behavior.," *Psychoneuroendocrinology*, vol. 19, no. 5–7, pp. 593–601, 1994.

[42] J. S. Winston, B. A. Strange, J. O'Doherty, and R. J. Dolan, "Automatic and intentional brain responses during evaluation of trustworthiness of faces.," *Nat. Neurosci.*, vol. 5, no. 3, pp. 277–283, 2002.

[43] E. Fehr and C. F. Camerer, "Social neuroeconomics: the neural circuitry of social preferences.," *Trends Cogn. Sci.*, vol. 11, no. 10, pp. 419–427, 2007.

[44] T. Kugler, T. Connolly, and E. E. Kausel, "The effect of consequential thinking on trust game behavior.," *J. Behav. Decis. Mak. doi101002bdm614*, vol. 22, pp. 101–119, 2009.

[45] B. King-Casas, D. Tomlin, C. Anen, C. F. Camerer, S. R. Quartz, and P. R. Montague, "Getting to know you: reputation and trust in a two-person economic exchange.," *Science*, vol. 308, no. 5718, pp. 78–83, 2005.

[46] G. Miller, "Neuroscience. Economic game shows how the brain builds trust.," *Science*, vol. 308, no. 5718, p. 36, 2005.

[47] M. R. Delgado, R. H. Frank, and E. A. Phelps, "Perceptions of moral character modulate the neural systems of reward during the trust game.," *Nat. Neurosci.*, vol. 8, no. 11, pp. 1611–1618, 2005.

[48] T. Baumgartner, M. Heinrichs, A. Vonlanthen, U. Fischbacher, and E. Fehr, "Oxytocin shapes the neural circuitry of trust and trust adaptation in humans.," *Neuron*, vol. 58, no. 4, pp. 639–650, 2008.

[49] A. Dimoka, "What Does the Brain Tell Us About Trust and Distrust? Evidence from a Functional Neuroimaging Study," *Manag. Inf. Syst. Q.*, vol. 34, no. 2, 2010.

[50] R. Riedl, M. Hubert, and P. Kenning, "Are there neural gender differences in online trust? An fMRI study on the perceived trustworthiness of eBay offers.," vol. 34, no. 2, pp. 397–428, 2010.

[51] W. C. Drevets, J. L. Price, M. E. Bardgett, T. Reich, R. D. Todd, and M. E. Raichle, "Glucose metabolism in the amygdala in depression: relationship to diagnostic subtype and plasma cortisol levels.," *Pharmacol. Biochem. Behav.*, vol. 71, no. 3, pp. 431–447, 2002.

[52] J. L. Price and W. C. Drevets, "Neurocircuitry of mood disorders.," *Neuropsychopharmacology*, vol. 35, no. 1, pp. 192–216, 2010.

[53] J. L. Price and W. C. Drevets, "Neural circuits underlying the pathophysiology of mood disorders.," *Trends Cogn. Sci.*, vol. 16, no. 1, pp. 61–71, 2012.

[54] M. Wilson and J. Hash, "Building an information technology security awareness and training program," in *NIST Special publication*, 2003.

[55] G. Magill, "The crucial role of stewardship in health care ethics.," *Health Care Ethics USA*, vol. 8, no. 2, pp. 2–3, 2000.

[56] B. D. Cone, C. E. Irvine, M. F. Thompson, and T. D. Nguyen, "A video game for cyber security training and awareness.," *Comput. Secur.*, vol. 26, pp. 63–72, 2006.

[57] G. C. I. Thorton and R. A. Mueller-Hanson, *Developing organizational simulations: A guide for practitioners and students*. Mahwah, NJ, 2004.

[58] J. Cannon-Bowers and C. Bowers, "Synthetic learning environments: On developing a science of simulation, games, and virtual worlds for training," in *Learning, training, and development in organizations*, S. W. J. Koslowski and E. Salas, Eds. 2010, pp. 229–261.

[59] E. Salas, J. L. Wildman, and R. F. Piccolo, "Using simulation-based training to enhance management education.," *Acad. Manag. Learn. Educ.*, vol. 8, pp. 559–573, 2009.

[60] E. A. Day, C. Blair, S. Daniels, V. Kligyte, and M. D. Mumford, "Linking instructional objectives to the design of instructional environments: The integrative design matrix.," *Hum. Resour. Manag. Rev.*, vol. 16, pp. 376–395, 2006.

[61] R. D. Rogers, A. M. Owen, H. C. Middleton, E. J. Williams, J. D. Pickard, B. J. Sahakian, and T. W. Robbins, "Choosing between small, likely rewards and large, unlikely rewards activates inferior and orbital prefrontal cortex.," *J. Neurosci.*, vol. 19, no. 20, pp. 9029–9038, 1999.

[62] S. Baron-Cohen, T. Jolliffe, C. Mortimore, and M. Robertson, "Another advanced test of theory of mind: evidence from very high functioning adults with autism or asperger syndrome.," *J. Child Psychol. Psychiatry.*, vol. 38, no. 7, pp. 813–822, 1997.

[63] J. Staggs, R. Beyer, M. Mol, M. Fisher, B. Brummel, and J. Hale, "A perceptual taxonomy of contextual cues for cyber trust.," *Proceeding Colloq. Inf. Syst. Secur. Educ. CISSE*, vol. 2, pp. 152–169, 2014.

[64] A. Herzberg and A. Jbara, "Security and identification indicators for browsers against spoofing and phishing attacks.," *ACM Trans. Internet Technol.*, vol. 8, no. 4, 2008.

[65] S. Egelman, L. F. Cranor, and J. Hong, "You've Been Warned: An Empirical Study of the Effectiveness of Web Browser Phishing Warnings," *Proceeding twenty-sixth Annu. CHI Conf. Hum. factors Comput. Syst. - CHI '08*, p. 1065, 2008.

[66] J. Hong, "The state of phishing attacks," *Commun. ACM*, vol. 55, no. 1, p. 74, Jan. 2012.

[67] A. Adelsbach, S. Gajek, and J. Schwenk, "Visual spoofing of SSL protected web sites and effective countermeasures.," *Inf. Secur. Pract. Exp. pp Springer Berlin Heidelb.*, pp. 204–215, 2005.

[68] D. K. McGrath and M. Gupta, "Behind phishing: an examination of phisher modi-operandi.," in *Proceedings of the USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET)*, 2008.

[69] M. Cova, C. Kruegel, and G. Vigna, "There is No Free Phish: An Analysis of 'Free' and Live Phishing Kits," in *Proceedings of the 2nd Conference on USENIX Workshop on Offensive Technologies*, 2008.

[70] C. E. Drake, J. J. Oliver, and E. J. Koontz, "Anatomy of a Phishing Email," in *Proceedings of the First Conference on E-mail and Anti-Spam (CEAS)*, 2004, pp. 1–8.

[71] G. R. Goldstein, "What is a clinical scholar?," *J. Prosthodont.*, vol. 20, no. 7, pp. 501–502, 2011.

[72] M. H. Kabat, R. L. Kane, A. L. Jefferson, and R. K. DiPino, "Construct validity of selected Automated Neuropsychological Assessment Metrics (ANAM) battery measures.," *Clin.*

Neuropsychol., vol. 15, no. 4, pp. 498–507, 2001.

[73] S. Furman, M. F. Theofanos, Y. Y. Choong, and B. Stanton, "Basing cybersecurity training on user perceptions," *IEEE Secur. Priv.*, vol. 10, pp. 40–49, 2012.

[74] R. D. Rogers, E. M. Tunbridge, Z. Bhagwagar, W. C. Drevets, B. J. Sahakian, and C. S. Carter, "Tryptophan depletion alters the decision-making of healthy volunteers through altered processing of reward cues.," *Neuropsychopharmacology*, vol. 28, no. 1, pp. 153–162, 2003.

[75] S. C. Brown and L. A. Mitchell, "An observational investigation of poker style and the five-factor personality model.," *J. Gambl. Stud.*, vol. 26, no. 2, pp. 229–234, 2010.

[76] D. S. Owens and D. Benton, "The impact of raising blood glucose on reaction times.," *Neuropsychobiology*, vol. 30, no. 2–3, pp. 106–113, 1994.

[77] N. Awad, M. Gagnon, and C. Messier, "The relationship between impaired glucose tolerance, type 2 diabetes, and cognitive function.," *J. Clin. Exp. Neuropsychol.*, vol. 26, no. 8, pp. 1044–1080, 2004.

[78] R. Stephens and R. J. Tunney, "Role of glucose in chewing gum-related facilitation of cognitive function.," *Appetite*, vol. 43, no. 2, pp. 211–213, 2004.

[79] A. H. Miller, V. Maletic, and C. L. Raison, "Inflammation and its discontents: the role of cytokines in the pathophysiology of major depression.," *Biol. Psychiatry*, vol. 65, no. 9, pp. 732–741, 2009.

[80] T. W. W. Pace and A. H. Miller, "Cytokines and glucocorticoid receptor signaling. Relevance to major depression.," *Ann. N. Y. Acad. Sci.*, vol. 1179, pp. 86–105, 2009.

[81] B. J. Brummel, J. Hale, and M. J. Mol, "Training cyber security personnel," in *The Psychosocial Dynamics of Cyber Security*, Taylor & Francis, 2016.

[82] R. E. Beyer and B. J. Brummel, "Implementing effective cyber security training for end users of computer networks," in *SHRM-SIOP Science of HR Series: Promoting Evidence-Based HR*, 2015.

[83] M. Deutsch, "Trust and suspicion," *J. Conflict Resolut.*, pp. 265–279, 1958.

[84] R. Axelrod and R. Iliev, "Timing of cyber conflict," *Proc. Natl. Acad. Sci.*, vol. 111, no. 4, pp. 1298–1303, 2014.

[85] P. Bobko, A. Barelka, and L. M. Hirshfield, "The construct of state-level suspicion: a model and research agenda for automated and information technology (IT) contexts.," *Hum. Factors*, vol. 56, no. 3, pp. 489–508., 2014.

[86] P. J. Bobko, A. Barelka, and L. M. Hirshfield, "The construct of state-level suspicion: a model and research agenda for automated and information technology (IT) contexts," *Hum. Factors*, vol. 56, no. 3, pp. 489–508, 2014.

[87] L. Hirshfield, P. Bobko, A. J. Barelka, M. R. Costa, G. J. Funke, V. F. Mancuso, V. Finomore, and B. A. Knott, "The Role of Human Operators' Suspicion in the Detection of Cyber Attacks," *Int. J. Cyber Warf. Terror.*, vol. 5, no. 3, pp. 28–44, Jul. 2015.

[88] L. Hirshfield, R. Dora, C. Webster, and P. Bobko, "Predicting Trust, Deception, and Suspicion during Online Interactions With a Keylogger," *Submitt. to J. Behav. Inf. Technol.*, 2015.

[89] C. Solinger, L. Hirshfield, S. Hirshfield, R. Friendman, and C. Leper, "Beyond Facebook Personality Prediction:," in *Social Computing and Social Media*, G. Meiselwitz, Ed. Springer International Publishing, 2014, pp. 486–493.

[90] N. Sommer, L. Hirshfield, and S. Velipisalar, "Our Emotions as Seen through a Webcam," in

Foundations of Augmented Cognition. Advancing Human Performance and Decision-Making through Adaptive Systems, Springer, 2014, pp. 78–89.

[91] L. M. Hirshfield, P. Bobko, A. Barelka, S. Hirshfield, S. Hincks, S. Gulbrunson, M. Farrington, and D. Paverman, "Using Non-Invasive Brain Measurement to Explore the Psychological Effects of Computer Malfunctions on Users During Human-Computer Interactions," *Adv. Human-Computer Interact.*, 2014.

[92] L. 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," 2011.

[93] J. Escalante, L. M. Hirshfield, and S. Butcher, "Evaluating Interfaces Using Electroencephalography and the Error Potential," 2013.

[94] D. W. Johnson and F. P. Johnson, "Joining Together: Group Theory and Group Skills," Englewood Cliffs, NJ: Prentice-Hall, 1975.

[95] K. Hoff and M. Bashir, "Trust in automation: Integrating empirical evidence on factors that influence trust," *Hum. Factors*, 2014.

[96] J. B. Walther, "Interpersonal effects in computer-mediated interaction a relational perspective," *Communic. Res.*, vol. 19, no. 1, pp. 52–90, 1992.

[97] P. J. Bobko and H. Odle-Dousseau, "Preliminary Report on the Development and Psychometric Analysis of a Twenty-Item, Self-Report Measure of State Suspicion," 2014.

[98] L. M. Hirshfield, P. J. Bobko, A. Barelka, S. Hirshfield, M. Farrington, S. Gulbronson, and D. Paverman, "Using Non-Invasive Brain Measurement to Explore the Psychological Effects of Computer Malfunctions on Users During Human-Computer Interactions," *Adv. Human-Computer Interact.*, 2014.

[99] D. Grimes, D. S. Tan, S. E. Hudson, P. Shenoy, and R. P. N. Rao, "Feasibility and Pragmatics of Classifying Working Memory Load with an Electroencephalograph," in *Conference on Human Factors in Computing Systems (CHI 2008)*, 2008.

[100] S. K. Vuyyuru, V. V. Phoha, S. S. Joshi, S. Phoha, and A. Ray, "Computer user authentication using hidden markov model through keystroke dynamics," *Manuscr. Submitt. to ACM Trans. Inf. Syst. Secur.*, 2006.

[101] R. A. Dora, P. D. Schalk, J. E. McCarthy, and S. A. Young, "Remote Suspect Identification and the impact of demographic features on keystroke dynamics," in *SPIE Defense, Security, and Sensing: Cyber Sensing*, 2013.

[102] M. Pusara and C. E. Brodley, "User re-authentication via mouse movements," in *Proceedings of the 2004 ACM Workshop on Visualization and Data Mining for Computer Security (VizSEC/DMSEC'04)*, 2004, pp. 1–8.

[103] D. A. Schulz, "Mouse Curve Biometrics," in *Biometric Consortium Conference 2006, Biometrics Symposium*, 2006, pp. 1–6.

[104] A. A. E. Ahmed and I. Traore, "A New Biometric Technology Based on Mouse Dynamics," *IEEE Trans. Dependable Secur. Comput.*, vol. 4, no. 3, pp. 165–179, 2007.

[105] C. Feher, Y. Elovici, R. Moskov, L. Rokach, and A. Schclar, "User identity verification via mouse dynamics," *Inf. Sci. (Ny).*, vol. 201, pp. 19–36, 2012.

APPENDIX A: ACRONYMS

Acronym	Definition
AFOSR	Air Force Office of Scientific Research
AIS	Assured Information Security, Inc.
BRI	Basic Research Initiative
CMC	Computer-Mediated Communicated
CNO	Computer Network Operations
CT	Click Time
CTS	Cyber Trust and Suspicion
D5	Deny, Disrupt, Degrade, Deceive, and Destroy
DC	Distance to Click
DLL	Dynamic-link Library
EEG	Electroencephalogram
FCD	First Click Distance
FCT	First Click Time
fMRI	Functional Magnetic Resonance Imaging
fNIRS	Functional near-Infrared Spectroscopy
GSR	Galvanic Skin Response
ID	Interval Distance
IT	Interval Time
KHT	Key Hold Time
KIT	Key Interval Time
KPL	Key Press Latency
KRL	Key Release Latency
M3	Third Moment
M4	Fourth Moment
OS	Operating System
PCA	Principal Component Analysis
RESCHU	Research Environment for Supervisory Control of Heterogeneous Unmanned Vehicles
SC	Scattering Coefficient
SCD	Second Click Distance
SCT	Second Click Time
SU	Syracuse University
TC	Time to Click
TCM	Trajectory Center of Mass
TCrv	Trajectory Curvature
TDC	Traveled Distance during Click
UTEP	University of Texas at El Paso
VCrv	Velocity Curvature
WSS	Winter Survival Study