March 2020

An Empirical Assessment of the Effectiveness of Deception for Cyber Defense

Kimberly J. Ferguson-Walter
University of Massachusetts Amherst

AN EMPIRICAL ASSESSMENT OF THE EFFECTIVENESS OF DECEPTION FOR CYBER DEFENSE

A Dissertation Presented

by

KIMBERLY J. FERGUSON-WALTER

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

February 2020

College of Information and Computer Sciences

AN EMPIRICAL ASSESSMENT OF THE EFFECTIVENESS OF DECEPTION FOR CYBER DEFENSE

A Dissertation Presented

by

KIMBERLY J. FERGUSON-WALTER

Approved as to style and content by:

Brian Levine, Chair

David Jensen, Member

Phillipa Gill, Member

Shannon Roberts, Outside Member

Dana LaFon, Member

James Allan, Chair
College of Information and Computer Sciences

DEDICATION

For my parents and grandparents, who always taught me to value knowledge and education, my husband, who never stopped encouraging me to finish, and my children, who brighten my life. LYA

ACKNOWLEDGMENTS

Thank you to my committee for their insightful questions and recommendations on my research. I want to give special thanks to Prof. Brian Levine, Prof. David Jensen, Dr. Dana LaFon, Dr. George Coker, Mary Berlage, Prof. Eliot Moss, and Dr. Jose Romero-Mariona whose support helped ease my nontraditional path for completion. This was only possible thanks to my friends, and family who convinced me it was possible: Dr. Erich Walter, Dr. Mike Santos, Dr. Dave Steuerman, Dr. Hanna Wallach, Prof. George Konidaris, Dr. Sarah Osentoski, Prof. Phil Thomas, and Prof. Andy Barto, and all the others who cheered me on along the way. I also want to voice my appreciation to all of my co-authors and collaborators, without whom this research would not have been achievable, especially, Maxine Major, Dr. Sunny Fugate, Dr. Temmie Shade, Andrew Rogers, Dr. Dirk van Bruggen, Prof. Robert Gutzwiller, and Chelsea Johnson.

ABSTRACT

AN EMPIRICAL ASSESSMENT OF THE EFFECTIVENESS OF DECEPTION FOR CYBER DEFENSE

FEBRUARY 2020

KIMBERLY J. FERGUSON-WALTER

B.Sc., UNIVERSITY OF CALIFORNIA IRVINE

M.Sc., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor Brian Levine

The threat of cyber attacks is a growing concern across the world, leading to an increasing need for sophisticated cyber defense techniques. The Tularosa Study, was designed and conducted to understand how defensive deception, both cyber and psychological, affects cyber attackers Ferguson-Walter et al. [2019c]. More specifically, for this empirical study, cyber deception refers to a decoy system and psychological deception refers to false information of the presence of defensive deception techniques on the network. Over 130 red teamers participated in a network penetration test over two days in which we controlled both the presence of and explicit mention of deceptive defensive techniques. To our knowledge, this represents the largest study of its kind ever conducted on a skilled red team population. In addition to the abundant host and network data collected, we conducted a battery of questionnaires, e.g., experience, personality; and cognitive tasks, e.g., fluid intelligence, working memory; as well as

physiological measures, e.g., galvanic skin response (GSR), heart rate, to be correlated with the cyber events at a later date. The design and execution of this study and the lessons learned are a major contribution of this thesis. I investigate the effectiveness of decoy systems for cyber defense by comparing performance across all experimental conditions. Results support a new finding that the combination of the presence of deception and the true information that deception is present has the greatest effect on cyber attackers, when compared to a control condition in which no deception was used. Evidence of cognitive biases in the red teamers' behavior is then detailed and explained, to further support our theory of oppositional human factors (OHF). The final chapter discusses how elements of the experimental design contribute to the validity of assessing the effectiveness of cyber deception and reviews trade-offs and lessons learned.

TABLE OF CONTENTS

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

AN INTRODUCTION TO CYBER DECEPTION

Humans are subject to cognitive limitations and bias, which can lead to lower quality decision making and flawed behaviors. While cyber attackers often take advantage of these limitations (e.g., spearphishing, and spam), it has been less explored for cyber defense. As network owners, defenders can use their home-field advantage to present information to attackers in specific ways to take advantage of and exacerbate innate human biases to delay, disrupt, or deter the attack. Deception for cyber defense advances that goal: to rebalance the asymmetric nature of computer defense by increasing attacker workload while decreasing that of the defender through strategic interactions with the human behind the cyber attack.

Creating a system that is always protected and secure is a far-reaching goal. While it is important for researchers to continue to move systems closer towards this absolute security, it is also essential to create techniques so a system can defend against an attacker who circumvents the current security defenses. Many techniques have been developed to increase the speed and accuracy of detecting intrusion activity at the perimeter with the aim of making a cyber defender's job easier Buczak and Guven [2016], Modi et al. [2013]. However, beyond *a priori* hardening of systems, less research has been done on techniques to make attackers' tasks harder. Deception for cyber defense is an emerging area of research aimed at just that. Deception holds promise as a successful tactic for making an attacker's job harder because it does more than just block access: it can also cause the attacker to waste both time and effort. Moreover, deception can be used by a defender to impart an incorrect belief in the attacker, the effects of which can go beyond any static defense.

In scenarios involving network intrusion, an attacker only knows what is perceived through observation of the target network. The intruder is often thousands of miles away from the network to which he or she is attempting to gain entry. Networks often unintentionally provide more information to an attacker than defenders would like. However, the network owner also has the opportunity to reveal information he or she desires the attacker to know, including deceptive information. Because network information is often complex and incomplete, it provides a natural environment in which to imbed deception since, in chaos, there is opportunity. Deception can alter the mindset, confidence, and decision-making process of an attacker, which can have more significant effects than traditional defenses. Furthermore, using deception for defensive purposes gives the defender at least partial control of what an attacker knows. This control sets the stage for defenders to shape attacker reactions for a variety of purposes.

While many different types of cyber deception techniques exist, as described in two recent textbooks on the topic Heckman et al. [2015], Rowe and Rrushi [2016], our research has selected decoys systems as its focus. Industry has started to adopt and advertise these techniques; examples include: Canary Thinkst Applied Research [2019], CyberChaff™ Galois [2019], ShadowPlex Acalvio [2019], Illusive Platform Illusive Networks [2019], DeceptionGrid™ TrapX Security [2019], and, Fidelis Deception® Fidelis Cybersecurity [2019]. However, as far as can be determined, the commercial solutions are missing the critical components of human-subjects testing for evaluation of effectiveness. It is essential to understand attacker behavior and reactionsto best defend against them. This scientific endeavor goes beyond cyber deception by investigating how to play on an attacker's cognitive biases and cognitive load to make attacking systems more difficult. We also examine how to use experimentally collected data to help design an adaptive defensive system.

Traditional cyber security techniques have led to an asymmetric disadvantage for defenders. The defender must detect all possible threats at all times from all attackers and defend all systems against all possible exploitation. In contrast, an attacker only needs to find a single path to defenders' critical information. We believe this asymmetry can be re-balanced using cyber deception to change the attacker's perception of the network environment, and lead attackers to false beliefs about which systems contain critical information or are critical to a defender's computing infrastructure. Our contributions hinge on performing rigorous experimentation.

While the effect and effectiveness of deceptive technologies have been hypothesized for more than a decade, scientifically rigorous studies of the comparative effectiveness of attackers on systems with and without deception are lacking. The goal of our research is to start to fill that gap, and provide a scientific assessment of the effectiveness of decoy systems. Another objective is to examine how to use deception to better understand and influence an attacker that has already infiltrated a network to ultimately delay, deter, and deny an attack. This can be done by taking actions that motivate the attacker to respond in specific ways to enhance the ability to detect, identify, understand, and thus defend against said attacker.

Our pilot studies Ferguson-Walter et al. [2017], the rigorous Tulsarosa Study Ferguson-Walter et al. [2019c] and the collected data we describe in Chaper 2 are a significant contribution to the cyber defense community by furthering our understanding of deception as a defensive tactic. The experimental design and data analysis is a key component of the contribution of this thesis. To our knowledge, it is to date the largest controlled experiment with skilled attackers that held constant the tools and exploits available to the attacker, the network topology and vulnerabilities on the network, and the time participants had to launch attacks. In addition, the amount of data collected during and after the attacks is unprecedented (516 Gigabyte (GB) of data, plus an additional 1537 GB (compressed) of screen capture video). This thesis

provides some results that address the main hypotheses of the experiment, including the effectiveness of decoy systems for cyber defense and the impact of knowledge of deception; and, the scale and breadth of data collected in this controlled study will allow for further future analyses in future work.

There are many research questions of interest that studies such as the Tularosa Study can help address. What elements of an experimental design contribute to the validity of assessing the effectiveness of cyber deception? How effective is cyber deception against skilled attackers? How does an attacker's evasion of detection change based on the presence of cyber deception? How does an attacker's cognitive attributes such as frustration and confusion change in the presence of cyber deception? How does an attacker's success in performing cyber attacks change with the presence of cyber deception? Is just the suspicion of cyber deception being present enough to influence attacker behavior? How does an attacker's evasion of detection change based on the suspicion of cyber deception? How does an attacker's cognitive attributes such as frustration and confusion change in the suspicion of cyber deception? How does an attacker's success in performing cyber attacks change with the suspicion of cyber deception? Is cyber deception more or less effective when the attacker is aware it is being used? How does cyber, versus psychological, deception affect an attackers feeling of frustration, confusion and self-doubt? How do increased levels of frustration, confusion, and self-doubt impact performance of cyber attackers? Are cognitive biases observed in cyber attacker behaviors? Which cognitive biases can be intensified to disrupt cyber attacks? We will focus on addressing a subset of these questions as part of our contributions summarized below. The main focus of this thesis is to assess the effectiveness of a cyber deception technique and its impact on cyber attack behavior and to address whether knowledge concerning use of deception as a defense reduces its impact.

Summary of Contributions

- We perform a literature review and pilot studies Ferguson-Walter et al. [2017] and then design and conduct a rigorous experiment with skilled red teamers to assess the effectiveness of cyber deception Ferguson-Walter et al. [2019c].

- We compare performance on the cyber task between control and experimental conditions. We address the following performance measures: resources expended, number of vulnerabilities identified, successful exploits, ability to evade detection.

- We compare performance on the cyber task between control condition and psychological deception condition.

- We compare performance on the cyber task between conditions where participants were informed about deception to where they were not informed.

- We compare level of emotional and cognitive effects reported between control group and experimental conditions and then compare level of cognitive effects reported to performance on the cyber task across all conditions.

- We catalog the types of cognitive biases observed in the cyber deception experiments, providing corroboration to our theory of oppositional human factors Gutzwiller et al. [2018] aiding cyber defense.

CHAPTER 1

BACKGROUND

Long before computers existed[1], information protection through deception was widely demonstrated. In the 5th century, B.C., Sun Tzu wrote that "all warfare is based on deception" Sun-tzu and Griffith [1963]. Deception is the provision of misinformation that is realistic enough to confuse an adversary's situational awareness and to influence and misdirect the adversary's perceptions and decision processes. The ultimate target of deception is the adversary's mind, and it has been asserted that altering an enemy's perception of reality through defensive deception can potentially level the cyber battleground Climek et al. [2015], Ormrod [2014].

1.1 Related Work

In an extensive review of deception and surprise, Whaley defined the types of deception employed in kinetic military operations throughout the ages, including hiding the real by masking, repackaging, and dazzling and revealing the false by mimicking, inventing, and decoying Whaley [2007]. This taxonomy has begun to be applied to the cyber domain, with the concept of deception for cyber defense recently gaining traction Almeshekah and Spafford [2014], Brzeczko et al. [2014], Heckman et al. [2013], Provos [2004], Vollmer and Manic [2014]. One common technique investigated is the use of honeypots to lure attackers to allow containment and observation of

[1]Sections in this chapter are based on published work: K. J. Ferguson-Walter, D. S. LaFon, and T. B. Shade, *Friend or Faux: Deception for Cyber Defense*, Journal of Information Warfare (JIW), vol. 16, no. 2, p. 28-42, 2017.

them Bringer et al. [2012]. This technique has been explored and expanded upon in many ways Campbell et al. [2015a]. Researchers have also created a technology that presents an adversary with a false network topology Trassare et al. [2013]. Others have created a framework for deception to assist in the analysis of deceptions, whether it involves people or computers, networks of people or computers, or people paired with computers Cohen and Koike [2003]. Honeypot effectiveness has been tested using cyber security games, revealing how different setups may cause attackers to change their operations to avoid negative outcomes Aggarwal et al. [2016]. Research on honeypots have motivated many other types of deceptive techniques that can benefit cyber defense. Honey-patches are similar to decoys, in that they provide a false vulnerability to an attacker, they then automatically redirect an attacker who attempts to exploit that vulnerability to a honeypot Araujo et al. [2014]. Based on the success of honeypots, Rowe et al., made real systems look like honeypots to deter attacks and confuse attackers Rowe et al. [2007]. This idea is a critical one, since there is a limit to what can be done to make fake systems look more real (before they simply become real systems), and working to make real systems look more fake is a good strategy for ensuring it is difficult for attackers to determine which systems hold real value.

Research on honeypots and deceptive content has taken several paths. Honeypots Canali and Balzarotti [2013], Lim, Sze Li Harry [2006] and deceptive content Bowen et al. [2009], Michael et al. [2004] have been shown to attract users on the Internet. Patterns of attacker behavior have been investigated through deploying honeypots on the Internet Nicomette et al. [2011] and in testbeds Rowe et al. [2007]. Previous research has discussed that the configuration of honeypots can either encourage or discourage attacks and identified some relevant design elements Frederick et al. [2012]. Deep learning has been presented as one method of automatic evaluation

of cyber deception techniques Ayoade et al. [2020]. The accumulated research has demonstrated that techniques designed to deceive users are feasible.

To date, our contributions, described in Chapter 2, are the only research to focus on rigorously evaluating the efficacy of decoy systems. Few cyber deception experiments have been executed, and those that have been executed tend not to have rigorous experimental control or a large enough sample size of participants that generalize to the desired population. Participants in studies with larger participant pools typically use unknown parties from the Internet Michael et al. [2004], Nicomette et al. [2011], Wagener et al. [2009].

A honeypot was deployed on the Internet for 419 days to characterize attack behaviors Nicomette et al. [2011]. This interesting study observed over 500,000 SSH connections and were able to identify stage one automated malware versus stage two human attacks based on human characteristics such as typos. While they were successful at better understanding the dictionaries used in dictionary attack in the wild, they noted that the attackers interacting with the honeypot tended to be script kiddies who were not familiar with the Unix access rights and did not delete the history files.

Placing deception on internet-facing network nodes does attain adversarial activity, but it is lacks the internal validity of a controlled study. Additionally, it does not allow for insight into the participants by way of reports or interviews. An alternative strategy is to design controlled studies using students from universities pursuing technology-related degrees Cohen et al. [2001], Rowe et al. [2007]. However, this lacks external validity, as the participant pool does not generalize well for predictive results of sophisticated cyber attackers. Students lack the experience and mindsets that would parallel the sophisticated adversaries these defenses are employed to deceive. In our study, we look to address these issues by utilizing the closest analogous group to malicious cyber adversaries available for scientific testing — red teamers —

and bringing in a larger number of participants in hopes of providing the statistical power and reliability to detect effects.

There are several efforts within the research community working to address the gap in empirical assessments of cyber deception techniques and strategies. There are three main rigorous approaches to this problem: 1) simplifying cyber scenarios such that a non-expert can be used as a reasonable participant; 2) creating realistic models and simulations based on human behavior; 3) conducting studies with skilled participants. There has been effort in abstracting realistic cyber attack environments for experimentation with non-experts. Simulation tools, such as HackIT Aggarwal et al. [2020], which simplifies and simulates network reconnaissance and attack, have been developed and tested with undergraduate students to test the effect of introducing deception at different timing intervals. These computer science undergraduates were found to attack the honeypots more often than real machines. However, there were not enough participants to make significance claims. Examining how to make models more closely resemble real human behavior has also been examined within the context of these simplified cyber games such as the FlipIt game Basak et al. [2018]. In this two-player competitive game, there are resources each player wants to control and they must spend an action (which has a cost) flipping that resource to gain or keep control. For their study, 155 participants were recruited using Amazon Mechanical Turk and asked to fill out the Short Dark Triad personality scale, which scores on the following traits theorized to be relevant to some criminal hacking communities: narcissism, Machiavellian, psychopathy. Their results indicated that are strategic differences between different types of attackers and defenders.

Our work focuses on using skilled participants in controlled experiments. In Chapter 5.1 we investigate how existing Capture-the-Flag exercises compare. Each different type of deception may require a separate experimental design. Recent work describing the Moonraker Study provides a design to assess host-based deception, using com-

puter specialists as the participants pool Shade et al. [2020]. The participants in the study are all unaware of the deception being used for defense, with the design including a training cover story. While the Tularosa Study focused on decoys and includes conditions explicitly made aware of the deception, there are some congruent results between the studies discussed in Section 3.4.

1.2 Decoy Systems

While both are cyber deception techniques used for defensive purposes, decoy systems, which is our research focus, differ from honeypot technology in several critical ways. Traditionally, the main purposes of a honeypot is to draw an attacker away from the true network and gather information about the attacker and the threats he or she poses. Decoy systems tend to be embedded within the true network; and while they can also capture some information (but less than high-interaction honeypots) about attackers who interact with them, this capture of information is not their primary purpose. Their primary purpose is to obfuscate the true network assets and confuse the attacker about the true network topology. Attackers are known to recon networks just to gather an understanding of the infrastructure so as to be better prepared should they want to perform a specific attack on that network in the future. Our red team tests, discussed in Chapter 2, demonstrate that decoy systems confuse attackers and make them unsure or about the true topology.

While honeypots and decoys both lure an attacker by looking more enticing than the real assets, decoy systems can also make the real assets harder to notice by taking up a large share, if not the majority, of the address space. This deception increases the likelihood that an attacker will interact with a decoy and, thereby, trigger an alert. Decoys can also be used to make a homogeneous network appear more heterogeneous, thus making the true attack surface much less obvious. For example, if the network to protect only has servers running Linux RedHat5, decoys can be configured to

not only add more RedHat5 servers, but also some other varieties of Linux, some Windows servers, and a population of client machines. As a result, not only will the amount of assets attackers need to consider be increased, but so will the number of different types of potential attack vectors of which the attackers believe they can take advantage.

Traditionally, honeypots are designed to be isolated from the real network they intend to protect. They appear to contain information or resources of value to the attacker. In this way, honeypots draw and hold an attacker's interest to and within the honeypots themselves, instead of to and within real network assets. With decoy systems, the fake is interspersed with the real, and the decoy assets do not need to be isolated to be effective. These shell assets can be low-fidelity, looking real from the outside — from "far away" as tested by network scanning tools and other red team activity. This is different from high-fidelity honeypots on which attackers' time is wasted once they enter the honeypot because of this, *pocket litter*, detailed information and realistic fake user activity, must be meticulously created and updated by hand, a process that takes both a great deal of time and resources. By design, most decoy systems do not require pocket litter because the intent may not be for an attacker to enter a decoy asset. Attackers are detected just by interacting with the decoy. This can allow for decoy systems to be a lightweight and inexpensive solution that can be easily deployed, configured, and maintained. This ease of deployment and reconfiguration makes this technique a prime candidate for combination with an artificial intelligence system, described further in Section 1.4.

While it is likely to be an uncommon occurrence, there are use cases where a legitimate user might accidentally interact with a decoy. We anticipate that this behavior would look different than an unauthorized hacker interacting with a decoy. For example, a benign user might accidentally mistype an IP address when trying to SSH into a server on which they have permissions. However, we do not expect

the authorized user to attempt multiple usernames and passwords, or use password hacking software against the machine. This network and keystroke behavior will look different between the mistaken and the malicious interaction.

Furthermore, if an attacker determines that they are in a traditional honeypot, they will leave and likely re-attempt to attack the real network. With decoy systems, even if the attacker identifies one decoy or knows it is present, the attacker still needs to put in the work and time to differentiate each asset as real or fake and to take extra precautions not to trigger an alert. Recent work has investigated combining honeypot and decoy systems as well as using artificial intelligence to move towards autonomous cyber deception Al-Shaer et al. [2019].

1.3 Adaptive Cyber Defense

Advanced cyber-defenses need to be able to respond to attacker activity in cyber time—at the same speed as network traffic and cyber attacks. This requires intelligent systems that can automatically react to malicious behavior and evolve their defenses over time as attacks change. The artificial intelligence controlling the defensive system must be able to look ahead and *dynamically* consider how an attacker might behave in the future before taking a defensive action. The concept of adaptive or active cyber defense Denning [2014] — where a system automatically prepares and implements predictive defensive strategies or reacts to detected suspicious activity without human intervention is gaining acceptance but has not yet been widely put into practice. Cyber deception is also an emerging research area in cyber defense. Adaptive cyber deception is a relatively new but inevitable extension of prior work, which cuts across the computer security, behavioral science, and artificial intelligence communities.

There are many reasons why defensive techniques using cyber deception should be adaptive. For example, surprise is one likely important element that can affect the attacker's decision processes and actions. When an attacker experiences unexpected

results, they may decide to change strategies or retry the same techniques, either of which will disrupt or delay their progress, giving defenders more time and opportunity to react appropriately. Static cyber-deception techniques may cause surprise at first, but over time this effect will wear off, as the attackers become familiar with these techniques and learn what to expect from them. If the techniques are adaptive, they will detect when the attacker has developed a response to the deception, and will alter the deception accordingly. Surprise is only one example of how adaptive cyber deception can negatively impact an attacker and disrupt their progress. There are likely to be many more ways to affect an attacker (which we are currently investigating), such as causing frustration, confusion, and self-doubt. These might cause an attacker to increase the number of errors they perform making them easier to detect, delay their attack until further defenses are in place or a critical task is complete, and even deter an attacker from pursuing a particular target all together.

1.4 An Adaptive Decoy System

A 2015 Gartner Report on deception techniques included the following key finding "Deception as an automated responsive mechanism represents a sea change in the capabilities of the future of IT security that product managers or security programs should not take lightly" Pingree [2015]. However in 2019 adaptive cyber-defense systems are still in their infancy, and cyber deception is just a small piece in the cyber defense landscape. We observe both a need to focus on adaptive cyber deception systems and a gap in current research, and have proposed using game theory and reinforcement learning to pursue autonomous cyber deception systems which can

decide when, where, and how to best use deception based on attacker behavior Fugate and Ferguson-Walter [2019].[2]

Our studies discussed in Chapter 2 suggest that decoy systems can be highly effective at disrupting network reconnaissance, confusing an attacker by using their cognitive biases against them, which can then increase the attacker's cognitive load. We theorize that these effects can be multiplied by allowing the decoys to be adaptive to each adversary's specific strategies and preferences. Furthermore, these initial studies indicate that cyber deception may be as or more effective when the attacker is actually informed that there is deception being used on the network for defensive purposes.

Implementing an adaptive cyber-defense strategy in a real-world cyber environment necessitates capabilities that may not be deployed in a typical network. In particular, it requires sensors, actuators, and a means of logically connecting inputs to outputs, making decisions as to how and when to adapt.

1. **Sensors** collect information to detect behavioral-based adversarial activity such as detecting scanning activity and logon attempts. More advanced sensors could detect activity such as the attacker attempting to use stolen passwords and could extend to post-exploitation activities, particularly when network assets contain honey-tokens.

2. **Actuators** take an automated action on the network or host as directed. Actuation of decoys involve configuration changes, creating new decoys, changing decoy parameters, modifying service banners, and other deceptive activities. Further decoy adaptations could include changing the IP address, opening or closing ports, adding or removing services, or even spoofing a different operating

[2]This section is based on published work: Fugate, S. and Ferguson-Walter, K. *Artificial Intelligence and Game Theory Models for Defending Critical Networks with Cyber Deception.* AI Magazine, Spring 2019.

system. Not only are these specialized tasks not normally managed by modern enterprise network management tools, but these tasks must be automated to rapidly respond to suspicious activity.

Furthermore, cyber deception techniques can be used to do more than delay, confuse and surprise an attacker. Cyber deception can be used to influence the attack in more direct ways. For example, the defender may want to learn something specific about an attacker or collect information about a specific type of attack. Deception can be used to entice or convince an attacker to take an action that, unknown to the attacker, actually benefits the defender in some way. This is important for cyber defenders, since as we move forward into more adaptive cyber-defensive systems, we must consider the natural co-evolution of multi-step, multi-stage attack/defense situations. These advanced defenses must take a strategic view, where moves are considered many steps ahead of both attacker and defender actions; this has been referred to as cyber co-evolution Willard [2014].

1.4.1 Perceived Payoffs and Automated Strategy Selection

We have begun to investigate both game theory Roy et al. [2010] and reinforcement learning Sutton and Barto [1998] techniques which may provide a good solution when creating an adaptive decoy system Bilinski et al. [2019]. Cyber deception is a powerful tool for defenders because it allows them to manipulate the game-board which has traditionally only been a possibility for attackers. However, as the owners of the network, cyber defenders should be able to control the information the network distributes and potentially change the way the network behaves. In our estimation, this type of game manipulation is able to give the defender an asymmetric advantage over an attacker. The game-board can be manipulated in several ways which can have various effects on the attacker.

By changing the game-board, as the attacker perceives it, the defender is able to limit the strategies available. If the attacker has the wrong information about a system, the strategies they think are applicable to attack will likely fail. One major advantage that cyber deception provides to a defender is the ability to change the perceived payoff of a set of actions to the attacker. Each player is selecting actions trying to maximize a long-term payoff. The payoff is an estimation of how good or bad the outcome is for that player. Recall that many game theory games are structured as zero sum where the payoffs for each outcome add up to zero across the players.

Since the defender can control the information the attacker uses to make his decisions (and form his game tree), the defender can manipulate the payoffs the attacker associates with certain paths. For example, a defender can make a system look more vulnerable or more interesting. This will cause the attacker's perceived payoff for that machine to be much higher than the true payoff. Furthermore, if the defender is using decoys or honeypots, the attacker's perceived payoff may be very high, while the true payoff is instead very high for the defender. This negative true payoff for the attacker is due to the time and energy wasted on a fake system.

For a defender to make wise decisions about how to best protect their network and systems, there are several useful things they need to know. First and foremost, the defender will be more effective if they know when they are being attacked. They can use proactive defenses including pre-set cyber deception techniques, but the effect will be greater if they can also adapt those defenses based on details of a current attack in real time. In addition to knowing an attack is occurring, details about the attacker and their actions will help in the defense. Learning preferences of the attacker, e.g., they tend to attack Linux machines; attacker attitudes, e.g., they are noisy and not cautiously avoiding detection; and patterns of behavior, e.g., the attacks occur at certain times of day, will aid the defender in customizing strategies to adapt the game-board and launch the optimal cyber deception.

Since skilled cyber defenders are outnumbered by attacks, and manual responses are too slow, adaptive defense systems are a logical technical next step. However, the artificial intelligence that fuels these kinds of control systems are only as effective as the feedback available such that the system can learn which automated responses had good outcomes and which did not. Understanding what a realistic payoff or reward function should be is a critical research question that needs to be addressed. Future work is needed to apply the results from our Tularosa Study and propose realistic payoffs from which an adaptive deception system can learn.

CHAPTER 2

EXPERIMENTAL DESIGN AND METHODOLOGY

To date, there is little experimental evidence of how effective cyber deception can be or how it may compare to other defenses. It is important for the community to investigate and understand the usage of cyber deception and the effects it can have to better protect information systems. To begin to measure the (positive) impact on cyber defenders, and the (negative) impact on cyber attackers, we have focused on decoys systems and conducted a rigorous human subjects experiment.

2.1 Cyber Deception Using Decoy Systems: Pilot Studies

This section will discuss results from initial pilot studies [1], whose results motivated the rigorous study discussed in later sections. The goal of these pilot studies were to evaluate a decoy system which provides realistic, lightweight decoys on a real network to maximize the chance of an attacker being detected and mitigated quickly, as well as delaying and disrupting an attacker's forward progress. Several such systems are available commercially, e.g., Canary, CyberChaff, Deception 2.0, DeceptionGrid, Deceptions Everywhere, DECOYnet, and Threatstream, but the concept is always the same — the large number of false assets provides an asymmetric advantage for defenders by: 1) Reducing the chance a real asset will be attacked; 2) Distracting an attacker from real assets and content; and 3) Forcing attackers to take additional actions, thus slowing them down and increasing the likelihood of revealing themselves.

[1] This section is based on published work: K. J. Ferguson-Walter, D. S. LaFon, and T. B. Shade, *Friend or Faux: Deception for Cyber Defense*, Journal of Information Warfare (JIW), vol. 16, no. 2, p. 28-42, 2017.

We completed a series of four pilot studies which suggest that such a decoy system can be highly effective at disrupting reconnaissance, confusing attackers by leveraging their biases against them, causing attackers self-doubt, and increasing attackers' cognitive loads. In tests where the attacker is made aware of the deception techniques, results indicate that the defensive effectiveness remains while confusion and paranoia are actually increased. These pilot studies cannot be a substitute for full rigorous experimentation. However, the pilot studies' results, as described below, offered scientific insight into the effects of deception for cyber defense and lay the groundwork for the scientifically rigorous Tularosa study.

As a first effort to apply scientific principles to security research and provide empirical evidence of the effectiveness of decoy system technology, we designed and conducted a series of pilot red-team exercises following guidance from the relevant institutional human-subject research review boards. The value of these exercises is clear: there is a need to measure human reactions, not just network behavior. In these pilot studies, a three-subject red team was given basic instructions to recon the network, define its topology, and locate and exfiltrate content of interest. Since the environment was a real operational network with real users and availability concerns, the team was also instructed that no distributed denial of service (DDOS) attacks or other attacks that could disrupt the network were allowed. The subjects were asked to think out loud to explain their actions, thoughts, and strategies, all of which were recorded. A semi-structured interview questionnaire was given at the end of the day.

A data collector was assigned to each red team member to monitor responses and ask questions for clarification. The collectors were monitored to ensure adequate standards of administration were maintained. The research team received training on how to collect data from the human subjects in a standardized manner. The purpose of the training was to identify and reduce bias and primacy effects in data collection. This helped prevent the collectors from asking leading questions and from

allowing the subjects to veer off track; the training also facilitated the maintenance of continuity throughout the study.

For data analysis purposes, a questionnaire was completed by the subjects, notes were taken by the observers, the audio of subjects and observers for all experiments were recorded, command capture of each subject on his or her laptop was obtained, and alerting results from the decoy system were collected as was network capture on the subnet. All four pilot studies were conducted on the same network in order to measure the effectiveness as a defensive tool as well as the effect of deception on an attacker. The same three professional red team members were used in all four pilot studies, which were held roughly six months apart on the same network.

2.1.1 Pilot Study One: Cyber Unaware (Present-Uninformed)

. The goal of Study One was to evaluate whether low-fidelity decoys were robust enough to deceive and disrupt a red team who was unaware of the deception. This was a two-day exercise. On the first day, 71 decoy assets were present on a network of 310 real assets. On the second day, decoys were not deployed. The subjects were not made aware of the purpose of the experiment.

On day one (with 81% real network nodes and 19% decoy assets), the subjects spent at least 38% of their time attempting to penetrate decoy assets. The decoy systems detected and alerted on all scanning and exploit activity related to the decoys. The subjects verbally expressed confusion during the scenario and questioned their tools, their skills, and themselves rather than the authenticity of the network. For example, one subject stated, "Are they connecting... No. ... I'm going to do it by hand".

The decoys were hand designed to span the spectrum of assets that might interest an attacker, from looking more vulnerable than true assets to looking like the most hardened and secure assets. In Pilot Study One, the subjects first investigated and

attacked what they considered to be the most vulnerable looking assets — the decoys. In fact, 83% of the total exploits launched were against decoy assets.

On day two (with 100% real network assets), the subjects briefly investigated the disappearance of decoy assets, which they thought to be real. During the post-experiment interview, the subjects stated their belief that day one was the control experiment and that in day two a new defensive "masking" capability had been added. Until the post-experiment debrief, the subjects believed the decoys assets were legitimate.

This pilot study supported the hypothesis and indicated that additional benefits of decoy systems may include:

- Improved detection of attacker activity through monitoring of scanning activity and interaction with illegitimate network assets;
- Slowing the adversary by increasing both the number of types of network assets and the complexity of the network;
- Encouraging changes in attacker behavior to address the increased scope, variety, and complexity of the network;
- Causing attacker confusion about the network as well as the exploit activities/successes. If an attacker does identify fakes assets on the network, he or she might conclude the whole network is not real and leave, thereby protecting the network.

2.1.2 Pilot Study Two: Cyber Aware (Present-Informed)

. It is often assumed that deceptive techniques can only be effective if a subject is unaware of the deception. Pilot Study Two was designed to examine whether decoys would still be effective if the red-team subjects were aware of the decoys' presence on the network. Thus, the hypothesis was as follows: using decoys to populate an existing

network with decoy machines, routers, printers, and other devices will influence the subject who is aware of the deception in the same manner as Study One's hypothesis.

After Study One, the subjects were briefed on the existence of decoy systems and the methodology of the deception without technical details. Study Two was set up to replicate Study One, except the subjects were told to expect decoys on the network.

During Study Two, it was noted that subjects continued to fault their arsenal of tools as well as their own skills when they did not achieve their goals. Additionally, it was noted that they spent a significant amount of time attempting to decide for themselves what was "false" on the network and what was "real". The effort spent on determining the veracity of the network nodes slowed down their intended goals to infiltrate and attempt exfiltration from the secure network. In Study Two, when the subjects were aware of the deception, they spent much more time before doing anything outside the landing point provided. This means they were slower to move and attack.

To measure the increased uncertainty or incorrect belief, subjects were asked to label which assets were real, which were fake, and which were suspicious. Thirteen percent of the decoys were identified as fake and 4% as real. Ten percent of real machines were identified as fake, and 2% as suspicious. This provides evidence of increased uncertainty since, without decoys deployed, the subjects would be certain that every machine they touched was a real machine. Additionally, the subjects attributed "strangeness" in the network to deception, even when deception was not involved. As one subject voiced, "Those names sound fishy to me.... I'm not taking any of the users ... because I don't trust this right now".

Many benefits of the deception being known were noted. For example, when the subjects were unaware of the deception in Study One, they focused on attacking the most vulnerable looking assets on the network. However, when the deception was known in Study Two, the subjects avoided the most vulnerable looking machines,

assuming they were put there to deceive them when, in fact, they were real. One subject stated, "So ... legacy OS's, right? What would be the only reason to have one on the network?... Honeynet."

In essence, the lack of deception awareness in Study One leveraged the subjects' confirmation bias in that they were expecting all nodes on the network to be real. So, when presented with decoys, subjects assumed they were real. In Study Two, this same confirmation bias was exploited in that subjects expected unusual stimuli in the network environment to confirm the presence of fake assets — even when they were not fake; these stimuli confirmed to the subjects an indicator of falseness. Both conditions leverage confirmation bias to the defenders' advantage. Furthermore, an awareness of deception, as in Study Two, adds to the cognitive load of the adversary in that he or she now has to evaluate each interaction with the network to assess its authenticity. The initial findings suggest there is great usefulness in allowing an intruder to be aware of deception on the network to help protect true vulnerable assets; however, more investigation is required to understand the full significance of this effect.

2.1.3 Pilot Study Three: Psychological Deception (Absent-Informed)

. In Study Three, the adversary's awareness of deception on the network was further investigated by examining if simply a belief that deception was present could provide some benefit to defenders. It was hypothesized that simply the subjects' belief that a decoy system was on the network would influence the subjects in the same manner as Study Two, even when no deceptive technology was deployed on the network.

The subjects made a variety of interesting observations including many attempts to assess the veracity of the various assets. Additionally, cognitive biases surfaced which convinced subjects that certain nodes absolutely looked fake. As one subject

commented, "I know that one is fake, so I am not going to bother with it". Post-exercise, the subjects were debriefed to assess what about the network landscape conveyed to them that certain nodes were fake and others real.

The purpose of Study Three was to assess the effect of the subjects' belief that deception was deployed on the network, even when it was not. A significant amount of the subjects' time was spent focused on their beliefs that certain nodes were intended to spoof them. Study Three confirmed that it might be enough for the adversary to simply believe a deceptive defensive tool is deployed on a secure network to delay, impede, and dissuade malicious activity. Further research is required on this topic.

2.1.4 Discussion of Pilot Studies

Pilot Study Four investigated whether the decoy system would have any effect if the subjects knew the technical details of how it worked. This study did not have bearing on the Tularosa study, and so will not be described here.

Based on post-experiment, semi-structured interview questionnaire results, observable data collected indicate that during Pilot Study One, subjects first questioned their tools and techniques before questioning the validity of the network. In Pilot Studies Two and Three, when the subjects were expecting deception, they spent a large amount of time trying to differentiate real from fake, greatly delaying and obstructing their exploit and exfiltrate goals. It was noted that the selection of targets for exploitation was affected and delayed by knowledge of the presence of the decoy system, and attack strategies were also changed. As noted by one subject, "My perspective is that anything that looks like an exploitable target in the network is not a valid target". One subject consistently and incorrectly indicated confidence in identification of decoys, while the other subjects indicated confusion and noted that the decoys were harder to identify than expected. The subjects' confidence levels were noted as extremely high when assessing a real asset as fake. These preliminary

findings suggested that more rigorous scientific design would be useful in quantifying subjects' confusion, perceptions, and confidence.

For future experiments, it is crucial to continue to include observation of the subjects to understand any changes/effects such as frustration or confusion. However, in real-world applications of cyber defense, such access to the attacker, which would allow the researcher to observe emotional/physical reactions to defensive interventions, is seldom possible. What can be collected is the behavior on the computer/network. A behavioral mapping effort is critical to be able to infer human reactions of the operators from data collected on the computer. These inferences will not hold in all cases or for all people; but with enough rigorously collected experimental data from which to draw the inferences, these inferences could be extremely useful in determining whether real-world defenses are having the desired effect on a real attacker.

While decoy systems are available for purchase from several cyber security vendors, currently, commercial entities are hesitant to perform additional testing and evaluation of their products for several reasons:

- This level of examination can take a great deal of time, money, and resources;
- There may be concern that the results might suggest their product works poorly when upheld to rigorous assessment;
- There may be fear that even if their product works well, putting an exact measurement of how well it works may make their product appear less competitive compared to products whose companies are free to make marketing claims without research to support their claims.

2.2 Cyber Deception Using Decoy Systems: Tularosa Study

Based on the Pilot Studies described above, we created a large-scale rigorous experiment[2] by designing a network penetration testing exercise on a simulated computer network with different conditions examining how participant performance changes if a cyber deception techniques is used and whether the participant is aware of the deception. We also sought to examine the attackers' cognitive, emotional, and physiological responses to the defenses, which is the human subjects portion of the study. This information was collected to better understand when the deception is effective (using more than just self-reports by the participants) and better correlate the impact on the human (e.g., stress, confusion, frustration) with the technical data captured during the penetration testing task on the network and the client.

Our Tularosa research study Ferguson-Walter et al. [2019c] (held at Sandia National Laboratory, and named after the New Mexico town of Tula Rosa) included 139 professional industry red teamers and in addition to the cyber task data collected, the design used a range of personality indices, physiological measures, and cognitive tests to understand attackers' mental models, decisions, and behaviors. Each index, measure, and test was hypothesized to correlate with performance on the network penetration task in the presence of deceptive defenses as described in Section 2.2.1.4. In many cases, our cognitive tests do not correspond to information that is directly available about attackers on an individual basis in a real-world setting. However, understanding the correlations between these factors and task performance may allow us to categorize attackers and mount a tailored response.

---

[2]This section is based on published work: K. Ferguson-Walter, T. Shade, A. Rogers, E. Niedbala, M. Trumbo, K. Nauer, K. Divis, A. Jones, A. Combs, R. Abbott, *The Tularosa Study: An Experimental Design and Implementation to Quantify the Effectiveness of Cyber Deception*, 52nd IEEE Hawaii International Conference on Systems Science (HICSS), 2019.

Two Days of Scenario-Based Penetration Testing

	Day 1 Conditions **(between participants)**	***Day 2 Conditions*** ***(*****within participants; changes from Day 1 in bold*****)***
Participants Split into Four Groups: A-D	A1 No decoys; Not Informed ("Control")	A2 **DECOYS ADDED**; Still not informed
	B1 No decoys; Informed ("Psychological")	B2 Still no decoys; **NOT INFORMED**
	C1 Decoys present; Not Informed ("Cyber Unaware")	C2 **DECOYS REMOVED**; Still not informed
	D1 Decoys present; Informed ("Cyber Aware")	D2 **DECOYS REMOVED, NOT INFORMED**
Planned Comparisons	***Comparisons Between Groups: Day 1***	***Comparisons Within Group: Day 1 vs Day 2***
	• A vs B: Whether information matters in the absence of decoys • A vs C: Whether decoy presence matters in the absence of information • B vs D: Whether decoy presence matters when information on deception is given • C vs D: Whether information on deception matters when decoys are present	• A1 vs A2: What happens when a decoy is added (without any other indication deception techniques may be in play)? • B1 vs B2: After being previously told deception techniques may be in play, do these beliefs and behaviors persist when the warning is not given a second time? • C1 vs C2: Does the effect of the decoys disappear when the decoys are no longer present, or does it persist without any indication deception techniques may be in play? • D1 vs D2: Does the effect of the decoys disappear when the decoys are no longer present, and when indications of deception techniques are no longer given?

Figure 2.1: Experimental conditions. Each day, a decoy system was either present or absent on the network. Participants were either informed or not informed that cyber deception tools might be present on the network.

2.2.1 Design

Subjects participated in a network penetration task. Modeled after our pilot studies Ferguson-Walter et al. [2017], we altered cyber deception, i.e., decoy system present versus decoy system absent, and explicit mention of deception i.e., informed versus not informed, between each group of participants on the first day of testing. We also manipulated within participants between the first and second days of testing. Figure 2.1 contains further details on these conditions. In addition, participants completed a series of questionnaires and cognitive tasks (see Section 2.2.1.4).

2.2.1.1 Conditions

We pseudo-randomly assigned participants to one of the four conditions (see Figure 2.1). Participants were run in cohorts of varying numbers with all members of a given cohort assigned to the same condition when possible, but with some cohorts mixing conditions to achieve a balanced number of participants between conditions. Each day, a decoy system was either present or absent on the network. Participants were either informed or uniformed that cyber deception tools might be present on the network. The influence of decoy presence and explicit mention of deception can be measured between participants using performance on Day 1. Persistence of effects

within participants, when the environment changes, can be measured by comparing Day 1 to Day 2 within a condition. There is no control condition on Day 2 since once participants have been exposed to the task, even with a new network, it is no longer an independent control.

- **Condition A: Day 1 No Deception (Control); Day 2 Cyber Deception.** Condition A provides a separate control condition to allow comparison between groups of participants experiencing cyber or psychological deception and those who do not. We also employed comparison within groups by having subjects encounter a different type of network on the second day. The topic addressed by adding the decoy system on Day 2 is the effect of encountering a deceptive network following exploration of a similar non-deceptive network. This might occur if an organization uses decoys only on a portion of their network.

- **Condition B: Day 1 Psychological Deception; Day 2 No Deception.** Condition B is designed to research the effect of (between-subjects relative to the control on Day 1) and persistence of (within-subjects across Day 1 and Day 2) psychological deceptive influence. The decoy system is not included on the network on either day. On Day 1, participants are informed that deceptive defense techniques may be in use on the network. On Day 2, participants are not explicitly told whether deceptive defense techniques are in use. This condition aims to discover what the effect is if an attacker assumes that deception is on the network due to acquired information and then moves to a new network with no cyber deception.

- **Condition C: Day 1 Cyber Deception; Day 2 No Deception.** Condition C allows us to examine the effect of (between-subjects relative to the control on Day 1) and persistence of (within-subjects across Day 1 and Day 2) cyber deceptive influence without explicit mention of deception. On both days, par-

ticipants are not informed of the use of deceptive defense decoy techniques. On Day 1, decoys are present on the network; on Day 2 they are not. This condition aims to discover what the effect is if an attacker experiences cyber deception on one network and then moves to a new network.

- **Condition D: Day 1 Informed Cyber Deception; Day 2 No Deception.** Condition D is similar to Condition C, except it addresses the effect of and persistence of cyber deceptive influence with prior awareness. A defensive decoy system is present on the network on Day 1, and participants are informed that deception might be present on the network. On Day 2, the decoys are no longer present on the network. Participants are not explicitly told whether deceptive defense techniques are still in use. This condition aims to discover what the effect is if an attacker is aware of and experiences cyber deception on a network and then moves to a new network without cyber deception.

2.2.1.2 Cyber Range

The Cyber Range was designed to emulate a semi-realistic enterprise network. Each participant's environment was also instrumented so that data could be collected during the study yet not be intrusive for the participants. The simulated enterprise environment consisted of a number of servers and workstations running both Microsoft Windows and Linux. A wide spectrum of operating system versions were installed. Microsoft Windows instances included various versions of clients and servers: Windows 7 through Windows 10 clients and Windows Server 2008 through Server 2016. Linux systems were comprised of Ubuntu 14.04 and 16.04 Desktop and Server. See Figure 2.2 for further details.

Active Directory services were installed on a Windows Server 2016 Enterprise system to emulate a typical corporate controlled authentication system. A DNS was also installed to provide name services for all of the clients and servers in the network.

Operating System Version	Instances	Notes
Windows 7 (SP0 and SP1)	4	Most vulnerable clients
Windows 8.1	4	
Windows 10	11	
Windows Server 2008	3	Various services
Windows Server 2012	1	Web server
Windows Server 2016	2	Domain Controllers (130 domain users)
Ubuntu 14.04 Desktop	19	
Ubuntu 14.04 Server	4	
Ubuntu 16.04 Server	2	

Figure 2.2: Distribution of Windows and Linux systems.

There were also a number of other common services provided on the network such as web servers, database servers, file servers. A total of 50 systems were installed in the environment with an even split of 25 each for each operating system—Windows and Linux. Twelve of the 50 were servers (6 Windows servers and 6 Linux servers). In the cyber deception conditions (C1, D1, and A2), there were 50 decoys in addition to these systems.

A Network Time Protocol (NTP) server was configured and installed within each participant's environment to provide time synchronization of all of the machines, allowing for reliable timestamping for data collection. The NTP server was designated as out of scope for the participants since this was part of the experimental support infrastructure and would jeopardize the data collection effort if attacked.

To provide a more realistic environment, we created 130 domain user accounts to provide a lived-in network appearance for the participants. Of the 130 domain user accounts, there were 15 domain administrators to simulate the IT staff for the fictional organization. Separate Organizational Units (OUs) were created to simulate actual business organizations, e.g., IT, Sales, HR, Staff, since it is common to create OUs to mirror primary business functions. A number of files were also created and placed in the user accounts and log files were populated by having staff members

perform activity on the network during the creation phase. For example, several domain administrator accounts were selected and used during this phase to generate log activity that would appear to be authentic. Each domain user account had a password that met the minimum-security requirements. Each Windows system was then joined to the domain to allow authentication services to work properly within the network. The IP addresses were randomized within a typical Class C subnet and MAC addresses were generated to represent typical vendor, e.g., Dell, HP, Intel.

The decoys used as the cyber deception component in this experiment were based on lightweight virtualization and were configured to replicate operating system and services of typical assets residing in an enterprise network. The decoys were configured to mimic both Linux and Windows services similar to those in the Cyber Range. These decoys respond to typical network port scans and provide almost identical feedback to those of real desktops. Since actual services were not running on these decoys, any attempt by a subject to logon failed and was logged as an indicator of unauthorized activity. Examples of some of the services are Apache web server, DNS, SSH, and FTP. Separate environments were designed to facilitate experimental conditions with and without decoys. With exception of the presence or absence of decoys, the environments were designed to be as similar as possible to allow for easy comparative performance analysis.

We provide each subject with a laptop to use during the experiment that was connected to the cyber range via a dedicated network. These laptops were configured with Kali Linux which provides a robust environment for penetration testers with over 600 security-related tools. Some of the most commonly used tools in this distribution are Nmap (port scanner), Metasploit Framework (penetration testing), and ZAP web application security scanner. In addition, the laptops were configured with their own offline Kali Linux repository with 65 Gigabytes (GB) of binary packages that include additional tools and software that could be easily installed by the subjects. The use of

the offline repository enabled us to disconnect the laptops from the internet (ensuring no PII was accidentally collected) while still enabling the subject to install additional software if needed throughout the study.

2.2.1.3 Cyber Data.

We collected several data sources from the participants' attack clients during the study. *Netflow* and *tcpdump* recorded full packet capture from their machines for post-experiment review of their network activity. A keylogger and video screen capture were used for the duration of the experiment to record their host-based operations. Participants were encouraged to keep a running log of findings via a Mattermost chat client during the experiment, giving real-time insight into what parts of their activities they thought were notable as they experienced them. Additionally, we retrieved data from the participants' laptops after the experiment was over. Several logs from the Kali Linux operating system were collected, including logs of the processes run, the system notifications, daemon logs, authentication records, and default package logs. The shells used by the participants had their history aggregated to reveal commands entered. All notes stored by the participants on the attack client were collected as well. If deception was present in their environment, we also collected the logs server-side from the decoy system that tracked instances of the decoys being triggered. These logs tracked four primary interactions with the decoys: single packets to a single host (touch), multiple packets to a single host (probe), single packets to multiple hosts in succession (scan), and interactive login attempts (intrusion).

2.2.1.4 Individual Measures

In addition to the network penetration task, participants completed a series of questionnaires and cognitive tasks. This section highlights the tasks selected and justification for their inclusion. See supplemental materials in the appendices for more

details including the cyber task instructions (Appendix B), full schedule (Appendix C), and all questionnaires (Appendix A).

- **Task-Specific Questions.** We designed three sets of questions to measure participants' experiences during the experiment. The questions provided a data stream on task performance in addition to data collected directly during the network penetration task. These questions included a daily briefing consisting of open-ended questions about participants' experiences during the network penetration test, with participants in the informed condition explicitly asked about "the nature of deception on the network, if found". On Day 2 participants were asked about their experience across both days and to rate tools available to them and their prior knowledge. In addition, each day participants were given a Cyber Task Questionnaire (CTQ) in which they were asked to rate and explain the level of confusion, self-doubt, confidence, surprise, and frustration they felt during the cyber exercise, with the Day 2 version including a question about belief in the presence of deception on the network.

- **General Questions about the Individual**. We designed questions to measure general information about an individual such as their demographic information and cyber security experience. These items are of particular interest because they may help diagnose whether given effects found in the data set are due to the experimental manipulation or a particular individual's background (even given random assignment to condition). They could also help explain the factors relevant to particular performance characteristics, e.g., initial moves of a participant with over twenty years of experience versus two years of experience. We also asked participants who experienced cyber deception to complete a questionnaire designed to assess their responses to deception in a network penetration context.

- **Cognitive Battery**. General cognitive ability ()i.e., I.Q.) is traditionally the best predictor of individual job performance across job categories and situations Ree and Earles [1992], Schmidt [2002]. Measurement of additional, specific cognitive abilities may provide additional predictive value in the context of particular jobs, reflecting the specific processing required in these domains. This includes circumstances in which initial selection on general cognitive ability already occurs as part of an employment screening process Lubinski [2000]. Furthermore, non-cognitive attributes, e.g., personality characteristics, may provide additional predictive power Schmidt and Hunter [1998]. Therefore, the battery for this study includes a number of tasks and questionnaires that go beyond general cognitive ability to allow a more comprehensive understanding of the abilities and attributes that are thought to characterize red teamers or be predictive of performance in the domain of network penetration (e.g, Campbell et al. [2015b], Egelman and Peer [2015], Summers et al. [2013]).

  Cognitive tasks included the Shipley-2 Shipley et al. [2009] as a measure of overall cognitive ability, the Sandia Progressive Matrices (SPM; Matzen et al. [2010]) as a measure of fluid intelligence, i.e., those aspects of intelligence that allow for adaptive reasoning and problem solving, the Over-Claiming Questionnaire (OCQ; Paulhus et al. [2003]) as a measure of ability to distinguish real from fictional items and decision-making confidence, the Operation Span (O-Span; Unsworth et al. [2005]) task as a measure of working memory, i.e., ability to maintain information in memory and inhibit distractors, the Remote Associates Task (RAT; Cropley [2006]) as a measure of convergent creative thinking, i.e., generating atypical links between concepts to generate a solution to a problem, and a set of insight and analytical problems to solve Wieth and Burns [2006]to assess proficiency at generating incremental solutions (analytical problems) and at reframing problems and approaching them from different

perspectives (insight problems). Personality assessments included the Big Five Inventory (BFI; John and Srivastava [1999]) as a measure of openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism, the General Decision-Making Style Inventory (GDMSI; Scott and Bruce [1995]) as an indicator of the way in which individuals approach and make decisions, the Indecisiveness Scale (IS; Rassin et al. [2007]) to determine if participants tend toward intuitive speeded decisions or toward gathering as much information as possible, and the Need for Cognition (NfC; Cacioppo et al. [1984]) as measure of individuals' tendencies to pursue and enjoy the process of thinking.

We also asked participants to complete the Karolinska Sleep Diary (KSD; Åkerstedt et al. [1994]) to assess sleep quality for the night prior to administration, as some participants were required to travel prior to participation and may have experienced sleep disturbances which could impact task performance.

- **Physiological.** We collected physiological data using Empatica E4 wrist-based devices. The Empatica E4 collects heart rate information (including heart rate variability via blood volume pulse), motion-based activity (accelerometer), peripheral skin temperature (infrared thermopile), and galvanic skin response (electrodermal activity sensor). Physiological signals like these have been used to characterize and predict cognitive and physical states in a variety of settings. Analysis of the physiological data is beyond the scope of this thesis.

2.2.2 Implementation

We received an approval on the experimental design from all relevant institutional ethics review boards (IRB)[3]. No personal identifying information (PII) was collected

[3]The IRB determined that the portion of the tasks that aligned with normal red team activity are not human subjects research (HSR) and thus could be included in contracted work. However, the portion that collected data about the participants, their cognition, and their physiology is HSR and thus was completely voluntary.

and all experimental data was anonymized with subject IDs. No cyber task performance or HSR information was provided back to any of the participants' employers.

2.2.2.1 Participants

Prior to traveling to our site for the two-day study, participants were provided the statement of work which indicated that they would be participating in a series of capture the flag style events to measure the effectiveness of defensive software on a simulated network. Participants could request, in advance, software tools, reference information, e.g., technical documents, and other computer files they wished to be included for the event. This ensured participants would have access to preferred tools and that all participants had access to the same tools. They were aware their actions would be monitored during the task and that they would be asked to complete a series of reports and questionnaires. Finally, they were informed that they would be required to sign a nondisclosure agreement stating they would not reveal information about the task, the network vulnerabilities, and the defenses encountered (including to other participants).

Upon arriving to the study, we asked participants whether they would also like to be part of a human subjects research study (HSR) as part of the cyber exercise. Those who opted in provided physiological and cognitive data in addition to the network penetration task and task-specific questionnaires; they were offered a $25 Amazon gift card for their participation. Six participants did not volunteer for the HSR portion. Those who opted out wrote an extended red team report, such that participants spent the same amount of time in the study regardless of the decision they made.

Data was collected on 138 professional red teamers, 132 of whom agreed to participate in the HSR portion of the study. For details on the demographics of the participants see Figure 2.3. The vast majority of our participants were male with English as their primary language. Most were under 35 years old and had a bachelor's

Measure	Level	%
Age	Less than 35 years	52
	35 - 50	37
	Over 50 years	11
Education	High School	11
	Associates/Tech	23
	Bachelors	43.3
	Masters	22
	PhD	0.7
Gender	Male	94
	Female	6
Langage	English	95
	Other	5

Figure 2.3: Demographics Information

degree as their highest level of education. Responses on the experience questionnaire indicated that participants were fairly evenly split between Linux and Windows users, although some chose to write-in Mac or a combination of operating systems. Most tended to work in groups of two to three people for engagements that last one to two weeks. However, there was substantial variance in these responses. For further details see Figure 2.4. The participants indicated the highest level of their expertise and years of experience in cyber security, network reconnaissance, and generalized defense practice (Figure 2.5). This is the skill set most necessary for the cyber task presented in the Tularosa study.

As Figure 2.6 shows, over half of participants indicated their typical teams had expertise in network reconnaissance, network penetration, host penetration, generalized defense practice, and incidence response, with the network reconnaissance and penetration categories being identified by over 80% of responders. Once again, reverse engineering was least common, with less than 30% of participants indicating expertise in those areas on their teams (however over 60% had access to experts in that field).

Question	Answer	Count	% of Responses
Operating system	*Linux*	55	44
	Windows	45	36
	Mac	9	7.2
	(Combination)	16	12.8
Team size	*Individually*	25	19.5
	2-3 people	56	43.8
	4 or more people	35	27.3
	(Other)	12	9.4
Duration of engagement	*1-2 days*	15	12
	3-days - 1-week	18	14.3
	1-2 weeks	28	22.2
	2-weeks - 1-month	23	18.3
	Over 1 month	22	17.5
	(Other)	20	15.9

Figure 2.4: Experience Questionnaire: operating system, team size, and typical duration of engagement.

		Rating		
Question	**Sub-Category**	**N**	**Mean**	**Stdev**
Level of expertise (1 = novice, 5 = expert)	*Cyber security*	128	3.64	0.93
	Network penetration		2.92	1.08
	Host penetration	128	2.93	1.1
	Network reconnaissance	128	3.39	1.12
	Incidence response	128	2.79	1.15
	Generalized defense practice	127	3.38	1.16
	Network protocol reverse engineering	128	2.02	1.05
	Binary reverse engineering	128	1.77	0.99
Involvement in each phase of engagement (1 = least, 5 = most)	*Reconnaissance*	128	3.38	1.35
	Weaponization	129	2.74	1.36
	Delivery of weaponized bundle	129	2.69	1.36
	Exploitation	128	3	1.33
	Installation of malware	126	2.73	1.38
	Command and control channel for remote manipulation	128	2.78	1.44
	Actions on objectives	126	3.27	1.32
Match to typical engagement (1 = least, 5 = most)	*Compliance testing*	129	3.02	1.52
	Blue team training	129	2.53	1.35
	Demonstrate the needs for increased security investments	125	3.32	1.33
	Whiteboarding/gaming/tablet op exercises	129	2.65	1.27
	Post-attack remediation effort	128	2.84	1.26
	Vulnerability analysis	128	2.62	1.32
	Security architecture review	129	3.27	1.28
	Persistent adversary emulation	129	2.59	1.46
experience	*Cyber security*	128	7.87	5.61
	Network penetration	128	4.26	3.91
	Host penetration	128	4.11	3.74
	Network reconnaissance	128	5.04	4.06
	Incidence response	126	3.79	4.29
	Generalized defense practice	129	6.9	6.27
	Network protocol reverse engineering	128	1.82	2.67
	Binary reverse engineering	128	1.51	2.24

Figure 2.5: Experience Questionnaire responses: level of expertise, involvement in each phase of engagement, typical engagement, and years of experience.

Expertise Category	Team		Access	
	Count	*Percentage*	*Count*	*Percentage*
Network penetration	111	88.1	105	83.3
Host penetration	102	81	104	82.5
Network reconnaissance	115	91.3	103	81.7
Incidence response	64	51	95	75.4
Generalized defense practice	91	72.2	95	75.4
Network protocol reverse engineering	38	30.2	84	67
Binary reverse engineering	34	27	79	62.7

Figure 2.6: Experience Questionnaire: Team composition and expert access. Expert access was calculated as responding positively to either the team or access questions.

2.2.2.2 Cognitive Battery/Personality Assessment Findings

Following the cyber task on each day, participants completed a number of cognitive tasks and personality assessments. This battery was designed to both assist in characterizing red teamers and in controlling for performance on the network penetration task by providing measurements of cognitive abilities and personality attributes previously considered to be predictive of performance in this domain (e.g., Campbell et al. [2015b], Egelman and Peer [2015]). According to personal communication Jones and Trumbo [2019], the cognitive battery and personality assessment results from the original analysis Ferguson-Walter et al. [2019c] have been updated and new analysis was performed which has included seven additional participants from an additional session that occurred while the publication was in review. Updated results are reported here.

For all personality assessments, scores from the current work were compared against other data sets to achieve a greater understanding of how red teamers as a specialized population may differ from more general populations (e.g., college undergraduates). For all measures, means and standard deviations from our sample and comparison samples were calculated, as were mean difference scores, and an effect size (Cohen's d) was computed. Independent samples two-tailed t-tests were conducted

Task	Item	N	Score	Mean	Std. Dev.
Big Five Inventory	Extraversion	127	POMP	63.57	16.33
	Agreeableness	127	POMP	76.52	12.29
	Conscientiousness	127	POMP	77.1	13.78
	Neuroticism	127	POMP	47.87	13.66
	Openness	127	POMP	79.48	10.53
General Decision Making Style Inventory	Rational	127	Score	21.43	2.83
	Intuitive	127	Score	16.7	3.5
	Dependent	127	Score	15.94	3.94
	Avoidant	127	Score	10.5	4.79
	Spontaneous	127	Score	11.65	4.19
Indecisiveness		129	Score	26.34	6.75
Cognition		114	POMP	76.96	10.83
Overclaiming	Real	111	Endorsement (0-6)	5.5	1.6
	Fake	111	Endorsement (0-6)	2.7	2
Sandia Matrices	Day1	110	% Correct	42.8	0.2
		110	RT (all responses) (*s*)	21.4	6.25
		110	RT (correct only) (*s*)	19.76	6.06
	Day2	119	% Correct	46.9	0.22
		119	RT (all responses) (s)	17.86	7.14
		119	RT (correct only) (s)	17.62	6.21
Operation Span		118	Score (all in set correct)	41.71	18.39
		118	Score (any in set correct)	57.69	12.94
		118	Error (math)	8.04	7.45
		118	Error (speed)	1.81	3.3
		118	Error (Accuracy)	6.23	5.73
Shipley-2	Vocabulary	124	% Correct	32	5
	Abstraction	124	% Correct	15	3
Remote Associates Task		120	% Correct	23.9	11.4
		120	RT (*s*)	7.06	1.73
Problem Solving	Analytical	110	% Correct	19.4	29
	Insight	110	% Correct	45.5	32.2

Figure 2.7: Mean scores, standard deviation, and sample size for cognitive battery and personality questionnaires. RT refers to reaction time (in seconds); POMP refers to Percentage of Maximum Possible

to assess any statistical differences between groups. Assumptions of normality were not violated, however Welch's correction for unequal variances was applied since the sample sizes were often very different between groups.

For the General Decision-Making Style Inventory (GDMSI) and Need for Cognition (NfC), our participant characteristics were compared against a sample set of 1,919 U.S. adults recruited via Amazon Mechanical Turk to assess attitudes toward privacy and security in the cyber domain Egelman and Peer [2015]. Results suggest that for the rational subscale of the GDMSI, our sample displays a more rational decision making style relative to the comparison sample ($t_{(143)} = 4.20$, $p < .001$). For the avoidant subscale of the GDMSI, our sample shows a less avoidant style ($t_{(143)} = 4.97$, $p < .001$). For the spontaneous subscale, our sample shows a less spontaneous style ($t_{(143)} = 2.25$, $p < 0.05$). These results indicate that network penetration professionals approach decision-making scenarios with a relatively high emphasis on a thorough, more targeted, planned search for and evaluation of alternative approaches while avoiding postponement of decision execution. An analytical and decisive approach has been suggested in prior characterizations of this group Campbell et al. [2015b].

Results for the NfC scale suggest that our sample exhibits a higher need for cognition than the comparison sample ($t_{(141)} = 7.77$, $p =< .001$), indicating that red teamers have a greater tendency than the comparison sample to pursue difficult problems and to enjoy the process of thinking, which is consistent with what prior interviews have implied Summers et al. [2013].

For the Indecisiveness Scale (IS), our participant results were compared against those from 291 undergraduate students Rassin et al. [2007]. Results suggest that our sample is less indecisive than the comparison sample ($t_{(153)} = 6.93$, $p < .001$). These findings support the GDMSI result of a less avoidant style of decision making and are consistent with the notion that network penetration professionals tend to be decisive when presented with decision situations.

Our participant results on the Big Five Inventory (BFI) were compared against a dataset Srivastava et al. [2003] of 132,515 Internet users living in the United States and Canada, aged 21-60. The sample of network penetration professionals exhibited higher scores on Agreeableness (predilection toward trust and compliance; $t_{(126)} = 3.92$, $p =< .001$), Conscientiousness (level of efficiency and organization; $t_{(123)} = 4.98$, $p < .001$), and Neuroticism (an irritable, unhappy disposition; $t_{(126)} = 10.59$, $p =< .01$), relative to the comparison dataset. These results are consistent with results from our host-based cyber deception experiment Shade et al. [2020] which utilized a different population—computer scientists, system administrators and other computer specialists who passed a pre-screen test designed to select those with the skills needed to perform red teaming activities. The combination of these results may indicate that these traits of being more agreeable, more conscientious, and less neurotic, are not unique to the Red Teaming population, but may apply to the broader population of people with a predilection towards these types of computing profession. However, if the pre-screen adequately selected computer specialists with skills similar to Red Teamers (who could become good Red Teamers with proper training), it is another plausible explanation for sharing the same traits. Although In both cases, while the scores for the samples were slightly higher they remained within the average range for each scale.

Similar to the approach taken for the personality data, our cognitive task data was compared against additional data sets. Our scores on the Sandia Progressive Matrices were compared against a sample of 171 undergraduate students for the Day 1 session and 160 undergraduate students for the Day 2 session Clark [2014], matching groups who answered the same subset of problems. Our scores on the Working Memory Operation Span were compared against that of 6,236 college students (all under age 35, precluding matching of age bins from our sample; Redick et al. [2012]). A significant effect was found on Day 1, where our sample took more time on average

to solve problems compared to the normative set ($t_{(183)} = 2.06$, $p < .05$). No effects were found on Day 2.

For the Insight and Analytical Problem Solving task and the Remote Associates Test (RAT), the average solution rates (and reaction time for the RAT) were calculated based on the data of college students (see Wieth and Burns [2006] for the Problem Solving Task; see Bowden and Jung-Beeman [2003] for the RAT) for the subset of problems used in the current study, thereby allowing an extrapolation of number of correctly solved problems in the data to be compared against our data. Standard deviations were not reported for the solution rates, so the only statistical comparison that could be attempted was on the RAT reaction time to produce correct solutions, which yielded a small effect where the Tularosa sample was faster than the comparison sample ($t_{(162)} = 2.04$, $p < .05$).

No comparative dataset was available for the Karolinska Sleep Diary, so comparisons were not possible. However, in Figure 2.8, a trend improvement in sleep quality was observed from Day 1 to Day 2 (paired-$t_{(57)} = 2.74$, $p < .01$), suggesting subjects slept more efficiently after Day 1 compared to prior to beginning the experiment, which could be due to travel. See Figure 2.9 for a description of all significant effects from the comparative analyses.

2.2.2.3 Procedure

Our study took place over two consecutive days with up to ten participants per session, with sessions run between October and December 2018. Each participant was assigned to an individual work station in the same room but divided into private, cubicle-style spaces. A proctor was always present in the room to answer questions and ensure participants worked independently. We attempted to group similar conditions, e.g., informed, during sessions to minimize cross-contamination. Participants worked on the same network environment within a given day, e.g., morning and af-

	Day 1 (N = 103)		Day 2 (N = 93)	
Sleep Question	*Mean*	*Stdev*	*Mean*	*Stdev*
Bed time	11:14:00	1:58:00	11:08:00	1:44:00
Wake time	5:54:13	0:51:00	6:09:40	0:55:40
Time needed to fall asleep (minutes)	21.43	21.39	14.51	13.52
Number of times woke during night	2.28	2.49	1.62	1.92
Time awake during the night (minutes)	18.45	39.22	14.65	39.87
Time in bed (minutes)	402.56	94.11	437.94	81.6
Total time asleep, accounting for wake time (minutes)	365.38	104.05	408.08	88.8
Number of awakenings per hour	0.37	0.47	0.27	0.45
Sleep efficiency (ratio of time asleep to time in bed)	0.9	0.13	0.93	0.11
Sleep quality (1-5 rating)	3.31	1.24	3.82	0.95
Refreshed (1 - 5 rating	3.24	1.24	3.53	0.93
Soundness of sleep (1 - 5 rating)	3.34	1.3	3.67	1.11
Sleeping throughout night (1 - 5 rating)	3.47	1.46	3.85	1.27
Ease of waking (1 - 5 rating)	2.32	1.36	2.55	1.27
Ease of falling asleep (1 - 5 rating)	2.4	1.44	2.14	1.27
Dreams (1 - 5 rating)	2.04	1.37	2.06	1.32

Figure 2.8: Sleep Quality Question: summary statistics on responses across both days.

	Tularosa			Normative Sample			Mean Diff	Effect Size		
Task	**Mean**	**SD**	**N**	**Mean**	**SD**	**N**	**(Tularosa - Normed)**	**Cohen's d**	**t-Statistic**	**Significance**
BFI-44 Agreeableness[1]	70.76	15.39	127	66.40	17.79	132515	4.36	0.26	3.92	p < 0.001***
BFI-44 Conscientiousness[1]	71.46	17.24	127	63.84	18.02	132515	7.62	0.43	4.98	p < 0.001***
BFI-44 Neuroticism[1]	34.94	17.10	127	51.02	21.34	132515	-16.08	-0.83	10.59	p < 0.001***
GDMSI Rational	21.43	2.83	127	20.34	2.84	1919	1.09	0.38	4.20	p < 0.001***
GDMSI Avoidant	10.50	4.79	127	12.68	4.81	1919	-2.18	-0.45	4.97	p < 0.001***
GDMSI Spontaneous	11.65	4.19	127	12.51	3.76	1919	-0.87	-0.22	2.25	p < 0.05*
Indecisiveness Questionnaire	26.34	6.75	129	30.65	3.15	291	-4.31	-0.82	6.93	p < 0.001***
Matrices Day 1 RT †	19760.32	6059.75	110	18380.00	4441.85	170	1380.32	0.26	2.06	p < 0.05*
Need for Cognition[1]	76.97	10.64	120	68.95	15.30	1919	8.02	0.61	7.77	p < 0.001***
Remote Associates Test RT †	7057.01	1729.80	120	7566.09	1684.30	76	-509.08	-0.30	2.04	p < 0.05*

[1] = POMP Transformed

† = RT for Accurate Trials Only

Figure 2.9: Comparative Analyses: descriptive values and statistics for responses on cognitive tasks and questionnaires. Only significant effects are reported here. Significant differences are indicated as $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

ternoon on Day 1, but a new variation of the environment across days, e.g., Day 1 versus Day 2. They were given an attack laptop for the network penetration task and an additional laptop with internet connectivity for research and the cognitive battery. The items in the cognitive battery were either completed with printouts or using E-Prime 3.0 software.

2.2.2.4 Scenario

When introducing the scenario in the task briefing at the start of the day, the following was the exact wording given to explain the task:

> You will conduct recon on the network and locate vulnerable services, misconfigurations, and working exploits. Specifically, your task is to provide actionable intelligence about the company network which can be used by the follow-on team over the next 3-6 months. Your objective is to collect as much relevant information about the target network as you can in the allotted time without compromising future network operations. There may be deception on the network.

The underlined section was not underlined for participants; it indicates the portion that was only provided to participants in the informed conditions.

2.2.3 Discussion

In retrospect, the task briefing was too ambiguous. It was designed to be worded to *encourage* more realistic, stealthy, attacker behavior without being so specific that it changed each participant's natural tendencies. However, several participants interpreted the instructions to "provide actionable intelligence... which can be used by the follow-on team" as instructing them only to perform reconnaissance and not exploitation. Interesting, many of these participants did decide to attempt exploits in the end, in spite of this perception.

While there are cyber games and Capture the Flag (CTF) activities that occur every year, we believe this is the largest controlled experiment which held constant

the tools and exploits available to the attacker, the network topology and vulnerabilities, and the time participants had to launch attacks. In addition, the amount and variation of data collected is unprecedented (516 GB of data, plus an additional 1537 GB of screen capture video)[4].

2.2.3.1 Design Decisions

The results of our pilot studies, described in Section 2.1, indicated that cyber deception had a measurable impact on attacker performance, with more time spent on decoys than real machines and self-reported confusion of which were the decoy machines. They also investigated whether just the belief that cyber deception is in use can negatively affect attacks. The Tularosa design is built upon the results of those pilot studies. Many of the aspects of the experiment were kept the same, but key changes were made to ensure a more rigorous experimental design. Other aspects were changed due to necessity rather than a focused improvement to the methodology; we discuss some trade-offs below.

We kept the general design similar. The four conditions in the Tularosa Study were also present in the earlier pilot studies. However, the pilot studies used the same red team across all conditions. Additionally, the pilot studies placed the control condition on the second day (after the cyber deception unaware condition). This was completed on the second day to minimize a learning effect. We felt it was important to have a separate, true control condition that could account for any learning effects or a priming effect (where experience with conditions that include deception or self-doubt may build over time). To achieve that, the Tularosa Study moved from a within-subject design to a between-subject design.

Furthermore, we updated the design to take advantage of the participants' time while on site by adding a second day to each condition. The second day allowed us to

[4]The data is currently only available for government researchers.

better examine the persistence of cyber deception. For example, if an attacker attacks a network in which they are affected by deception for defense, do any of the cognitive effects, e.g., caution, frustration, suspiciousness, self-doubt, persist when they move to a new target network (that may have no deception in use)?

Notably we made some calculated trade-offs in the design decisions for the Tularosa Study that are less realistic than the pilot studies, the first being that the red teamers were asked to work independently. During the pilot studies, it was very difficult to determine the success of any individual when he or she was working as part of a team. This would mean looking for 30 teams per condition, rather than individuals (or reducing the sample size). We opted for a larger sample size, trading off a degree of realism for those used to working as part of a team or relying on another person for specific, relevant expertise. For some participants, this may have hampered their performance; however, it was a restriction that was distributed across all participants and could actually mimic a team that trades off roles across different tasks or days.

The next trade-off stems from having such a large sample size in this domain. While the Tularosa Study generally collected vastly more data than the previously discussed pilot studies, the pilot studies did have trained observers in the room, encouraging participants to talk aloud during the exercise. This enabled audio recording (and eventually text transcription with easy access to time stamps) to correlate verbal comments expressing frustration or confusion back to the computer inputs and outputs at that time. This gave the research team the ability to ask for clarification and also directly link the commands attempted and interactions with certain machines with psychological attributes of the participant. While this was a major strength of the pilot studies, it simply was not scalable. Additionally, with each participant working individually, we deliberately tried to minimize any verbal discussion or questions during the study that could influence another individual's thoughts or actions. We had a proctor in the room at all times, noting any blatant signs of frustration

and noting the time. Additionally we had a Red Team Report at the end of each day. To try to get similar information that can be more easily correlated with the time stamps of the participant interactions during the cyber task, we asked participants to use a chat client to document their plans and progress, in real-time, during the task. In reality, there was a wide variance in how people used this tool and how well people followed instructions in general.

The final trade-off we will review may have caused more severe effects on the generalization of the results. The Tularosa Study was run on a fully simulated network designed just for this study. There were no real users. There was no unexplained messiness or policies not being followed unless designed by us. This is clearly not realistic and may reduce the measured effectiveness, especially of the psychological deception, where pilot studies suggest the effects of being informed of deception benefit from the natural messiness present in an enterprise network.

2.2.3.2 Experimental Validity and Limitations

We considered validity concerns, including internal, external, and construct validity when designing the Tularosa Study. Internal validity was supported by using the same proctors throughout the experiment who read from a standardized script for instructions and responses to participant questions. Participant time on task was cataloged and monitored throughout the session, as were breaks, and lunch. Duration of the cyber task was controlled, as was the lunch break, for all participants. Participants were not allowed to discuss the cyber task during breaks. Identical copies of the cyber range were presented to all participants for a given condition, and identical machines (Kali Linux and internet-ready reference laptops). We arranged ahead of time to include any publicly available tools requested by participants, however no proprietary or costly tools were allowed. Additionally, a large standard set of red teaming tools were provided. A within subjects' component was implemented,

whereby only cyber range deception was manipulated on Day 2. This design choice reduced the amount of individual variability across days and conditions inherent in between subjects' designs.

Many aspects of the Tularosa design support external validity. Since this was a tightly controlled laboratory study, the ecological validity could be called into question. For example, the standard set of tools provided could have hampered the performance of participants who were out of their comfort zones and unable to rely on tools they regularly use. As a proxy for one aspect of ecological validity, we asked participants to rate on a scale of 1 to 5 how they felt regarding the tools provided to them during the experiment (Appendix A.2). The mean rating was 3.51 out of 5, with a standard deviation of 0.93, suggesting that participants were largely satisfied with the tool selection provided to them. Participants were provided with a popular red teaming platform, Kali Linux, as well as internet access on a separate laptop for research. This experiment was designed to test the behavior of red teamers, and how this study would generalize to other populations who perform cyber attacks is unknown at this time. We subcontracted participants through various companies in several states around the United States, thus giving this project a broad, random sample within the specific population of professional red teamers. That said, this experiment was not an "in the wild" red teaming exercise, and thus proprietary tools were not allowed, participants had to work alone rather than in groups, and had a tightly controlled schedule. Finally, real-world cyber attack scenarios and ted teaming engagements typically exceed one day. Moreover, often the attacker will be the deciding factor of how long the engagement continues, which could change dynamically based on many relevant factors including interest, difficulty, and priority. We only allowed the participants to perform the task for one day per network. This was a monetary necessity but does diverge from the usual experience as evident in the data collected from the participants on the usual duration of engagements.

Construct validity is difficult to measure currently, as many planned future analyses will be required to determine if the deception led to altered cyber-behavioral performance. However, results discussed in Section 3.4 on self-reported suspicion of deception by condition did reveal associations between the cyber deception manipulation and suspicion. The data suggest an aggregate effect of the two deception manipulations, as the Cyber Aware condition showed the largest suspicion scores, whereas the Cyber Unaware condition produced an effect of roughly 80 percent that of the Cyber Aware condition. These data need to be scrutinized more carefully to disentangle the specific contributions of each of the deception manipulations.

2.3 Conclusions

Cyber deception is an inherently interdisciplinary domain. It sits at the intersection of computer science and the social sciences. Human behavior is at the root of cyber offense and defense. Understanding human behavior and leveraging this understanding for the defender's advantage are the foundations of defensive cyber deception. Deception techniques affect the operator behind the keyboard who is attempting to complete a mission and should have a stronger and longer-lasting impact than simply detecting or impeding attacker actions on the defended system. The pilot studies' results paint a picture of just how powerful deception can be for cyber defense. The reason is simple: attackers are usually human operators. Deception is one technique that focuses on affecting the operators themselves.

Cyber deception has been described as a "game changer" in cyber security Gartner Report [2015]—one that can allow the cyber defender to leverage the *home-field advantage* of owning and controlling the targeted network environment. We designed the Tularosa Study to empirically measure the effectiveness of cyber, and psychological, deception on an attacker's ability to perform reconnaissance and exploitation. While this chapter describes the experimental design, methodology, cyber range and

participant population, in Chapter 3 we discuss the data analysis completed to provide results addressing the main hypotheses. The scale and breadth of data collected in this controlled study will allow for further future analyses beyond those described in thesis. Furthermore, there are many cyber defense research questions beyond cyber deception we believe this data can help address.

CHAPTER 3

DATA ANALYSIS

In this chapter we review the initial data analysis performed on a subset of data collected from the Tulsarosa Study Ferguson-Walter et al. [2019c]. The analysis will address these hypotheses:

- Hypothesis $H1$: Defensive cyber, and psychological, deception tools impede attackers who seek to penetrate computer systems and infiltrate information. To address this hypothesis we compare performance on the cyber task between control and experimental conditions. We will compare results across all experimental conditions to assess which type of deception is most effective.

- Hypothesis $H2$: Defensive deception tools are effective even if an attacker is aware of their use. To address this hypothesis we compare performance on the cyber task between conditions where participants were informed about deception to where they were not informed.

- Hypothesis $H3$: Defensive deception is effective even if the attacker merely believes it may be in use, even when it is not. To address this hypothesis we will compare performance on the cyber task between control condition and psychological deception condition.

- Hypothesis $H4$: Defensive cyber, and psychological, deception causes increased confusion, and surpsie in the attacker. To address this hypothesis we compare level of cognitive effects reported between control group and experimental conditions.

The Tularosa dataset can address hypotheses 1–4 because it includes experimental conditions that differ in whether cyber deception defenses are or are not present (Absent versus Present) and whether or not participants were told deception may be present (Informed versus Uninformed). Please note that hypothesis $H5$ will be addressed in Chapter 4.

Before any of the questions of interest outlined in the hypotheses can be answered, we must first know what constitutes success. In the cyber domain, there are multiple, and sometimes competing, indicators of success or failure. Were subjects stealthy in their approach? Were they able to exfiltrate the important information? Did they appropriately identify and avoid decoy systems? Were they fast? These questions can all be addressed to some degree by the data collected, but a notion of the level of success will still be largely subjective.

While a subset of data was missing for various sources, excluding some participants from specific analysis, the total number of participants who were included for this analysis are as follows: 35 for AUPU, 28 for AIAU, 30 for PUAU, and 30 for PIAU, for a total of 123 professional red teamers. Since this analysis focuses on the day 1 cyber data, we include participants who were allotted the standard time to complete the cyber task.

A total of 139 participants were run through the experimental protocol in 17 different sessions. All five subjects from the first session day 10/23/2017 were excluded due to data collection issues effecting both network and keystroke capture. All ten subjects from session day 11/13/2017 (one of which chose to opt-out of the HSR portion) were excluded due to a late start on the Day 1 cyber task which caused a reduction in the allotted time. One subject who participated did not fit the selection criteria and was excluded.

The current analysis mostly focuses on between-group comparison across the four conditions on Day 1 to begin to address the main hypotheses. Analysis involving data

from Day 2 remains future work. There were two participants who were no-shows on Day 2 after participating in Day 1, and an additional seven who participated on both days, but whose Day 2 cyber data is missing due to technical difficulties. For these seven participants we are still able to use their self-report and cognitive data for future analysis.

3.1 Measures of Success

There is no widely known and acceptable metric for red teamer or hacking behavior. Often success if subjective and dependent on the specifics of the motivation and the person behind the attack. Measuring success of each red teamer in the Tularosa Study is also subjective. There are many different ways to measure success and many additional paths that can be taken for future work. In this section we discuss metrics for forward progress, effort wasted on decoys, and self-reported success. The following sections address additional critical metrics including: detection evasion, altered perception, and altered cognitive and emotional state.

3.1.1 Forward Progress

Forward progress is difficult to measure for this analysis because neither a specific end-goal nor explicit flags were provided for participants. Allowing participants to determine what is reportable also reveals what they perceive to be of significance. In this respect, the Tularosa Study differs from a typical Capture The Flag (CTF) exercise, where the "flags" take the form of computer files containing a specified keyword hidden throughout the system and are thus pre-determined by the designers of the exercise. In contrast, real-world network exploitation requires subjective valuation of objectives and risk of exposure, which is what our study targeted. The motivation behind this design decision is discussed in more detail in Section 5.1.4.

One proxy for forward progress which has shown to have significant differences across conditions is the number of keystrokes. There are many limitations in this measurement, as it does not take in to account think-time, or that some participants might be more productive and efficient while also typing less. However, we think this is a reasonable measurement to support the hypothesis that we will see a difference in forward progress across conditions since attackers cannot progress very far without interacting with the attack client. Since the assumptions to run a parametric test were not met because these data are non-normal, we ran the Kruskal-Wallis statistical test which provides a non-parametric one-way analysis of variance to check for a significant difference between conditions. While there is no statistical difference across the main four conditions, we do see a significant difference when combining the Absent and Present conditions (Kruskal-Wallis chi-squared = 2.7079, $p = .015$). This indicates that the keystroke count of participants in the Absent conditions are higher than those in the Present conditions which is consistent with the hypothesis $H1$ that decoys impede and delay forward progress.

We also considered key terrain, namely the domain controller (DC). While there was no statistically significant differences across the conditions in self-reported identification or exploitation of the DC, we do see a notable numerical difference in identification. Participants in Present conditions successfully identified the DC less than half the time and those in the Absent condition identified the DC more than half the time. Only three or four participants per condition reported successfully exploiting the DC. Interestingly, 100% of the participants in the Absent conditions who self-reported successfully exploiting the DC also reported exfiltration of critical information from the DC, in comparison to 50% of those in the Present conditions. It is unknown at this time if this is a difference in forward progress or a simply a difference in reporting. These non-significant findings are consistent with the hypothesis $H1$ that presence

of decoys impedes forward progress, future work will examine the ground truth of success as seen in the cyber data.

Our experimental design also allows us to consider different timing measurements. For investigating forward progress, we will consider the amount of time spent before attacking begins. While this could be caused by many different things, including thinking about or researching something that will actually end up making the attacks more effective, this timing analysis will just be a small portion of a bigger picture as we consider the data analysis results. First we consider the time until the first alert (of any type) is triggered by an interaction with a decoy. These *decoy alerts* are generated by the decoys and thus only exist in the Present conditions. Decoy alerts are the preferred alerting metric because there are no false-positives, by design, since no legitimate users or services would be interacting with a decoy.

Since the assumptions to run a parametric test were not met because these data are non-normal, we ran the Kruskal-Wallis statistical test to check for a significant difference between conditions with results as follows: Kruskal-Wallis chi-squared = 4.4416, $p = .035$. This indicates that the Present-Uninformed condition (mean = 20.59 minutes) took a significantly longer time to initiate an interaction with a decoy than the Present-Informed condition (mean = 11.74 minutes). While it might be assumed that information about deception can delay an attacker by making them think twice about what to do first, this result indicates otherwise. The data supports the opposite theory, that knowledge about the deception can make an attacker more aggressive and less cautious. Perhaps this behavior indicates that they are eagerly hunting for the deception. This is likely because the information of the deception was vague. In the Pilot Study participants were specifically told to expect decoy systems; this information slowed down their initial actions. In the Tularosa Study the information was simply that *deception may be present* which appears to have caused participants to want to quickly seek out the details/source of said deception.

This change in behavior could also benefit defenders, as less cautious behavior on the attacker's part may lead to faster detection and mitigation by the defender.

3.1.2 Attacker Resources Expended

One important measure of the effectiveness of decoys is the level of attacker effort directed against them. In the Tularosa Study, one key resource was the limited time available to the attackers. The study allows us to measure time spent attacking each host in multiple ways, as target IP addresses were captured in multiple data streams: captured network traffic; self-reports of attacker activity (Mattermost); shell commands entered; and unstructured notes taken, which will primary be future work. While participants varied in the verbosity of their notes and reporting, it is primarily the difference in interaction with real (versus decoy hosts) across conditions that matters. In addition, the deceptive hosts logged all network interaction. Any attacker resources expended on decoys also delays forward progress.

To investigate the question of increased effort expended in deception conditions, we consider several different measures. First, we examine the number of commands sent to real hosts versus decoys. We filter out shell commands that contained an IP address and then assess whether that targeted IP address is a real or decoy machine. Then we count the number of instances a real machine was targeted by each participant in each condition. There were no statistically significant differences in the number of decoys targeted across the Present conditions. However, results indicate a statistically significant difference in the number of real machines targeted indicating that fewer real machines were targeted in Present conditions than Absent conditions (Kruskal-Wallis chi-squared = 4.58, $p < .01$). This supports hypothesis $H1$ that the presence of decoys impedes forward progress and protect real machines from attack. This also helps build a case for the technical effectiveness of decoys for defense. A total of 710 commands included decoys IP addresses, so this is a minimum baseline for the

number of commands wasted on decoys. See Figure 3.1 for the average IP-containing commands per participant.

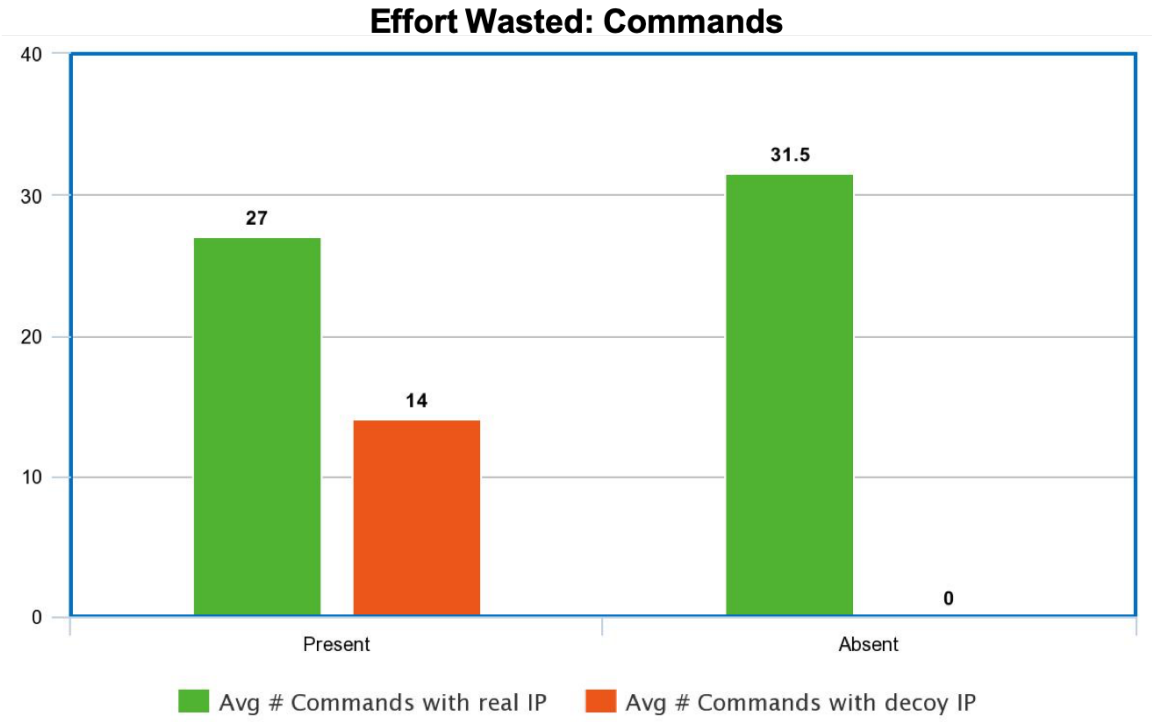


Figure 3.1: **Impeded Forward Progress: Average number of commands wasted on decoys.** Wasted resources demonstrates technical effectiveness of decoys. Significantly fewer commands with real IPs in Present condition is consistent with decoys impeding attacker forward progress.

Next, we consider the total byte count of all the packets sent from each participants' host. While the results (Figure 3.2) do not indicate a significant difference in medians among the four conditions, we note an increased variance in Present-Uninformed is noted across many data types and may be a feature caused by the unknown presence of decoys. This was also observed in the pilot studies where the deception caused some cyber attackers to become more cautious and work slower, it has the opposite effect on other participants, who become less cautious perhaps due to frustration. A change in behavior is evident, both observationally and in the cyber data. The increase in variance can be interpreted as indicating chaos injected into the

performance of participants when decoys are present. This supports the hypothesis $H2$ that the presence of decoys changes an attacker's behavior, whether or not they are aware of the deception.

Results indicate that significantly fewer bytes were sent to real machines in the Present conditions (Kruskal-Wallis chi-squared = 5.28, $p = .022$). These bytes totaled over 10GB and can be considered wasted attacker resources. This further displays the technical effectiveness of decoys for defense. See Figure 3.3 for the average megabytes per participant.

Next we discuss the number of packets sent to decoys. While number of packets and bytes are related, they are different measures, since packets can vary widely in size. Any packets or bytes sent to decoys is a waste of the attacker time and resources. It also increases the risk of them exposing themselves to defenders. Figure 3.4 indicates that in the conditions where decoys were present, 35% of the packets sent were to decoys. We see no statistical difference in number of packets sent when comparing the Informed and Uninformed conditions. While ideally we would hope to see 50% or more of the traffic targeting decoys (since 50% of the assets on the network were decoys), in this experiment, the simulated network and the decoys were not configured for maximum realism or interaction, so we would expect to see even more packets targeting decoys in real-life scenarios.

Wasted effort can also be seen through the number of snort alerts detected referencing a decoy IP (more details in Section 3.2). Any attacker activity that generated a snort alert on a decoy is wasted effort. Notice that for the Present conditions, more alerts are on decoys than real machines. Each Absent condition accounts for about a quarter of the snort alerts for the four conditions (as expected). However, each Present condition has about half the number of snort alerts on real machines than the Absent conditions. This is a very explicit example of effort wasted on decoys since if the decoys were not present, all of that effort is on real machines instead.

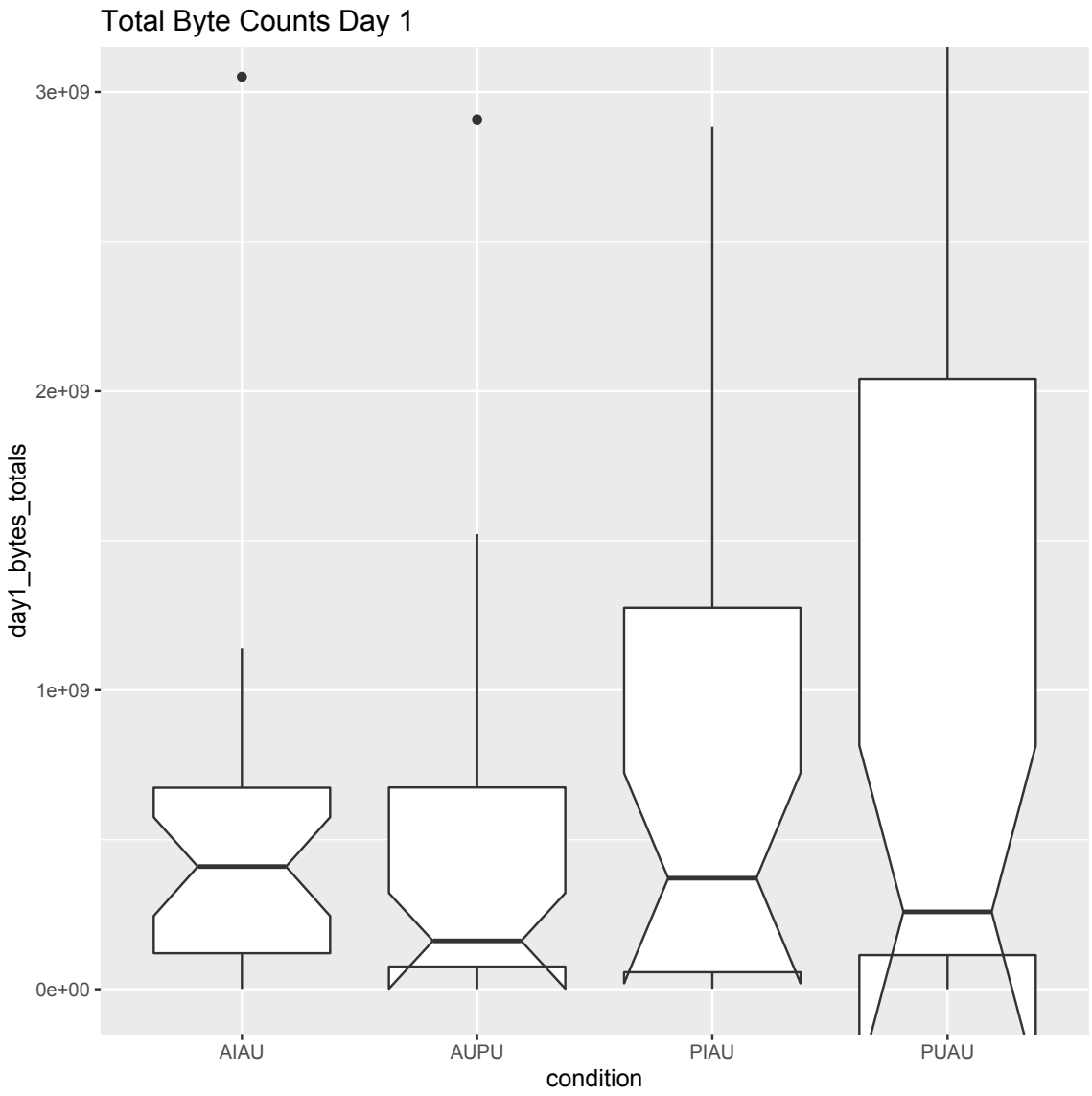


Figure 3.2: **Impeded Forward Progress: Between-group differences in total byte count across all conditions.** Results do not support a statistical difference in byte counts across conditions, but note the increase variance in the Present-Uninformed condition.

3.1.3 Self-reported Success

To further consider success on the cyber task, we evaluated the *Day 1 Red Team Briefing* requested from all participants upon completion of the cyber task on Day 1. There were several participants who either misunderstood that the briefing was required, or decided not to complete it (as indicated by the reduced total N).

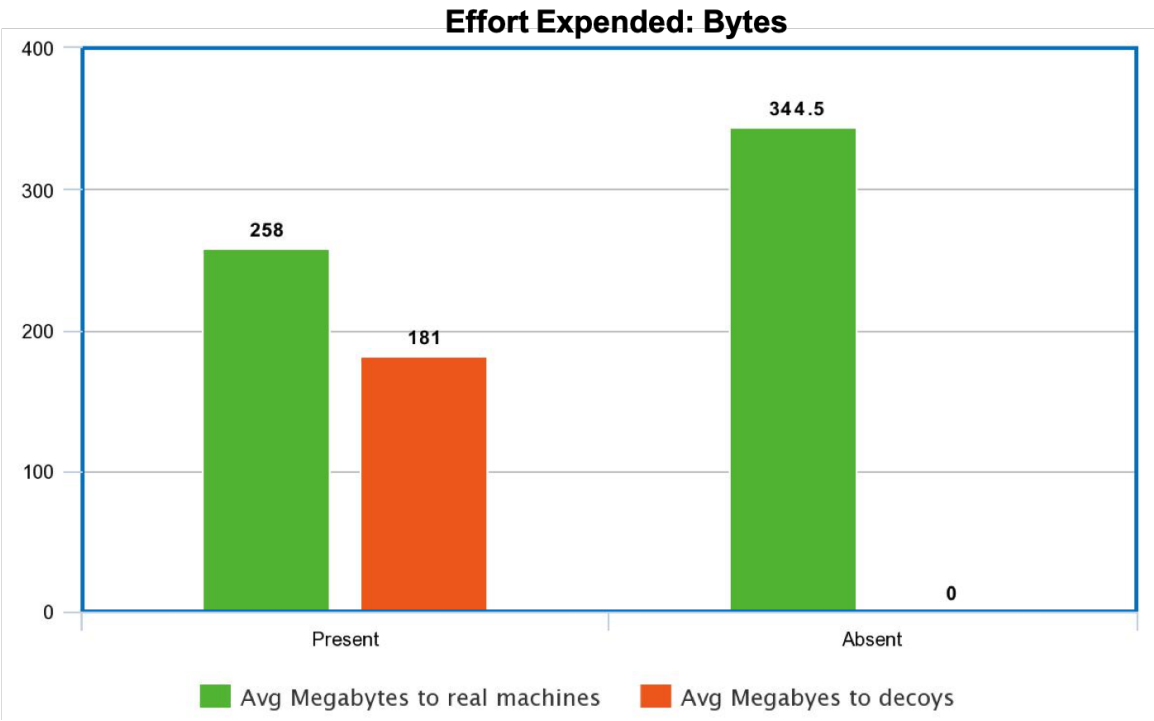


Figure 3.3: **Impeded Forward Progress: Average number of Megabytes wasted on decoys.** Wasted resources demonstrates technical effectiveness of decoys. Significantly fewer megabytes sent to real IPs in Present condition is consistent with decoys impeding attacker forward progress.

In the Tularosa Study, we also directed participants to self-report vulnerabilities they identified. Using a time-stamped Mattermost chat client, they followed a semi-structured reporting format (to include at least the IP address of the target host) and report all "potentially useful information about target systems on this network". This provides timestamped information on the thought process and beliefs of the participants captured during the cyber penetration task (rather than only in retrospect). These reports allow us to compare the number of reported vulnerabilities to the number successful exploited. This also allows us to monitor when the participant changes their mind about the value of a target or changes their stated strategy. It will also allow for future work examining what information the participants deemed

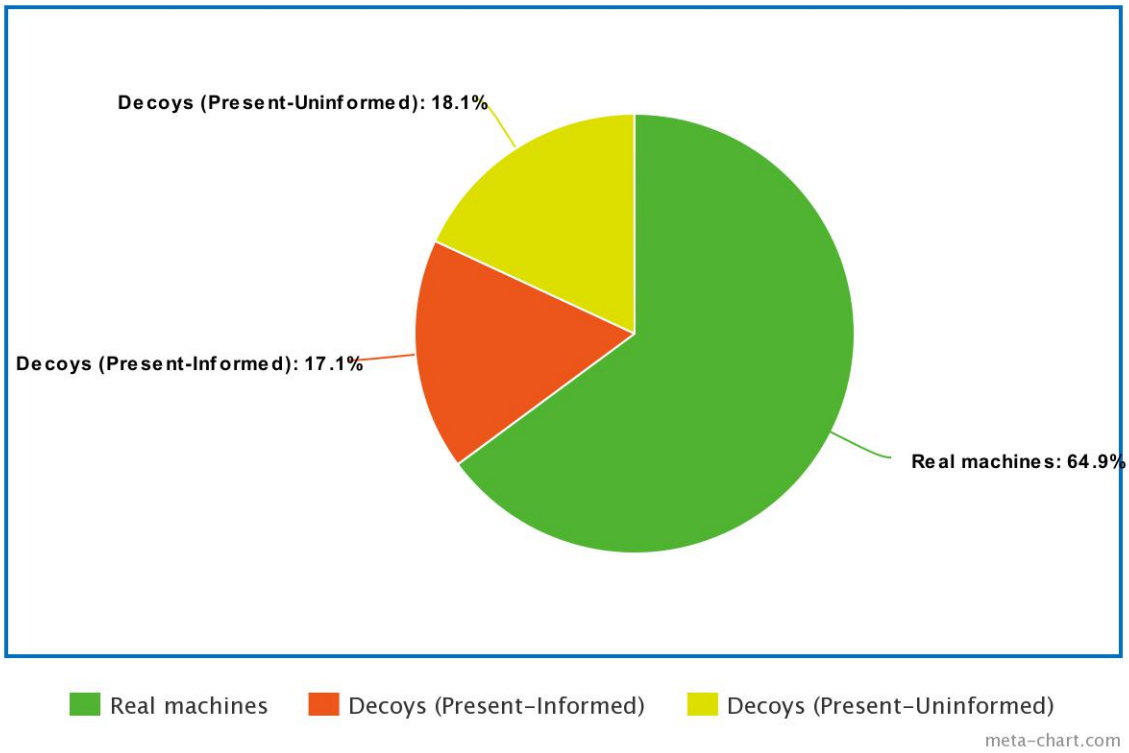


Figure 3.4: **Wasted Resources: Number of Packets sent to decoys.** No statistical difference between Present-Informed and Present-Uninformed. Note that packets sent to decoys can indicate wasted effort.

significant enough to transcribe into the final end-of-day report and how that related to their real time reporting and their experimental condition.

We examined the Mattermost messages and labeled each message that referred to an exploit as either a reported failed exploit attempt, a reported successful exploit attempt, or an identification of a vulnerability that could be exploited. We furthermore provided that quantity of the number of new machine IPs identified in the message, as well as the specific exploit type that was discussed. A statistical analysis using a Kruskal-Wallis test of these results shows a trend in the self-reported exploit successes such that the Present conditions reported fewer exploit success than the Absent conditions ($p = .076$), further supporting Hypothesis $H1$. Using the Dunn Test for multiple comparison after running Kruskal-Wallis test for stochastic dominance

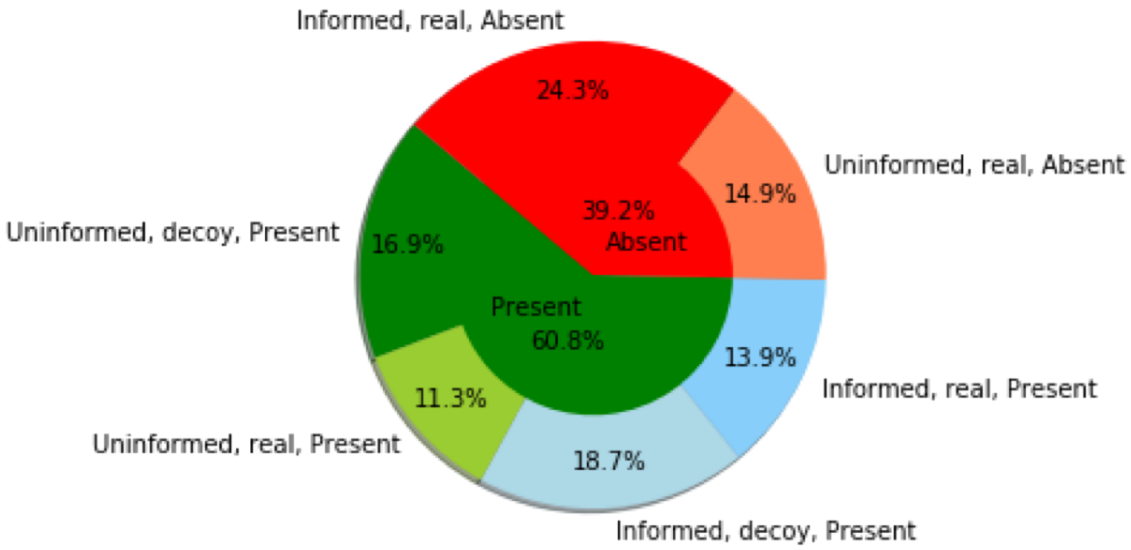


Figure 3.5: **Wasted Resources:Number of Snort alerts triggered.** Number of alerts generated per condition split by real and decoy IP addresses. Note that activity triggering snort alerts on decoys can indicate wasted effort.

(Kruskal-Wallis chi-squared = 8.182, $p = .042$), we also find a trend of the Present-Uninformed condition reporting more identified vulnerabilities that could be exploited than the Present-Informed condition ($p = .061$) which is consistent with the information about deception causing doubts about potential vulnerabilities discovered. This supports the hypothesis $H2$ that being informed of the deception negatively effects performance which could be related to increased second guessing and self-doubt ($H4$)

This labeling scheme does not incorporate the fact that some of the exploit attempts might have been against decoys and were thus, not truly successful. Additional analysis can be completed as part of future work to make this determination, which (since it can only decrease the true successes in the Present conditions) can only increase the significance between the difference in means between Present and Absent conditions. While self-report data is problematic due to subjectivity and inconsistencies, this can been seen as an additional indicator that the decoys impeded forward

progress, as hypothesized in $H1$. Future work could be done to investigate the belief structure of the attackers.

3.2 Evading Detection

Network penetration tests are not always focused on quiet versus loud attacker behavior due to the attacker emulation being expected and authorized by the network owners allowing access. However, these cyber professionals are often skilled at remaining undetected since, detectability adds extra realism or purpose to the task, and many techniques, tactics and procedures (TTPs) are learned with an understanding of the various levels of noise they create. Advanced cyber defense techniques are often focused on advanced persistent threats (APTs), which are quiet by definition—noisy attacks are easily detected, so often not persistent. To our knowledge, little research has been done to understand loud versus stealthy attacker behavior, or the characteristics of the attacker who performs them.

The scenario prompted the participants to attempt to avoid detection by stating: "your objective is to collect as much relevant information about the target network as you can in the allotted time without compromising future network operations". This should be their natural tendency since noisy actions such as rapidly executing numerous probes on each host is likely to result in immediate detection and expulsion from a network. In the Tularosa Study there is no response to detected activity for several reasons: first, the limited time available to participants recon and exploit the network; second, blocking participants from the network would preclude them from further participation (reducing the amount of data collected); and third, involving people in defending the network would add variability between participants which would be difficult to resolve during data analysis. We, instead, retroactively determine when each participant would have been detected by a typical intrusion detection system (IDS).

Built-in to the deceptive network defenses is the ability to detect and record attacks launched against the decoys. This allows us to determine what fraction of the attacks would have been detected in real time by the decoy system and determine how many exploits were successful prior to that time. Another question of interest is whether participants in the informed conditions were less aggressive, i.e., launched fewer scans and attacks, for fear of detection.

This section is focused on the level of detectability, measured as the number of alerts that were triggered by the participant's activity on Day 1 in the simulated network environment. We analyzed the raw number of snort alerts generated and also the ratio of snort alerts divided by the number of packets sent from the participants computer. This ratio was utilized to ensure that the results were not biased by the various levels of forward progress made by each participant. In other words, what percentage of attacker activity was detected?

Alerts were tracked by replaying the packet capture (PCAP) through community rules in Snort and Suricata. We could also detect whether the alert was triggered through interaction with a decoy or real asset in the simulated network environment. We took the aggregate of alerts, indicating detected activity targeting either a decoy or real asset in the simulated environment and divided that by the number of packets sent from the participants computer to achieve a detectability ratio.

Over half of the snort alerts in the Present conditions involved decoy targets; see Figure 3.5. For our analysis we did the typical outlier removal, removing outliers that were three standard deviations away from the mean (subjects removed were S51, S109, S112). This is count data with a one-tail distribution, and requires a non-parametric approach. Since the distribution is quantitative and non-normal, we ran the Kruskal-Wallis statistical test to check for a significant difference between conditions (Kruskal-Wallis chi-squared = 8.7481, $p = .033$). This was followed by pairwise comparisons using Dunn test which indicated a significant difference between

	Absent-Uninformed	Absent-Informed	Present-Uninformed	Present-Informed
mean	550	734	973	1138
N	32	26	28	27

Table 3.1: **Detectability: Means of snort alert counts across conditions.**

the Absent-Uniformed (Control) condition and the Present-Informed condition ($p = 0.007$) supporting our hypothesis $H4$. Table 3.1 displays the calculated means of each condition and the number of subjects (after outlier removal) available for this data analysis. There was a significant difference indicated that the participants in the Absent-Uniformed (Control) condition tended to trigger less snort alerts than those in the Present-Informed condition ($p = .026$). There was also a significant difference between the Present conditions and the Absent conditions when combined (Kruskal-Wallis chi-squared = 6.71, $p < .01$), which is consistent with the hypothesis that the presence of decoys increase the chance that an attacker reveals themselves to a defender, which can impeded forward progress, supporting hypothesis $H1$.

While there was a significant difference between the Present-Informed condition and the Control condition for raw number of snort alerts on Day 1 (Figure 3.6), this finding was not evident when using our detectibility ratio likely due the increased variance noted in Present conditions and discussed above. Since the number of bytes sent is correlated to the number of packets sent, it is not surprising to see similarly increased variance.

While there was no significant difference in the number of Snort alerts generated with a decoy IP as the target between the Present-Informed and Present-Uninformed conditions, we also analyzed the alerts generated by the decoys themselves. We examined each individual type of alert the decoys generated, as well as the total of all alerts per participant. First we removed all outliers that were above three standard deviations from the mean (S123 for scans, S55 for probes, S88 and S93 for touch).

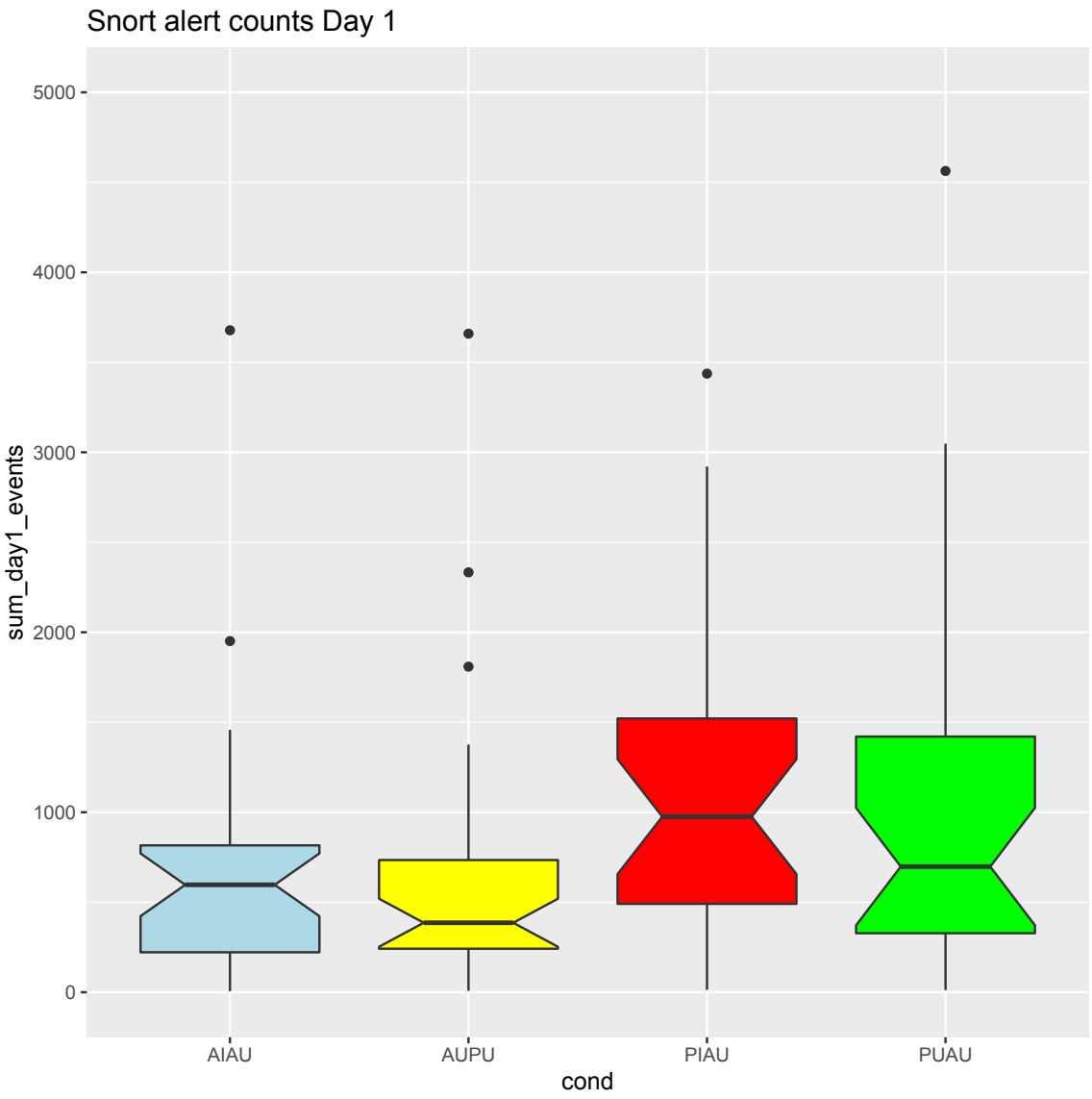


Figure 3.6: **Detectability: Between-group differences in number of snort alerts triggered.** Results indicate a difference in medians between Absent-Uninformed (control) and Present-Informed conditions on Day 1 that is statistically significant.

Since these alerts are generated by the decoys, we can only compare the Present-Informed and Present-Uninformed conditions. We used the Kruskal-Wallis statistical test and found significant differences across all alert types. This gives further evidence that decoy system alerts have utility above what a standard IDS can supply.

Recall that scan alerts are triggered on a decoy when an subject scans multiple decoy IPs within a short time period, while probe alerts are generated when a single decoy IP is probed for additional information. An intrusion alert for the decoy system is triggered only in response to an interactive logon response; many exploit attempts will trigger a probe alert instead. Both probe and intrusion alerts are triggered later in the kill chain than touch and scan alerts. The Present-Informed condition had significantly more touch alerts ($p = .006$) and scan alerts ($p = .005$), but fewer probe alerts ($p < .0001$) than the Present-Uninformed condition (see Figure 3.7). The Present-Informed condition having more scan alerts but fewer probe alerts is consistent with hypothesis $H2$ and the information of deception actually further reducing their forward progress.

Combining the alerts, we find the Present-Informed condition had significantly more total decoy alerts overall than the Present-Uninformed condition ($p < .0001$). This indicates that Information on the presence of deception reduced more aggressive behavior towards the decoys. We suspect that this is because the informed participants are less likely to continue to interact with any asset that seems suspicious, due to fear it may be deceptive. Intrusion alerts, which are generated by decoys after an interactive login attempt (e.g. SSH, RDP) were also examined. While the findings for total number of intrusion alerts (with outlier S88 removed), which are triggered by an interactive login attempt, were not statistically significant (since only 33 out of 60 participants in Present conditions generated intrusion alerts on Day 1), the Present-Uninformed condition had twice as many participants (N=22) who had at least 1 interactive login attempt on a decoy than the Present-Informed condition (N=11). A chi-square test of independence was performed to examine the relation between condition and the triggering of at least one intrusion alert. The relation between these variables was significant ($\chi^2 = 8.15, p = .0043$). Participants in the Present-Informed condition were less likely than those in the Present-Uninformed condition to attempt

an interactive logon to a decoy. While in isolation this result is consistent with the information of deception leading participant to identify and avoid it, other results, such as the misidentificaiton of machines in Figure 3.10 counter that. Instead, we purport that this is another indication of impeded forward progress by participants in the Present-Informed condition, thus supporting hypotheses $H1$ and $H2$.

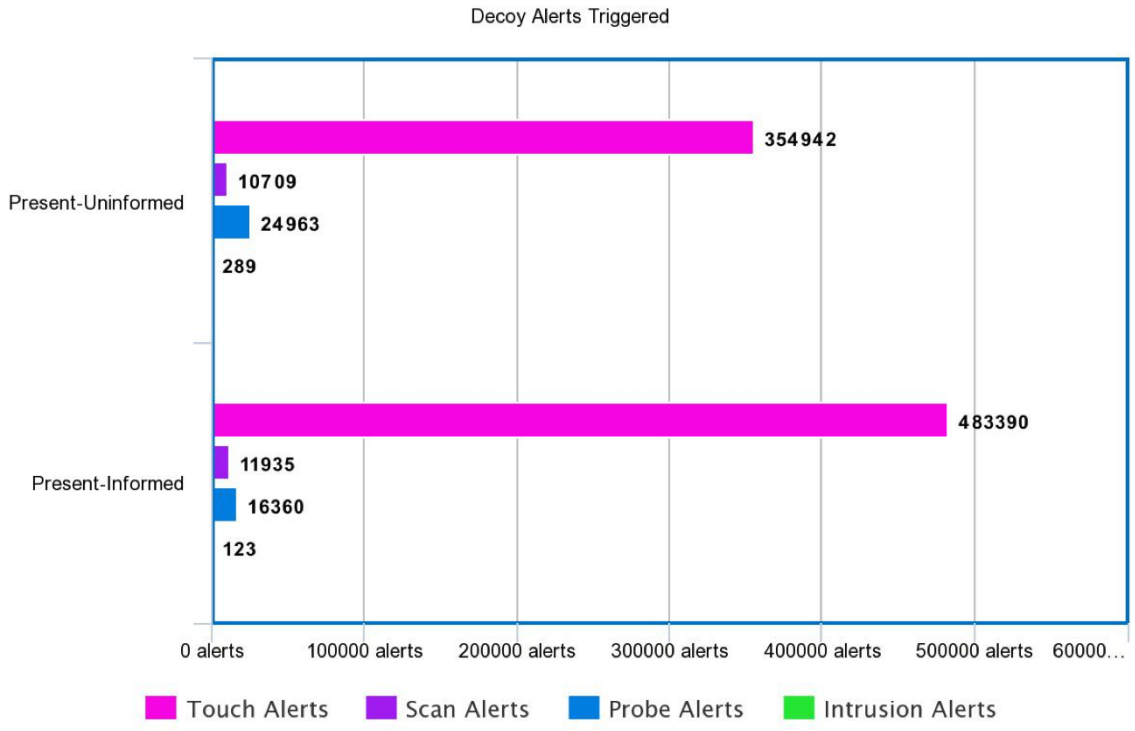


Figure 3.7: **Impeded Forward Progress: Decoy alert triggered by participants in the Present conditions.** The Present-Uninformed condition had significantly more fewer touch and scan alerts but significantly more probe and intrusion alerts indicating that they progressed further in the cyber kill chain that the Present-Informed condition.

There is a notable difference in the graphs of the severe and major Suricata alerts over the course of the cyber task (Figure 3.8). While this doesn't directly support any of our hypotheses, it does demonstrate a difference in the pattern of cyber behavior across the four experimental conditions. Note there is some consistency seen matching the previous finding in decoy alerts that the Informed conditions are more

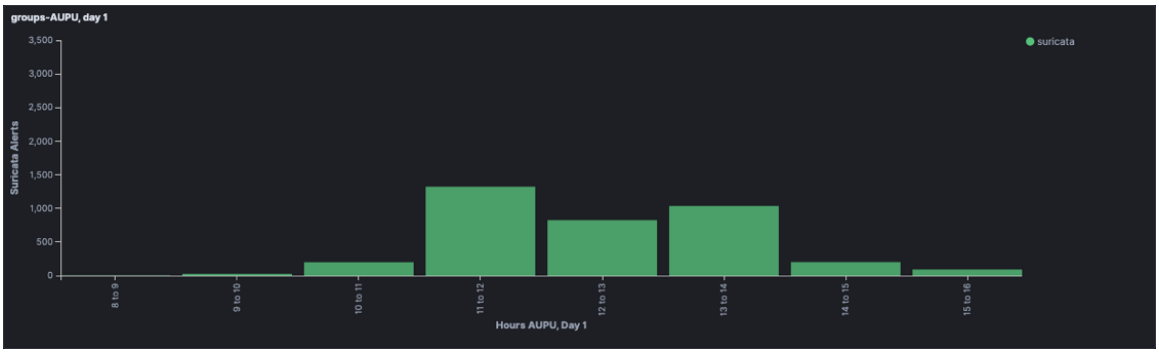
aggressive and trigger alerts faster. We theorize that this is a search to discover the nature/location of the deception.

3.3 Altered Perception

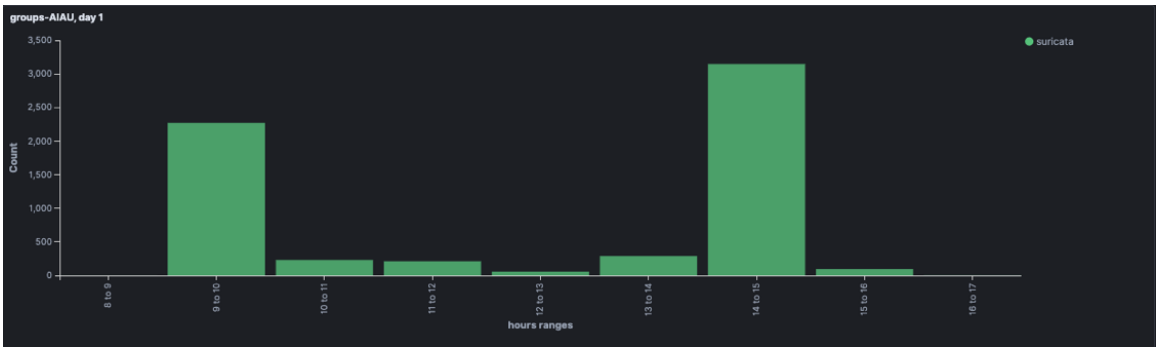
The main purpose of the Tularosa experimental design is to measure the effects of cyber deception–the most obvious being a observable difference between reality and perception caused by the deception. While comparing perceptions to ground-truth remains future work, we can begin to see difference in the perception of success across the experimental conditions.

We coded the narratives written in the end-of-day briefing as *Success* if the participant discussed more self-perceived successes than failures, as *Failure* if the participant discussed more self-perceived failures than successes, and as *Neutral* if the number of failures and successes discussed were equal or (more commonly) the briefing did not discuss failures/successes. A chi-square test of independence was performed and indicated as statistically significant difference in the number of reported failures in the Absent-Uninformed condition and the Present-Informed condition ($\chi^2 = 4.49$, $p = .034$) and displayed in Figure 3.9. Notice that the Present-Informed condition has half as many failures reported than the other conditions. This could be because the combination of being informed of the deception and having deception present acts as an excuse for the participants who no longer feel responsible for the failures and therefore report failures less and successes more. This idea would need to be further evaluated by looking at other metrics of self-reported failure compared to actual failures.

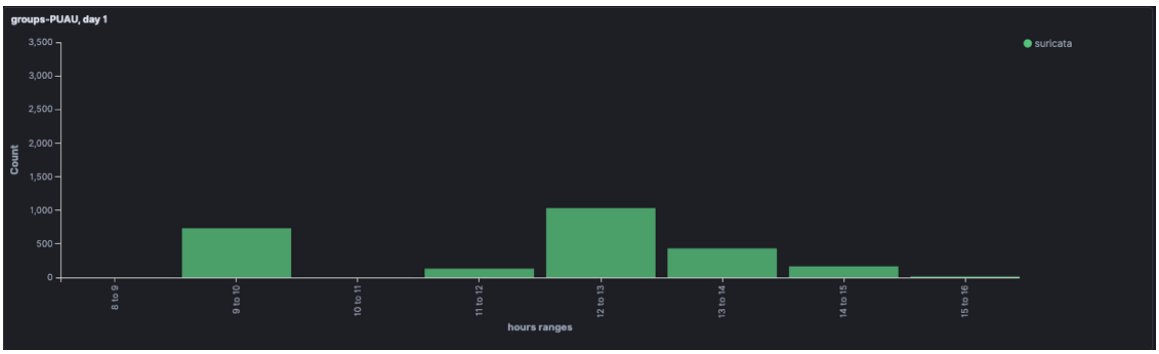
Any perceived deception by the Absent-Informed condition is clearly an example of a mismatch between perception and reality. While the data we captured did not make it easy to observe the measurable effect that information may have had on the Absent-Informed condition, we did see instances of blame being placed on



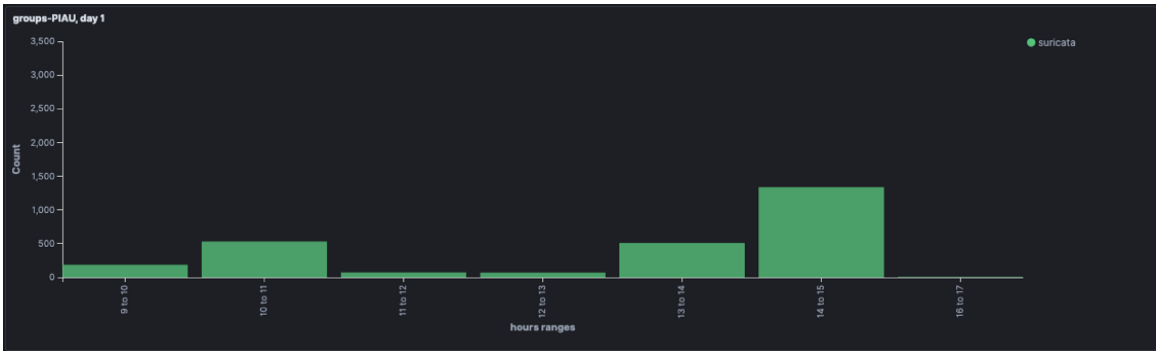
(a) Absent-Uninformed (Control): Day 1 Suricata Alerts per hour.



(b) Absent-Informed: Day 1 Suricata Alerts per hour.



(c) Present-Uninformed: Day 1 Suricata Alerts per hour.



(d) Present-Informed: Day 1 Suricata Alerts per hour.

Figure 3.8: **Change in Cyber Attack Behavior: Count of Suricata alerts.**

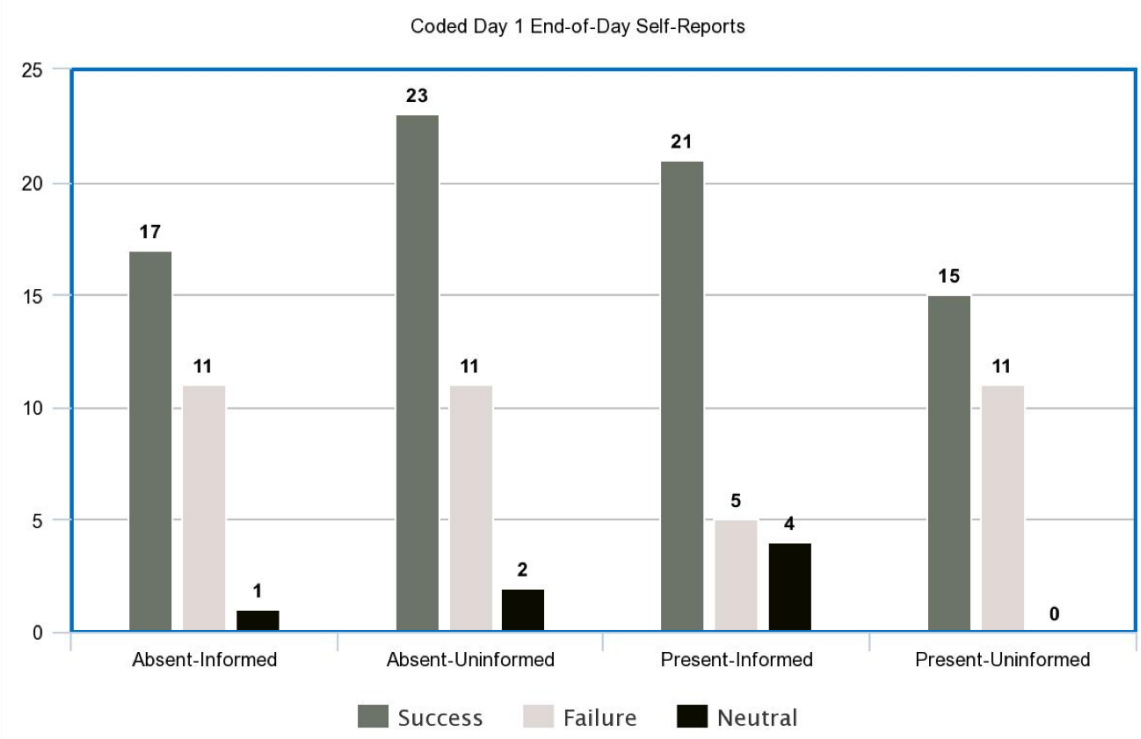


Figure 3.9: **Altered Perception: Self-reported failures and successes coded from the end-of-day briefing for day 1.** Significantly fewer failures reported in the Present-Informed condition, potentially due to attribution of failures on the deception.

the non-existent deception, such as this excerpt from the end-of-day briefing: *"This network was filled with deception and I spent the majority of the day going down rabbit holes that led me nowhere."* Outcomes of this study suggest that future experiments designed to assess the effect of psychological deception (when no cyber deception is present) should utilize a real network, or ensure that the simulated network has enough realistic messiness, mistakes, imperfect users, and policy mismatches, such that *real* things can be misattributed to deception. The simulated network for the Tularosa Study did not include these features, and thus, did not provide evidence of the potential effectiveness of claiming deception is present, when it is not.

The most commonly identified vulnerability reported by participants that was designed into the simulated network as easy attack vector is a vulnerability in the

Microsoft implementation of the Server Message Block (SMB) protocol, denoted by entry CVE-2017-0144 in the Common Vulnerabilities and Exposures (CVE) catalog. Many participants also referred to it by the Microsoft security bulletin identifier MS17-010. EternalBlue was a well-publicized exploit at the time of this study that exploits this vulnerability. The Mattermost reports were coded to identify when EternalBlue was reported as successfully performed by participants. The Present conditions have fewer EternalBlue successes reported, while the Informed conditions have fewer failures reported. Some of the decoys appeared to be vulnerable to EternalBlue. This further supports hypothesis $H1$ that the decoys impeded forward progress, since a successful EternalBlue exploit was a common tactic for progressing further in the cyber kill chain. The Informed condition reporting fewer failed EternalBlue exploits further supports the idea raised in Section 3.1.3 that the Informed conditions may be, correctly or incorrectly, blaming failures on the deception, and are therefor less likely to label/report it a failure. The results were as follows:

- Absent-Uninformed: 5 participants with a self-reported EternalBlue exploit success; 14 reported successes; 2 failures
- Absent-Informed: 4 participants with a self-reported EternalBlue exploit success; 15 reported successes; 1 failures
- Present-Uninformed: 3 participants with a self-reported EternalBlue exploit success; 15 reported successes; 5 failures
- Present-Informed: 6 participants with a self-reported EternalBlue exploit success; 8 reported successes; 0 failures

We also used the community rules for Suricata, another open source IDS, to detect the EternalBlue exploit in the PCAP data. While it is still possible that some launches of the exploit were missed, these results should be closer to ground truth than the

self-reports in regards to the exploit being launched against a real machine. When analyzed with Dunn Kruskal-Wallis multiple comparison we see a significant difference of the Absent-Uninformed (control) condition generating more EternalBlue exploit alerts than the Present-Informed condition ($p = .050$), and the Absent conditions generating more alerts than the Present conditions (Kruskal-Wallis chi-squared = .0697, $p = .014$). This further supports a decrease in forward progress in the Present conditions. Again we see the pattern of the least number of participant making forward progress by this metric in the Present-Informed condition. The numerical results were as follows:

- Absent-Uninformed: 17 participants generating an EternalBlue alert; 147 alerts generated
- Absent-Informed: 10 participants generating an EternalBlue alert; 107 alerts generated
- Present-Uninformed: 8 participants generating an EternalBlue alert; 87 alerts generated
- Present-Informed: 6 participants generating an EternalBlue alert; 26 alerts generated

It is worth reiterating the intrusion alert findings in this section, since attacking a decoy is another measure of altered perception. In the Present conditions the participants often perceived that the decoys were real vulnerable machines. Unsurprisingly, we see this even more in the Present-Uninformed condition, where due to confirmation bias (discussed more in Chapter 4), participants have little reason to question the veracity of the machines. In looking at the means of the number of intrusion alerts in the two conditions, while not statistically significant, we see a much larger mean in the Present-Uninformed condition caused by some outliers triggering 1000+

intrusion alerts. The Uninformed participants are much more likely to perseverate on a particular machine they think they should be able to exploit, even after many failed attempts.

We also used the Mattermost self-report data to label and count the number of individual machines that each participant misidentified, e.g., thinking a real machine was a decoy or vice versa. This is a clear measure of perception versus reality and demonstrates the progression of misidentification across the different conditions. This is an important metric because incorrectly identifying a real machine as fake, can lead to an attacker ignoring a true vulnerable target, and incorrectly identifying a decoy as real, can lead an attacker to waste time and resources on an irrelevant target giving defenders more information about the attacker and more time to rally defenses. There were a total of 248 misidentifications observed in the Day 1 Mattermost data accross all conditions with 109 participants contributing Mattermost reports on Day 1. Other misidentifications, which included incorrect operating system identifications leading to mismatched exploit attempts, as well as exploit attempts or high-value target judgment placed on non-existent IP addresses accounted for a total of 6 of the misidentifications, half in the Absent-Informed condition and half in the Present-Informed condition. This is consistent with the information that deception may be present leading to more errors by the cyber attacker. In both Present conditions, over 95% of the misidentifications were attributed to judging a decoy to be a real machine (by determining it to be a high-value target or attempting to exploit it). Participants were not specifically asked to document which machines they thought might be real or decoys to reduce introducing bias into the experimental design. Results displayed in Figure 3.10) demonstrate that, again, the combination of presence of decoys and information about deception shows the biggest effect, supporting hypothesis $H2$. There may also be a correlation between an increase in misidentifications and increased confusion or frustration supporting hypothesis $H4$.

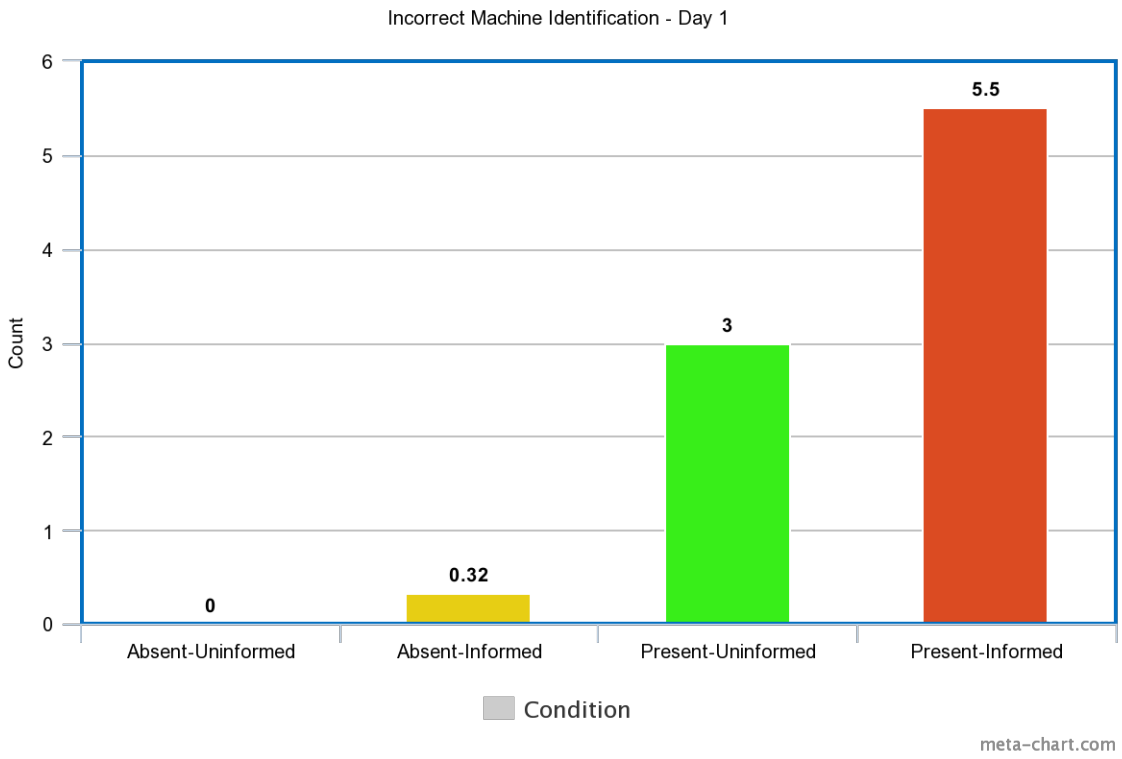


Figure 3.10: **Perception versus Reality: Average number of machines misidentified per participant in each condition.** Results suggest that both the presence of information and the information about the deception can effect misidentification of machines. A total of 254 assets were incorrectly identified acrosss all participants.

3.4 Cognitive and Emotional State

Next we analyzed the self-report likert scale responses provided by participants at the end of each day at part of the Cyber Task Questionnaire (CTQ) to address how the experimental treatment effected cognitive and emotional state. Recall that this data is part of the Human Subjects Research (HSR) data, and thus was not collected from participants who opted out. There was one session that started late (but still had time to completed the cyber task), and was dismissed before participants could finish the end-of-day questionnaires. For the analysis of the data discussed in this section there were 120 participants data available. For analyses involving Day 2 this was further reduce by the number of participants who did not return on Day 2. This

information will eventually be linked to various measures of success to determine if the cognitive state of participants effected their performance in the cyber penetration task. This can help inform cyber defense strategies in two ways, 1) determining usefulness of defenses causing or exacerbating a specific cognitive state, and 2) linking metrics that are measurable in the cyber data with changes in cognitive state. These extension remain future work.

One research question that we begin to address with the CTQ involves the participants' belief in deception given the manipulated two independent variables: Presence of cyber deception (absent versus present) and Information about deception (uninformed versus informed). Once the performance metrics are completed future work will correlate this belief in deception with performance on the cyber task.

Responses to the belief in deception question were coded using the following rating system: 1 = No, definitely no deception; 2 = Probably not, leaning toward no; 3 = Unsure, equal yes/no; 4 = Probably, leaning toward yes; 5 = Yes, definitely deception. Two raters completed the scoring, and scores were averaged across raters for analysis. Inter-rater reliability showed satisfactory reliability for Day 1 ratings (83% agreement, Cohen's $\kappa = .77$). At the end of each day, participants also reported the extent to which they felt confused, self-doubt, confident, surprised, and frustrated with the task on scales from 1 to 5.

3.4.1 Between-Group Differences on Day 1

We consider between-group differences on Day 1 to answer the research questions introduced in Figure 2.1 listed as *Planned Comparisons Between Groups*. Significant differences were evident regarding participant belief in deception presence, as follows. A two-by-two between-subjects ANOVA (Cyber Deception Presence: absent versus present by information: uninformed versus informed) showed that there was a main effect for Presence, $F(1, 61) = 12.36$, $p < .001$, where those in the Present conditions

reported more significantly belief in deception compared ($M = 3.60$) to Absent ($M = 2.19$), $p < .001$. There was also a non-significant trend for information, such that those Informed about deception tended to suspect more deception (M=3.21) compared to those Uninformed conditions (M=2.58), $p = .125$. See Figure 3.11 for comparisons between each of the four experimental conditions.

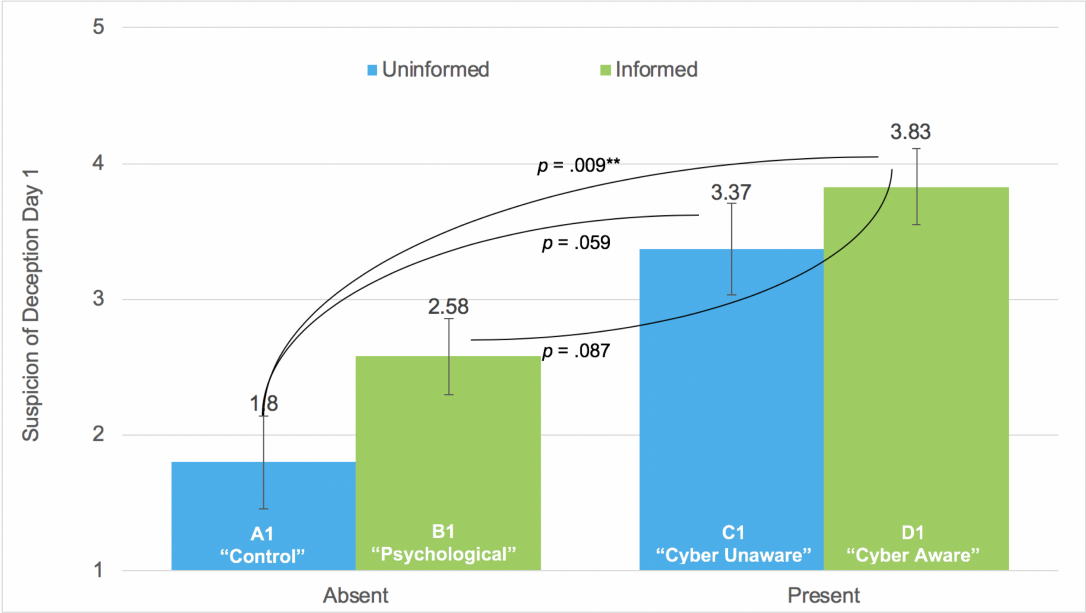


Figure 3.11: **Cognitive and Emotional State: Day 1 between-group differences in belief of deception.** Results suggest that both information and presence of deception have effect. There is a significant cumulative effect of Information and Presence, such that factual information in the presence of cyber deception instills the greatest belief in the presence of deception.

On Day 1, participants in the Present-Informed condition reported significantly higher confusion score than those in all other conditions (See Figure 3.12). Since the likert scale information is ordinal and discrete, with a limited range, we ran the non-parametric Kruskal-Wallis statistical test. This found a significant difference in confusion with $p = 0.014$ and Kruskal-Wallis Chi-squared of 10.687. This was followed by pairwise comparisons using Dunn test which indicated a significant difference between the Absent-Uniformed (Control) condition and the Present-Informed condition

($p = 0.007$) supporting our hypothesis $H4$. When analyzed further, we find that both Information and Presence of deception has a significant effect on self-reported confusion. Both the Informed participants have higher confusion (Kruskal-Wallis chi-squared = 5.047, $p = .025$) compared to Uninformed and those who had decoys Present have higher confusion (Kruskal-Wallis chi-squared = $6.47, p = .011$) than those who had no decoys.

This further supports the findings in our host-based deception experiment Shade et al. [2020] where we found a significant difference in the self-reported confusion when comparing Absent-Uninformed and Present-Uninformed (which were the only two conditions in the Moonraker Study.) Combing these results from the two separate studies supports $H2$ and the idea that the presence of cyber deception causes increased confusion, regardless as to whether the participants are informed of the presence of deception.

The Tularosa data also indicates significantly more self-reported surprise in the Informed condition compared to the Uninformed conditions (Kruskal-Wallis chi-squared $= 4.066, p = .044$, Figure 3.12). This is interesting because it would be easy to assume that if a participant knows to expect deception, they will be less surprised when something unexpected happens. Instead these findings indicate that information that deception is present can cause increased surprise, which is tangential the findings in our Moonraker Study Shade et al. [2020] where we found significantly more difference in the self-reported surprise when comparing Present-Uninformed to Absent-Uninformed. The Moonraker Study design forced a more narrow path to success than the Tularosa Study and included a different type of cyber deception; either reason could explain the difference in findings.

The Moonraker Study results indicate a significant difference in cyber task completion rate with participants in the condition with increased confusion completing the cyber attack task less often. Future work will determine whether the Tularosa

data results are consistent and the condition in which we measured an increase in confusion also shows a reduced completion rate. It has previously been assumed that information that deception is being used should be withheld from attackers to be the most effective. In contrast, the results discussed above match our observations in our Pilot Study Ferguson-Walter et al. [2017] and help us further to build a case that knowledge that deception is being used against you, can actually multiply many of the effects since attackers start to see deception even where it is not, thus increasing confusion and surprise above what is caused by the cyber deception.

3.4.2 Correlations Between Reported Cognitive and Emotional States

Based on our Pilot Study, frustration is an emotional state theorized to be effected by the experimental treatment. Although there was no significant difference in self-reported frustration scores across conditions, the Day 1 correlation results are consistent with a difference in cognitive effects across conditions that mirroring examples documented in the pilot studies. The Absent conditions showed significant positive correlations between reporting frustration and confusion ($r = .574$, $p < .001$ for Absent-Uninformed) and $r = .454$, $p < .05$ for Present-Informed)). This could indicate that the task itself had confusing aspects which led to frustration among participants. Both Informed conditions showed significant positive correlations between reporting frustration and self-doubt ($r = .391$, $p < .05$ for present and $r = .583$, $p < .01$ for absent), and negative correlations between self-doubt and confidence ($r = -.536$, $p < .01$ and $r = -.511$, $p < .01$ respectively). This may indicate that information of the presence of deception (regardless of the veracity of the statement) can cause self-doubt to the participant, which affects confidence. Both Present conditions showed significant positive correlations between reporting frustration and surprise ($r = .563$, $p < .01$ for uninformed and $r = .708$, $p < .001$ for informed), as well as self-doubt and confusion ($r = .535$, $p < .01$ for uninformed and $r = .381$,

$p < .05$ for informed), indicating that cyber deception may cause a cyber attacker surprise, and confusion about the network may lead to increased self-doubt when attacking. In the Present-Uninformed, suspicion of deception was negatively correlated with self-doubt ($r = -.535$, $p < .05$), mirroring what was discovered in the pilot studies Ferguson-Walter et al. [2017]; participants could be attributing task performance to feelings of inadequacy instead of to deception deployed on the network. However, in the Present-Informed condition, a positive correlation was observed with confidence ($r = .490$, $p < .05$), which could be reflecting the fact that since they were informed, and likely found deception evident on the network, they felt confident in their ability to negotiate it. In fact, an opposite, though marginal, effect was observed in the Absent-Informed condition, suggesting that being informed about deception but not finding anything on the network to support that claim resulted in less confidence about the attack strategy.

3.4.3 Within-Group Differences Between Days.

We also looked at the within-group differences by examining the change in self-reported cognitive and emotional state from Day 1 to Day 2. While there was not a significant different in self-reported frustration between-groups on Day 1, the change in self-reported frustration on Day 2 shows a significantly higher amount of frustration in the AUPU condition than the other conditions (Kruskal-Wallis chi-squared $= 10.526$, $p = .015$). The condition changing from Absent-Uninformed (Control) on Day 1 to Present-Uninformed on Day 2 showed increased frustration while the other conditions changing from psychological and/or cyber deception to Absent-Unaware on Day 2 indicated reduced frustration. The difference between self-reported frustration between Day 1 and Day 2 was significant when comparing AUPU to AIAU ($p = .025$), to PIAU ($p = .025$), and to PUAU ($p = .043$). We believe this difference is caused by the addition of decoys for this condition on Day 2. We see a decrease in

the other conditions from Day 1 to Day 2 due to a learning effect and an absence of decoys on Day 2. The other conditions have no decoys present in Day 2 and tend to report less frustration than Day 1.

In addition, the change in self-reported surprise from Day 1 to Day 2 showed a trend with Kruskal-Wallis chi-squared = 7.2963, and $p = .063$ between the AUPU condition and both the Informed conditions (AIAU, $p = .074$ and PIAU, $p = .068$). With the addition of decoys on Day 2 the AUPU condition reported increased surprise, while the conditions that were informed on Day 1 showed a decrease in surprise on Day 2.

Frustration has been shown to reduce effectiveness of cyber operators Dykstra and Paul [2018], as well as other stressors like fatigue and increased cognitive workload. While confusion and surprise was not included (since they are seldom factors without deception present), we believe that they will have similar effects to frustration and other types of stress. This related work focused on measuring and reducing stress to improve cyber operator performance supports our hypothesis that using deception to increase stress can reduce the effectiveness of cyber operators.

3.4.4 Word Count Analysis

In this section we discuss the results and implications of a word count analysis performed on the self report data. To better scope the results to align with the other findings discussed in this chapter, we focus on day 1 self-reports only. For each participant this includes all Mattermost reporting from Day 1 and the Day 1 end-of-day Red Team Briefing. It is important to note that there was no significant difference in the number of Mattermost messages across conditions. Figures 3.17 and 3.18 display a word cloud showing the most frequent words for all participants across all conditions from the Mattermost chat logs and Red Team Briefing for Day 1, respectively. These two different types of reports tended to provide different types of information.

The real-time Mattermost chat logs are where we can observe mistakes being made, remade, and corrected, the reasoning behind actions, frustrations and their causes. The end-of-day Red Team Briefing is where we tend to observe a summary of the strategy taken, as well as final outcomes and details of only the most memorable successes, failures, and frustrations. The differences between the types of reports can be seen in the difference Word Clouds. For example, notice the term *deception* appears in the end-of-day word cloud, but not the real-time word cloud, suggesting that this is something participants reasoned about retrospectively.

Upon cursory review of word counts, we selected keywords of interest which seemed to have many occurrences. We then considered the root of these words, and noted other versions and spelling that would need to be grouped with the root. For example, one root word of interest is *deception*, and when we counted for deception we included: *deceive, deceptor, deceptive, decoy, deceit, deception(s), and honeyX* (where X allows us to count honeypot(s), honeynet(s), etc.). The root words selected were: *deception, confusion, difficult, easy, failure, success, frustration, interesting, real, unable, exploit*. The results are from a binary indication of whether a participant used a term relating to the root word in any messages for that day. As elsewhere in this chapter, we use the Kruskal-Wallis statistical test to compare across conditions. When comparing more than two conditions performed post-hoc pairwise comparisons using Dunn (1964) Kruskal-Wallis multiple comparisons with p-values adjusted with the Benjamini-Hochberg method. Notable findings are described in Table 3.2.

Statistical differences and trends for the keyword of *Deception* on Day 1 align with the coded CTQ responses on suspicion of deception, discussed above, that the Present-Informed condition had the most belief that deception is present, and the Absent-Uniformed the least. They also align with the self-reported likert scores for confusion indicating that Present-Informed had the most confusion. These findings supports the idea that keyword counts can provide us with useful information for

concepts that may have not been directly address in the questionnaires. Word counts alone are likely not a strong enough indicator to confirm a hypothesis, but when grouped together to further confirm other results, they can be a powerful enhancement to further the narrative.

Success is mentioned significantly more for participants in the Present conditions ($p = .004$). This supports the hypothesis that the decoys are providing an easy target, and giving participants a false sense of success.

Another finding indicated that *Real* is mentioned significantly more for participants in the Informed conditions ($p = .026$). This is likely due to the informed participant taking the time to question, investigate, and report on what they perceive as real or fake. This supports the hypothesis that information about deception, even when it isn't really there, can distract (and thus delay) attackers from their true goal (which is not to determine real from fake).

We also note a non-significant trend that *Easy* is mentioned significantly more for participants in the Uninformed conditions ($p = .051$). This could further support the idea that psychological deception makes the task harder for participants, regardless as to whether cyber deception is actually present. Based on observations from the Pilot, this could be due to increase self-doubt and paranoia both in the case when the deception is really present and when it isn't.

3.5 Discussion of Data Analysis Results

In this chapter we discussed results from the data analysis that has been performed on the cyber data collected from the Tularosa Study. This analysis is focused on the cyber data and self-reports and only examines the conditions on Day 1 (unless otherwise stated). We do not examine the cognitive battery, the physiological measures, or fully evaluate persistence effects evident on Day 2. This is left for future work. A collection of all the significant findings discussed above can be seen in Table 3.3.

Root Term	Higher	Lower	p-value
Deception	Present-Informed	Absent-Uninformed	p = .00005***
		Present-Uninformed	p = .0005***
Deception	Informed	Uninformed	p ¡ .001***
Deception	Absent	Present	p = .032*
Deception	Absent-Informed	Absent-Uninformed	p = .068
Deception	Present-Informed	Absent-Informed	p = .056
Confusion	Present-Informed	Absent-Uninformed	p = .001**
		Absent-Informed	p = .001**
		Present-Uninformed	p = .016*
Confusion	Present	Absent	p = .002**
Success	Present-Informed	Absent-Uninformed	p = .044*
Success	Present	Absent	p = .012*
Success	Present-Informed	Absent-Informed	p = .057
Real	Informed	Uninformed	p = .026*
Exploit	Informed	Uninformed	p = .076

Table 3.2: **Word Count Analysis: Notable differences in number of subjects per condition having a self-report containing keywords on Day 1.** Significant differences are indicated as $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$.

Since our data is mostly non-normal, non-parametric test were usually used, which can have a lower power. Most notably the results are consistent with:

- H1: Presence of decoys cause delays in forward progress and increases detection.
- H2: A combination of the presence of deception and the knowledge that it is in use has the largest effect on cyber behavior and can cause increased confusion and surprise.
- H4: Information that deception might be present can effect the attackers cognitive state, decisions, and behavior.

While the results presented do not provide a strong argument Hypothesis $H3$ – the value of providing information that deception might be present when it is actually absent, based on observations and previous pilots, we still think this is a valuable concept for future research. The nature of the simulated network range used for the Tularosa

Study, did not have enough of the natural messiness provided by a real network with real users. We argue that this messiness is precisely what is needed to provide the plausible deniability and uncertainty that make the psychological deception effective.

Hypothesis H1:			
Data Source	**Higher Mean**	**Lower Mean**	**p-value**
Keystroke Count	Absent	Present	$p < .05$*
Commands w/ real IPs	Absent	Present	$p < .001$**
Bytes to real IPs	Absent	Present	$p < .05$*
Snort Alert Count	Present	Absent	$p < .01$**
EternalBlue Alerts	Absent	Present	$p < .05$*
Reported Exploit Successes	Absent	Present	$p = .076$ ns
Hypothesis H2:			
Data Source	**Higher Mean**	**Lower Mean**	**p-value**
Self-reported Confusion	Present-Informed	Absent-Uninformed	$p < .01$**
Snort Alert Count	Present-Informed	Absent-Uninformed	$p < .05$*
Time to first decoy alert	Present-Uninformed	Present-Informed	$p < .05$*
Decoy Touch Alerts	Present-Informed	Present-Uninformed	$p < .01$**
Decoy Scan Alerts	Present-Informed	Present-Uninformed	$p < .01$**
Decoy Probe Alerts	Present-Uninformed	Present-Informed	$p < .001$***
# with Intrusion Alert	Present-Uninformed	Present-Informed	$p < .01$**
All Decoy Alerts	Present-Informed	Present-Uninformed	$p < .001$***
EternalBlue Alerts	Absent-Uninformed	Present-Informed	$p < .05$*
Reported # of vulnerabilities	Absent-Uninformed	Present-Informed	$p = .061$ ns
Hypothesis H4:			
Data Source	**Higher Mean**	**Lower Mean**	**p-value**
Self-reported Surprise	Informed	Uninformed	$p < .05$*
Self-reported Confusion	Informed	Uninformed	$p < .05$*
Self-reported Confusion	Present	Absent	$p < .05$*
Suspicion of Deception	Present-Informed	Absent-Uninformed	$p < .01$**
Self-reported frustration change score across days	AUPU	AIAU PIAU PUAU	$p < .05$* $p < .05$* $p < .05$*

Table 3.3: **Summary of significant findings.** Significant differences are indicated as ***$p < .001$, **$p < .01$, *$p < .05$, ns for a non-significant trend.

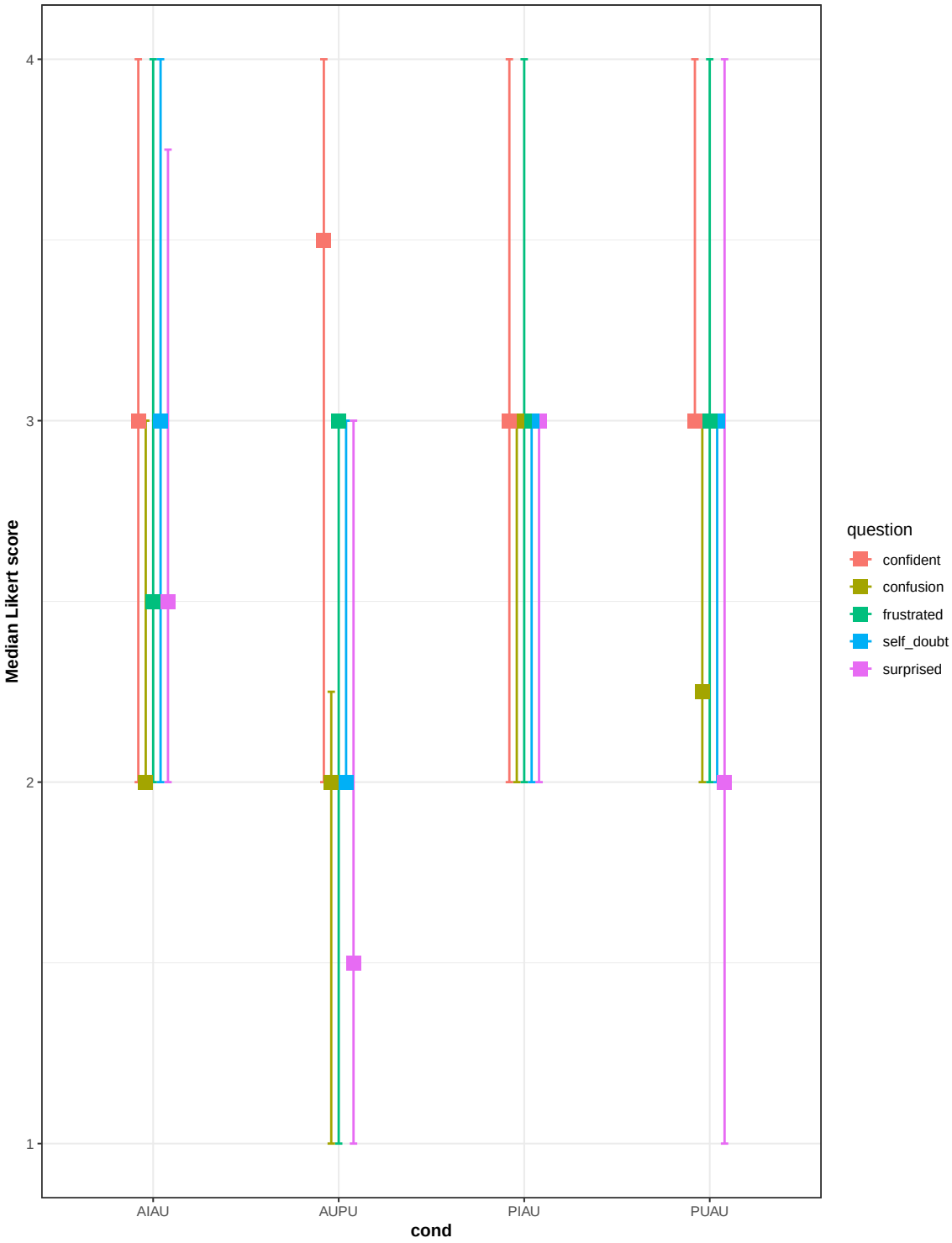


Figure 3.12: **Cognitive State: Day 1 self-reported cognitive and emotional state.** This is self-reported data provided at the end of day 1 indicates that the Present-Informed condition had significantly more surprise than the control condition (Absent-Uniformed) and significantly more confusion than all other conditions.

TSQ Preliminary Results: Absent-Uniformed Condition

Correlations Between Variables for AUPU Condition on Day 1						
	Confusion	Self-Doubt	Confidence	Surprise	Frustration	Deception
1. Confusion	—					
2. Self-Doubt	.245	—				
3. Confidence	-.050	-.317	—			
4. Surprise	.339	.471**	-.230	—		
5. Frustration	.574***	.115	-.047	.271	—	
6. Deception	.553	-.159	-.334	.351	.061	—

Note. $p < .05$* $p < .01$** $p < .001$***. Significant relationships are shaded for ease of viewing.

Figure 3.13: **Day 1 correlations between self-reported cognitive states for Absent-Uninformed (control).** Notable results suggest that in the Absent-Uninformed condition *frustration* is related to *confusion* and *surprise* is related to *self-doubt*.

TSQ Preliminary Results: Absent-Informed Condition

Correlations Between Variables for AIAU Condition on Day 1

	Confusion	Self-Doubt	Confidence	Surprise	Frustration	Deception
1. Confusion	—					
2. Self-Doubt	.312	—				
3. Confidence	-.514**	-.511**	—			
4. Surprise	.472**	-.111	-.170	—		
5. Frustration	.198	.583**	-.199	.189	—	
6. Deception	.256	.146	-.449	.095	.285	—

Note. $p < .05$* $p < .01$** $p < .001$***. Significant relationships are shaded for ease of viewing.

Figure 3.14: **Day 1 correlations between self-reported cognitive states for Absent-Informed.** Notable results suggest that in the Absent-Informed condition *frustration* is related to *self-doubt*, *lack of confidence* is related to *confusion* and *self-doubt*, and *confusion* is related to *surprise*.

TSQ Preliminary Results: Present-Uninformed Condition

Correlations Between Variables for PUAU Condition on Day 1

	Confusion	Self-Doubt	Confidence	Surprise	Frustration	Deception
1. Confusion	—					
2. Self-Doubt	.535**	—				
3. Confidence	-.236	-.240	—			
4. Surprise	.016	.197	-.424*	—		
5. Frustration	.338	.376	-.565**	.563**	—	
6. Deception	-.454	-.535*	.038	-.048	.069	—

Note. $p < .05$* $p < .01$** $p < .001$***. Significant relationships are shaded for ease of viewing.

Figure 3.15: **Day 1 correlations between self-reported cognitive states for Present-Uninformed.** Notable results suggest that in the Present-Uninformed condition *frustration* is related to *lack of confidence* and *surprise* and *confusion* is related to *self-doubt*.

TSQ Preliminary Results: Present-Informed Condition

Correlations Between Variables for PIAU Condition on Day 1

	Confusion	Self-Doubt	Confidence	Surprise	Frustration	Deception
1. Confusion	—					
2. Self-Doubt	.381*	—				
3. Confidence	-.244	-.536**	—			
4. Surprise	.292	.337	-.166	—		
5. Frustration	.454*	.391*	-.265	.708***	—	
6. Deception	-.070	-.211	.490*	-.086	-.187	—

Note. $p < .05$* $p < .01$** $p < .001$***. Significant relationships are shaded for ease of viewing.

Figure 3.16: **Day 1 correlations between self-reported cognitive states for Present-Informed.** Notable results suggest that in the Present-Informed condition *frustration* is related to *surprise*, and *lack of confidence* is related to *self-doubt*.

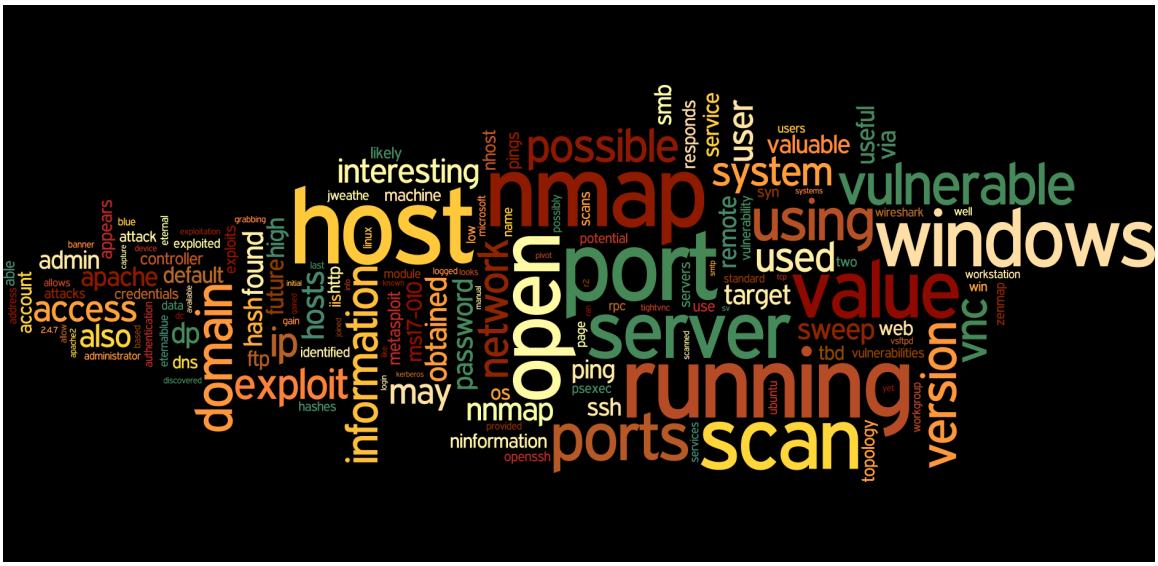


Figure 3.17: **Word Cloud: Displaying the top 150 frequent words across all conditions used in Mattermost reporting on day 1.**

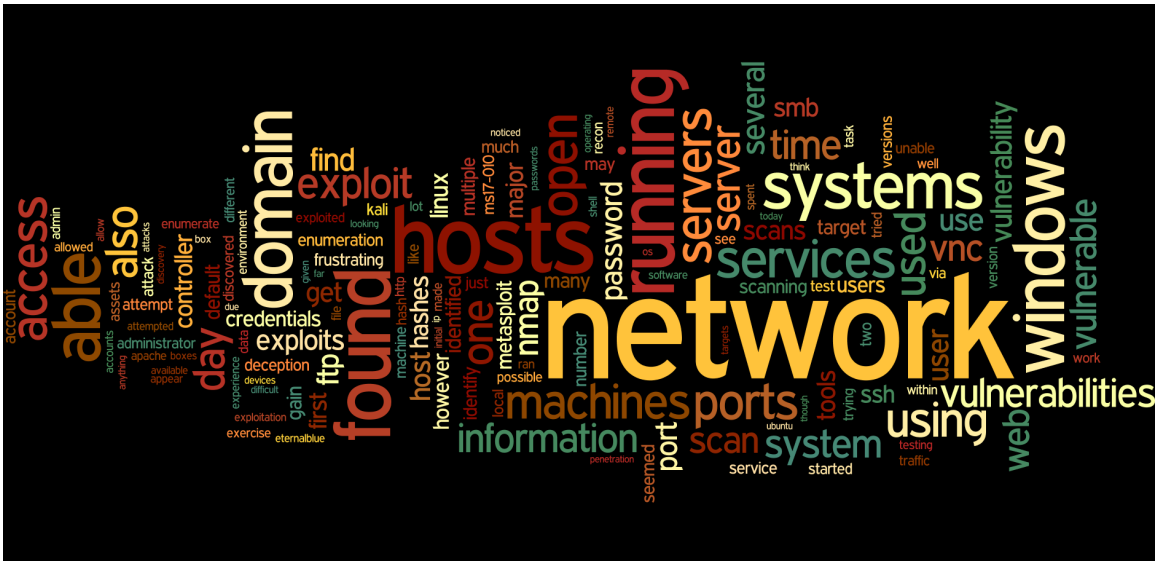


Figure 3.18: **Word Cloud: Displaying the top 150 frequent words across all conditions used in end-of-day reporting on day 1.**

CHAPTER 4

OPPOSITIONAL HUMAN FACTORS

Oppositional Human Factors (OHF)[1] is a new concept we introduced as the science of reversing traditional human factors and usability recommendations to make interaction with technology difficult for those with malicious intent. Inverting human factors can aid in cyber defense by flipping well-known guidelines and using them to degrade and disrupt the performance of a cyber attacker Gutzwiller et al. [2018]. There has been significant research on how cyber defenders currently perform tasks and how information should be presented to operators, cyber defenders, and analysts to make them more efficient and more effective. However, we can actually create these situations just as easily as we can mitigate them.

Oppositional human factors are a new way to apply well-known research on human factors to disrupt potential cyber attackers and provide much-needed asymmetric benefits to the defender. We will introduce preliminary experimental findings that provide new evidence of traditional attentional and decision-making biases present in red teamer behaviors. For example, well-known biases such as confirmation bias may disrupt red team decisions and goals, and simultaneously increase their risk of detection. Disrupting attention and decision making are two conceptual components we will describe in a growing OHF framework for cyber defense. We note that oppositional human factors compliment cyber deception practices and, in some cases, go beyond what would traditionally be defined as "deception" because the techniques

[1]This chapter is based on published work: R. Gutzwiller, K. Ferguson-Walter, S. Fugate, and A. Rogers, *"Oh, Look, A Butterfly!" A Framework for Distraction Attacker to Improve Cyber Defense*, Human Factors and Ergonomics Society (HFES), 2018.

can disrupt the thought processes, work processes, and mental state of attackers in novel ways and can have long-lasting and dramatic effects on the ability to successfully attack networks.

4.1 Oppositional Use of Human Factors

In conducting a cyber attack, operators may multitask when any two or more demands for their attention occur at the same time. Simultaneous mental demands emerge when multi-tasking and create mental resource interference Arrington and Logan [2005]. One could force an attacker to incur this kind of interference by engineering each task in the environment to impose loading of the same resource, e.g., overloading the auditory or visual channel. In this way any tasks that are normally separated in time could be made to happen at the same time, forcing these additional attentional costs which can be predicted explicitly by the multiple resource theory model of attention Wickens [2008]. Using the model "in reverse," we would actually seek to increase the interference of sets of tasks rather than decrease it.

As a simple example which addresses oppositional human factors in the context of existing defense techniques, an attacker who scans a network containing decoys will be faced with the tasks of assessing vulnerabilities in a larger number of potentially vulnerable systems. However, even if the number of decoys is large, this multitasking task, e.g., running a large number of banner-grabbing scripts, is unlikely to present a significant cognitive burden. If, however, our systems respond inconsistently, provide corrupted responses, or provide responses which change over time, an attacker is more likely to forego automated or scripted interactions and revert to manual instance-by-instance tests. Such an approach has many implications, but of interest here is the potential for the attacker to intentionally choose to manually perform multiple simultaneous actions, potentially overloading available cognitive resources and certainly

wasting some time interacting and validating vulnerability scans of systems which are specifically intended to increase attacker cost.

Not all cyber attackers work in the realm of multiple task performance, where multiple resource theory predictions hold. And often, to deal with multiple co-occurring tasks, human operators attempt to separate tasks and then switch rapidly between them, leading to more sequential information processing Salvucci and Taatgen [2011] as human attention appears limited by a single channel Liao and Moray [1993]. These task switches are costly when they are forced Kiesel et al. [2010], Monsell [2003], but also when they are made under voluntary control Arrington and Logan [2005], which requires the attacker to make on-the-fly judgments about where to allocate attention. These conditions also often lead to cognitive tunneling, a phenomena in which an operator becomes fixated on one ongoing task in the milieu e.g., diagnosing why a landing gear may be malfunctioning, and ignores or forgets to monitor and do work in others, even when they are important e.g., ensuring the plane is at a stable altitude. By inverting guidance on avoiding cognitive tunneling, we may be able to force the attacker to tunnel into worthless tasks by making them seem interesting or important — or at the very least disrupt attention if attacker focus is on tasks that progress the attack.

For example, a decoy system's purpose might be to allow an attacker to begin downloading a large file but then to degrade or to cease this file transfer when the attacker begins interacting with other, more valuable, information systems. The opportunity costs associated with initiating interaction with a new, unknown system versus the "known" value of the existing file transfer is likely to lead the attacker into tunneling into the task of determining the reason for the "high value" file transfer failure.

4.2 Attackers Exhibit Framing and Attentional Tunneling

Analysis of transcripts from our pilot studies suggest that many forms of cognitive bias were present and that the use of deception played an important role in creating bias where there would not have been bias otherwise. We have spent some time attempting to deconstruct these transcripts in respect to various cognitive bias effects, both attentional effects and decision making biases. While it can be difficult to precisely identify actual effects or to properly identify which form of bias is present due to deception (if any) we believe that it is worthwhile to examine several key examples.

When deception was present and the participants were not informed of the deception, the red team participants seemed to have a tendency to blame themselves for perceived discrepancies. This can be viewed as an *illusion of control* bias, where a perceived good outcome based on a poor decision can induce a false feeling of control over the situation Langer and Roth [1975].

> **Red Team Member A (RT-A):** "That did not work. Very disappointed. That is a very stable exploit. Very well known. Very well documented. I really thought I had something there, although while I'm at it... if we had one SP 0..."
> **Red Team Member B (RT-B):** "Try another exploit at it."
> **Red Team Member A RT-A:** "I might have done something dumb. Hold on."

In this exchange RT-A blames themselves when a well-known and stable exploit fails to work. Once the red team was informed of the use of deception, in future studies, they tended to make very different assessments from nearly identical behavior of systems on the network. In the not informed condition they simply move on to other systems or other exploits. However, in the informed condition they spend significant amounts of time attempting to find indicators that systems are real or fake. This suggests a *framing effect*, where people exhibit changes in behavior or decision making associated with the way a problem or situation is described, even though

the underlying environment is still the same Tversky and Kahneman [1981], which then appears to lead to the presence of *attentional tunneling*. The framing effect is all about how people view information: Windows XP Service Pack 0 originally framed as gain by the red teamers, as they are vulnerable, however once informed of deception this was framed as potential loss, assuming that vulnerable appearance indicates a decoy that will lead to detection.

Other examples from this study's transcripts also suggest that there may also be aspects of other cognitive bias effects. In particular, the following exchange is telling:

> **RT-A:** "Very well. Very well cloaked, I got to tell you... I'm very impressed with this one because I know for a fact that that [user name] is not valid because it does not show up in the infrastructure. So I use that script to automate the process of validating all my user names and so I know it's fake, and so I connected to the machine and I'm greeted with what looks correct. It looks like any other Windows 7 client machine including the proper greeting."
> **RT-B:** "Let me, let me finish poking and prodding at these servers that I know are fake."

At this point the red team members have a potentially valid suspicion that the system is a decoy. However, the machine is real and the user name is actually valid. This is an example of the *anchoring bias*—the tendency to rely too heavily and "anchor" to one trait or piece of information when making decisions during the task Tversky and Kahneman [1974]. The participant is relying to heavily on the information given at the start of the study that deception might be present, and continues to insist the machine is a really impressive fake instead of real. We also observed many instances where red team confidence in a system being fake increased throughout the day even when no additional evidence was available. This is a clear example of *irrational escalation*, gradually increasing their estimation of the correctness of a judgment based on incomplete and insufficient evidence.

Sunk-cost fallacy effects, wherein people justify increased investment of time or effort based on ongoing prior investment in spite of new evidence (available to the participant) suggesting that the decision was probably wrong Arkes and Blumer [1985], may in some cases have led to other tunneling effects, as seen in the example below. There were several instances where the red team participants would continue to return to discussions and analysis of hosts that were suspected to be fake. In some cases they accurately assessed a system as fake but continued throughout the day to re-assess their own judgment. This behavior is not in line with the explicit goal given to exploit and exile the system, and is instead a distraction from that goal.

> **RT-A:** "And we had a collection of reasons for our suspicion. Strike one, as one of our team members had noted, was that for the [user name] it is very difficult for somebody to have a last name that's made out only of consonants. So that was strike one against that thing. Strike two: in one of our collection notes we found this thing had a different IP address. Now it has another IP address. Not necessarily a strike against it, but looks interesting. And then strike three was, this particular kind of scan gathers more information after negotiating with that unit, with that piece of equipment, and it came back with MYGROUP unknown, looked very suspicious. So I want to gather that same information..."

Each of these examples from our pilot studies support our claim that the use of deception is likely to create new opportunities for causing cognitive bias in attackers to aid cyber defense. The cognitive bias effects which were observed were not specifically intended in the design of the decoy systems. In other words, these effects were not the intent of the decoy system's design, but they could be. If such a decoy system designs were using oppositional human factors guidelines, each of these effects (and others) could be leveraged for maximum effect using all of the tools and measures available.

4.3 Attackers Exhibit Confirmation and Anchoring Biases

In the pilot studies, red teamers were asked to exfiltrate data as a primary task. One of the subtasks that attackers engage in is determining whether hosts are real or fake. Shifting emphasis to this decoy identification makes some sense from the perspective of an attacker, especially one who is told that deception is present. Effort expended early in determining which machines are decoys might save time later, if they are able to learn to accurately distinguish decoys from real systems. But it also suggests that if a real system appears to be a decoy, due to a particular misleading but distinguishing characteristics, this same effort may be beneficial for cyber defense, biasing attackers away from real systems. Further benefit to cyber defenders stems from delaying or causing attackers to reveal themselves by increasing their noise or luring them to interact with decoys.[2]

A *take-the-best* heuristic strategy, wherein an attacker notes the most salient and accurate characteristics to distinguish real from fake, may be useful precisely because it is fast and accurate (mostly, but not always). Still this heuristic must be developed over time and exposure, and attackers may not get the amount of feedback necessary to build this heuristic, or the feedback may be delayed enough that it does not help long-term. Unrelated characteristics of machines might be observed by the attacker, labeled as indicating either real or decoy systems, and then used to inaccurately identify real systems as decoys. The removal of feedback both in confirming a system is truly real, and in the lack of executing certain defensive behaviors against an attacker, may be a good thing at times as it slows the skill acquisition of the attacker.

We found several examples of red team members suffering from *confirmation bias*, the tendency to search for, interpret, focus on and remember information in a way

[2]This section is based on published work: R. Gutzwiller, K. Ferguson-Walter, and S. Fugate, *Are Cyber Attackers Thinking, Fast and Slow? Evidence for Cognitive Biases in Red Teamers Reveals a Method for Disruption* , Human Factors and Ergonomics Society (HFES), 2019 Gutzwiller et al. [2019].

that confirms one's preconceptions Einhorn and Hogarth [1978], when making host determinations. They anchored or fixated on their initial suspicions about hosts and sought to confirm these findings e.g., all the real hosts look like X, all the fake hosts look like Y, and the determination is made in confirmation of prior knowledge.

For example, a red team used a corporate employee directory to check hostnames as part of their attempt to determine real from fake. Attackers believed — incorrectly — host names were based on a "first.lastname" naming convention. Using the corporate directory would be a reliable way to determine fake versus real hosts. However, it only confirmed their hypotheses, and was not reliable in this case.

Once the red team determined a host was fake they also continued looking for more evidence to confirm this, including examination of whether the font was different between what they believed were fake versus real hosts. They also believed that connection timing was a differentiating factor.

When told deception may be present, some participants went further in noting how a host was suspicious due to a variety of other factors, including (1) the last name was made of only consonants and, assuming host names are real people's names, it was reported as *odd*, (2) the host had a different IP address than earlier attacker notes listed, and (3) the output from various network scan showing UNKNOWN made them suspicious, even though this can happen with real machines and operating systems.

The assessment process of real versus fake hosts was particularly interesting in the pilot study in which there were no fake machines. Attackers were told there might be deception, but ignored or failed to search for evidence pointing to machines being real, instead searching and finding evidence that could be used to suggest a host is fake. In examples provided above, because of various suspicions, a host became a target for the red team to determine it was fake and the *secondary task* overtook the primary task of data exfiltration. In this case the host was real, not fake, therefore,

not only were the attackers tricked at the identification stage, but they spent more time investigating the host for the express purpose of labeling it as a decoy.

In our analyses, attackers also developed uncertainty when they were told that deception was present. A form of anchoring bias appears in the examples above but is even clearer below. When the red team was told that deception may be present, they treated the results of the directory comparison as "truth". Some participants became highly certain of their particular divination of real from fake — as in the lookup of identified machines in corporate directories and cross-referencing names. One subject was completely confident that the lookup was an accurate measure and proceeded to focus significant energy on the machines deemed to be fake, in spite of the stated goal: to exploit real machines. The anchoring bias — fixation on the lookup methods — suffered by attackers seemed to lead to tunneling behaviors on determining real from fake hosts. Together, these biases result in a belief in false knowledge and a tendency to forego useful work in lieu of further hypothesis testing. With enough time in such a situation, the attacker may eventually care less about correctly identifying machines, but this also benefits the defender. This bias was something the red teamers noticed and acknowledged, but were still effected by, as evident by this answer to a post-exercise question about how knowledge of deception changed their strategy.

> **RT-A:** "I think that right off the bat we were swayed to that being one of our very first tasks — identify what we think is real, compare and contrast assets [...] instead of just targeting the most interesting assets, we actually spent time comparing them."

4.4 Biases of Cyber Attackers

In the current analysis of data from the pilot studies, we found evidence that cyber attackers exhibit classic attentional and decision-making biases. As discussed, we have also found evidence for many biases, take-the-best heuristics and other illusions that

may lead attackers astray, but more biases are likely at play. There are many biases which could play important roles in achieving OHF effects, though we suggest caution in assuming all have bearing on cybersecurity.

The ever-changing nature of technology may drastically change the tasks of cyber attackers. Even the differences between deep-state actors versus "script kiddies" are large enough that some types of biases may exist in one of the populations and not in the other. This deserves more study.

We believe we can exploit these behaviors for the betterment of cyber defense as part of an OHF strategy, and in combination with decoy usage. Understanding attention allocation and decision-making related to malicious activity is potentially only a few elements of an OHF toolkit for cyber defense. The ability to predict how humans behave in the case of exhibiting bias may also help discriminate between actors.

Finally, one could view cyber deception as a form of OHF as well, because a normal working assumption of interaction with an IT system is that it is truthful and transparent about its operations. Deception violates usability guidelines by (usually) hiding goals and states of the system from its users. A key assumption applies for many of the various attention and decision making OHF techniques discussed above: a defender or defensive system has the ability to control what a would-be attacker will be able to view.

Current computing systems tend to freely share and to even broadcast computing system information such as TCP port status, the existence and version of running services, operating system information, and a plethora of other forms of technical data useful for auto-configuration and interconnectivity between systems and services. Each of these information sources could be easily manipulated by a clever defender to mask true system state and behaviors to manipulate attacker knowledge of network and system state. Such manipulations would likely be key tools for manipulating

attacker biases, particularly if they can be demonstrated to be empirically effective via experimentation.

Lastly, an astute reader may wonder if examining and gaining a comprehensive understanding of attacker cognitive biases might eventually lead to attackers learning how to better address and suppress their own biases — thereby diminishing or removing the intended defensive effects of OHF techniques. However, people exhibit bias blind spots Ehrlinger et al. [2005], Pronin et al. [2002], and a key characteristic of many bias effects is their tenacity: simply knowing about the cognitive bias is unlikely to prevent the occurrence Friedrich [1996], West et al. [2012]. Specific training to *de-bias* participants has occasionally been shown as effective, but is still not widely available Shaw et al. [2018].

4.5 Summary

Cyber red team members exhibited several biases and heuristics in our pilots. A challenge is to categorize these aspects of behavior and work to control the environment of the attacker to disrupt them. Future experiments can then test whether oppositional techniques will disrupt realistic attacker behavior.

One of the uniquely desirable traits of human factors is the focus on the human performer in a system. By examining the cognition of attackers, situated in the cyber ecosystem, one can apply cognitive theory to potentially disrupt their activity. Human factors approaches are concerned with similar ideas, but subsumed by their goals of improving rather than disrupting performance.

In fact, it is an open question whether an oppositional view will remain useful in the eyes of the human factors community. Others have proposed that the field should focus on providing joy and pleasure *hedonomics* Hancock et al. [2005]. We suggest that OHF may disrupt the entire hedonomics hierarchy, from basic usability to pleasurable design and affective experience. Attackers waive assurances to good

(and certainly, pleasurable) system usability and design when they violate laws and user agreements. Further, use of OHF may in fact be a form of safety in practice for cybersecurity, as allowing attackers to operate within the perimeter of our networks with full control of their cognition is a dangerous allowance. It should be noted that designers are certainly aware, as are businesses, that humans have certain limitations which can be exploited (for good, or not-so-good; Nodder, Chris [2013]).

We focused on oppositional methods here because, similar to cyber deception, they have potential to be asymmetrically beneficial to defensive operations. Study of cyber defense and the people who perform it reveal defense as a domain highly affectable by human factors D'Amico and Whitley [2008], D'Amico et al. [2005], Gutzwiller et al. [2015], Mahoney et al. [2010], Mancuso et al. [2015], and we suggest the incorporation of OHF techniques may be a useful path forward. While studying the cyber attacker is relatively uncommon in the human factors community, in comparison to improvements for defender training, visualization tools, and communications, we argue that it is likely to bear fruit.

A review of the literature reveals that studying cyber attackers from a cognitive viewpoint appears to be missing experimentation and analysis. The attentional and decision making examples provided above are from our pilot study, which provides a deep perspective on the decision making process of the red teamers through talk-aloud protocol and audio transcribing. The Tularosa Study begins to more rigorously investigate how deception changes an attacker's behavior and performance. While data describing the decision process for the participants is more limited, it does provide a breadth of over 130 red teamers' experiences to analyze. Assessment of the cognitive biases exhibited by the vast number of participants in this study is detailed below.

Both the pilot studies and the Tularosa Study were designed to asses cyber deception which is just one technique in the OHF framework. Future experiments, which

are out of the scope of this thesis, will be needed to further investigate the effectiveness other techniques. Strategically forcing poor usability and inducing decision making errors in malicious actors could reduce the impacts or success of a cyber attack or even act as a deterrence. But first, one must answer: are decision making and attentional biases observed in cyber attacker behaviors? In using our data to search for the answer, we hope to gather information to craft situations which can be used to create and test the disruption of attacker cognition and detail its outcomes for network defense.

4.6 OHF Experimental Methodology

It was suspected that biases, being broadly applicable and pervasive across domain, will be evident in observations of attacker behavior in cybersecurity scenarios[3]. To address our hypothesis $H5$,the focus was on five high-prominence cognitive biases in decision making, selected from a survey list of over one hundred for their application to cybersecurity, and for their likelihood of being induced by the experimental conditions. While prior work described above using a different, but similar red team experiment found evidence of bias and distraction Gutzwiller et al. [2019, 2018], the methods were casual observation. They did not employ strict qualitative methods. They also had no apparent a priori biases in mind to evaluate. Both are a necessary and common component in qualitative data analysis, and they are provided here on a larger, more complex dataset which used professional red teamers, realistic networks, and manipulated deception techniques (Ferguson-Walter et al., 2019).

A rigorous scoring process, defined in the methods, was used in which carefully defined biases were pulled into a rulebook for scoring. Multiple raters used the rule-

[3]This section is based on unpublished work: R. Gutzwiller, K. Ferguson-Walter, C. Johnson, L. Guo and M. Major, *"Evaluating Cybersecurity Red Team Cognitive Decision-Making Biases"* , 2019 (in draft).

book to score the briefing and chat data, which allowed for assessment of inter-rater reliability. The main hypothesis here was that given biases pervasiveness in complex environments, our anecdotal observations and subject-matter expert (SME) opinions, and prior experiments Gutzwiller et al. [2019] that evidence of all decision-making biases would be found in the given sample. An additional hypothesis was based on the deception manipulation (i.e., the presence of fake cyber assets, such as servers) in whether participants were told deception was present. Some biases may arise around the manipulation; framing effects are created by giving the participants different information about the deception on the network and would be more likely in "informed" conditions (or conditions following prior 'informed' sessions). Confirmation bias may be more likely to bias behavior in the condition of informed as well, because participants are prompted to consider whether various network items are real; testing whether they are real or not is then ripe for confirmation-based testing given other information.

4.6.1 Data Selection

To examine for cognitive biases, text-based outputs from participants were assessed. The data would reveal enough to aid future development of more subtle behavioral methods for assessment, as have been done elsewhere. For the current study, the approach to qualitative data analysis involved the identification and coding of themes that appear in text passages from the daily briefing, and Mattermost (chat) responses.

4.6.2 Biases Selection

An initial comprehensive list of cognitive biases in decision-making was developed, using prior experience, prior literature reviews, and Wikipedia. Following the initial terminology curation, we confirmed original citations and definitions of each bias or effect.

Next, biases were evaluated for their cyber security relevance to both red team operations. Two further characteristics of the biases affected our final selections for coding. The first was (1) likelihood of bias emergence given the participants' tasks performed in Tularosa. To consider this complex question, the combined team of cognitive psychologists and cyber security SMEs reviewed all biases and determined whether they were applicable to the contexts of the experiment, and whether the exhibition of the bias would show up in text or chat reporting. Not all biases are traceable in subjects reporting or communication. Second, we (2) considered whether a bias was robust enough to show up in a noisy, realistic environment, and thus we ignored a few of the narrower, laboratory-derived or small-effect size biases.

Next, we limit the number of biases scored to avoid overload of raters given the amount of material and time, to seven. The biases chosen were (a) confirmation bias Nickerson [1998], (b) framing effect Tversky and Kahneman [1981], (c) anchoring bias Tversky and Kahneman [1974], (d) sunk cost fallacy Arkes and Blumer [1985], (e) the availability heuristic Tversky and Kahneman [1974], the default effect Johnson and Goldstein [2003], and illusory correlations Chapman [1967]. The list was then turned into a rulebook by providing definitions and coding instructions.

4.6.3 Participants, Materials, & Data Inclusion Criteria

The Tularosa dataset consisted of 138 participants who each produced the following open-ended text files. There were three briefing reports and two days of unstructured chat responses. Each were reviewed for all included participants and scored for each of the five biases. For consistency with the data analysis described in the previous chapter, we only discuss results of Day 1 here.

1. **Mattermost Day 1**, unstructured chat with real-time activity notation and timestamps. Participants were given instructions on how to use the chat and what content to send ahead of the experience. "When you learn potentially

useful information about target systems on this network you will immediately report this information to your team via your inter-net connected laptop using the Mattermost website" and told to report "The last 2 octets of the IP address. why you believe the host is interesting, how you obtained this information, estimate its value to future operations".

2. **Briefing Day 1**, short, open-ended questions within survey. "Please take 15 minutes to brief us on your experience during the cyber task on DAY ONE (today). Please share any in-formation you think is relevant or important for a briefing. Specific questions to consider include: major vulnerabilities found, flaws in the network, success in exfiltrating assets, strategies you used, aspects of the network that were particularly frustrating and/or confusing, and nature of deception on network, if found."

3. **Overall Briefing**, short, open-ended questions within survey. "Please take 15 minutes to brief us on your overall experience during the cyber tasks across BOTH DAYS (today and yesterday). Please share any information you think is relevant or important for a briefing. Specific questions to consider include: information not included in either daily briefing, changes in strategy or approach between the days, differences noted between the days, suspicions about the networks, etc."

4.7 Data Labeling

The rules used for labeling data are detailed below. For each of the selected bias we provide a generic definition developed before data coding began and then a domain-specific example developed by a team of psychologists and cyber experts during coding to ensure consistency. The results provided in Section 4.8 are based

off a single expert's coding of biases in the data. Future work will include using additional experts for data coding.

Confirmation Bias (CB). Rule for inclusion is any of the following ideas related to:

- The tendency to search for, interpret, focus on and remember information in a way that confirms one's preconceptions
- An unwitting/unconscious, less explicit, one-sided case-building process for one's beliefs
- Example: In a Present condition, an incorrect value judgment or exploit on a decoy
- Example: In any condition, naively blaming failure on a non-existent source/feature

Framing Effect (FE). Rule for inclusion is any of the following ideas related to:

- Changes in behavior or decision making associated with the way a problem or situation is described ("framed"), even though the underlying environment is still the same.
- Changes in evaluations of probabilities and outcomes when the same problem is framed in different ways (e.g., positive or negative)
- Example: In an Informed condition, report asset as a fake/decoy when it is real

Anchoring Bias (AB). Rule for inclusion is any of the following ideas related to:

- The tendency to rely too heavily, and essentially "anchor" oneself, to one trait or piece of information when making decisions during the task.

- The situation where people fail to adjust far enough from their initial estimate to yield their final answer
- The tendency to be bias toward an initial estimate or belief, whether or not this initial estimate is relevant to the current decision(s) at hand
- Example: Perseveration on belief/perception from Day 1 to Day 2 (across days), e.g., Continue to look for deception throughout Day 2 while in AU
- Example: Perseveration on something after a change in the "world" has occurred, e.g., Continue to believe networks are same between Day 1 and 2

Sunk Cost (SC). Rule for inclusion is any of the following ideas related to:

- Justification for increased investment of time or effort based on ongoing prior investment *(*in spite of) new evidence (available to the participant) suggesting that the decision was probably wrong
- Continuing to work on a plan or action, despite knowing it is invalid or wrong
- Example: Perseveration on ONE machine when further actions are not useful and they know it's not a good idea

Default Effect (DE). Rule for inclusion is any of the following ideas related to:

- Evidence of a choice that was made among at least two options, one of which could be viewed as a default option, in which the choice selected was the default.
- Evidence of a choice made among options where the option selected corresponded to a default option that was deemed "easier" or requiring less effort (physical or mental)
- Example: Using tools in default mode rather than parameters to better match requirements, be more appropriate, etc. (this will be evident in other data sources but not self-reports)

Availability Heuristic (AH). Rule for inclusion is any of the following ideas related to:

- The tendency to overestimate the likelihood of events which have greater "availability" in memory (that are easier to call to mind). Usually related to how recently the memory was formed, or how emotionally charged they may be.
- Example: No examples found in self-report data, due to what was reported

Illusory Correlations (IC). Rule for inclusion is any of the following ideas related to:

- Evidence of perceiving a relationship (correlation) between two events, where no such relationship exists (or where there is no evidence of such relationship)
- Example: Seeing relationship where no such relationship is possible/exists
- Example: Blaming failures on non-existent reasons when *aware* that this relationship is *not true*

We also coded for tunneling behavior, which was often seen in association with the anchoring bias. An example of tunneling is: perseverating on *one strategy* across multiple machines or returning to same machine repeatedly. We also counted the number of incorrect identifications, namely, when a decoy was reported as real and when a real machine is reported as fake (the latter is also an example of framing effect in Informed conditions). For each of the cognitive biases as well as cognitive tunneling, we used a binary labeling scheme, labeling the bias as present (or absent) for each participant in each data source. However, for the incorrect IDs we wanted a count so we labeled each occurrence of an IP address being misidentified (but only once per IP). The analysis of this data was described in Chapter 3.

	CB	FE	A	SC	AH	DE	IC	TB	Total
Absent-Uninformed	2	0	0	0	0	0	0	0	2
Absent-Informed	3	2	0	0	0	0	0	0	5
Present-Uninformed	18	0	0	1	0	0	0	2	21
Present-Informed	26	6	3	0	0	0	0	6	41
Total	49	8	3	1	0	0	0	8	69

Table 4.1: **Mattermost Day 1: Coding for evidence of cognitive biases in real-time Mattermost chat logs for day 1.** Note the large number of confirmation biases seen in the Present conditions.

	CB	FE	A	SC	AH	DE	IC	TB	Total
Absent-Uninformed (N=35)	4	0	0	0	0	0	0	0	4
Absent-Informed (N=31)	3	6	0	0	0	0	0	3	12
Present-Uninformed (N=26)	5	0	0	0	0	0	0	0	5
Present-Informed (N=29)	7	11	0	0	0	0	0	1	19
Total	19	17	0	0	0	0	0	4	40

Table 4.2: **Red Team Briefing Day 1: Coding for evidence of cognitive biases in end of day report for day 1.** Note the large number of framing effect biases seen in Informed conditions. Most participants completed this report so there is little missing data (6 missing reports).

4.8 OHF Experimental Results

Given the rules and examples listed above, results are displayed in Tables 4.1, 4.2, and 4.3 displaying the number of cognitive biases evident in the different data sources analyzed from the Tularosa Study.

Examples of *confirmation bias* go beyond misidentification and can be seen in this example of participant in the Present-Informed condition naively blaming failure on a non-existent feature, since this is one-side evidence that their failures are not their fault. The tester did not identify the method of this obfuscation:

> "There was evidence of deception occurring on the network. Systems would be found to have a vulnerability, however, upon attempts to exploit, the vulnerability would disappear, only to reappear later."

One example of a *framing effect* observed in the Absent-Informed condition was seen in the Red Teaming Briefing. Even though the participant did not feel they

	CB	FE	A	SC	AH	DE	IC	TB	Total
Absent-Uninformed (N=12)	2	0	1	0	0	0	0	0	3
Absent-Informed (N=13)	1	1	3	0	0	0	0	1	6
Present-Uninformed (N=9)	1	0	0	0	0	0	0	0	1
Present-Informed (N=11)	0	0	0	0	0	0	0	0	0
Total	4	1	4	0	0	0	0	1	10

Table 4.3: **Overall Briefing: Coding for evidence of cognitive biases in overall briefing report written at the end of day 1.** Note the small number of biases evident in this report could be caused by the increase in missing reports (77 missing reports).

experience deception or were disrupted by it, because they were framed by being informed that deception may be present they are blaming their failures on the non-existent deception:

> " I would imagine that I missed the real deception on the network which is why I did not get very far today.

One participant in the Absent-Informed condition explicitly mentioned their *tunneling behavior* in the Red Team Briefing:

> "Overall, a very frustrating day – demonstrates that my skillset is dated, and I am getting tunnelvision when trying to exploit a system."

One example of *Anchoring* behavior leading to tunneling was identified in the Overall Briefing where a participant in the Absent-Informed condition on Day 1 continue to see deception throughout Day 2 and focused on determining what was legitimate instead of stated goal:

> "I began to question the legitimacy of all of the hosts that were shuffled. Once I began to realize that this was the case, I decided to focus on trying to identify legitimate hosts if any."

4.8.1 Discussion of Observed Cognitive Biases

The Tularosa Study was not designed to measure cognitive biases in Red Teamer behavior. It was designed to measure the effectiveness of deception against cyber

	CB	FE	A	SC	AH	DE	IC	TB	Total
Absent-Uninformed	8	0	1	0	0	0	0	0	9
Absent-Informed	7	9	3	0	0	0	0	4	23
Present-Uninformed	24	0	0	1	0	0	0	2	27
Present-Informed	33	17	3	0	0	0	0	7	60
Total	72	26	7	1	0	0	0	13	119

Table 4.4: **OHF Combined Results: Combination of counts of cognitive biases in real-time Mattermost chat logs for day 1, Red Team Briefing for day 1, and Overall Briefing.** Note the smallest number is seen in the control condition and the largest in the Present-Informed condition.

attacks. However, even though it was not the main purpose of the study, we see evidence of over 100 cognitive biases in the data we examined (See Table 4.4). This supports our hypothesis $H5$ that cognitive biases are prevalent in cyber attacker behaviors and can be induced to disrupt attacks. There were likely many more biases in play during the study that did not come across in the self-reports at all, or not strongly enough to be coded as a bias. We do see a difference in the total number of biases across the conditions (See Figure 4.1). This is largely explained by the coding rules, since, for example, evidence of the framing effect would be much easier to identify in the conditions where the Informed conditions where participants were framed by being told deception may be present. As another example, we see that many more examples of confirmation bias are identified in the present conditions. This is because most participants begin the study with a belief that all the machines on the network are real, and even when provided with the information that deception may be present, still seem to look for evidence to confirm their preconceptions that the machines are real and the information the scans provide on them are valid.

We see zero instances of Default Effect, Availability Heuristic, or Illusory Correlations. This does not indicate that these biases weren't present, rather that the kinds of information reported by subjects did not supply the evidence required for us to identify their presence. We can use Sunk Cost as an example to show why

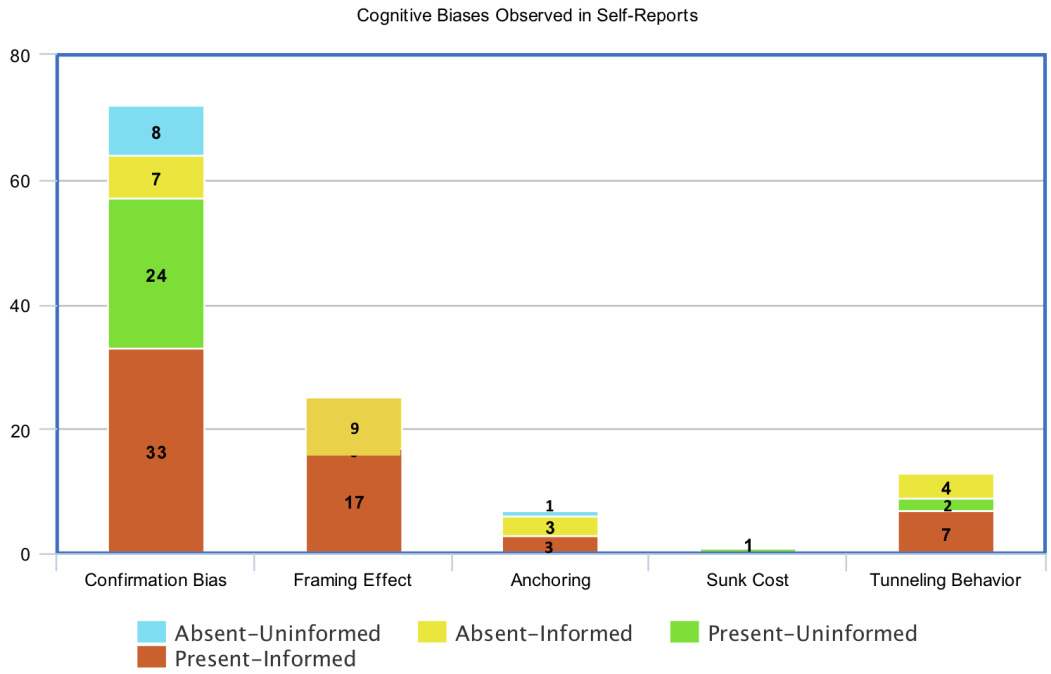


Figure 4.1: **OHF Combined Results.** Total counts of pre-selected cognitive biases observed in self-report data.

it would be unusual to have enough data reported to identify many of these biases. We only labeled one instance of Sunk Cost from a series of Mattermost data for one participant in the Present-Uninformed condition. At 10:16, the participant targets a specific machine and attempts to exploit it. They then report failure, stating that the target "appear to have crashed". They move on to fingerprint another machine. At 10:25 they report that the original target has recovered and states: "Crash further implies that target is vuln[sic] and the MS17-010 may have just failed as it does. Attempting once more." The next message says "Target has crashed again. No more attempts will be made against [target]." They move on to other targets, but at 12:19 they return to the original target and attempt to exploit it again. The outcome is the same: "[target] has crashed again... that sucks." This was labeled as Sunk Cost, because the participant perseverates on that one target, even when aware further action are not useful. Labeling this as Sunk Cost is clearly subjective, but the participant

specifically states they will not be attacking the target again (we assume because they do not think it will be useful to do so) and yet hours later they knowingly return to that same machine.

We see that different style of reporting is more likely to make different biases clear. For example, the Mattermost chat logs were intended for reporting information in real time. This allows us to collect information that the participant thought worthy of being reported in the moment, which might later by disproved, discarded or forgotten before the end of day reports. The end of day reports tend to focus more on successes and the overall briefings were designed to be information they would want passed on to team members for further action.

When looking across the 138 participants, we see that 52 (38%) participants exhibited at least one of the biases (or tunneling which can be viewed as an indicator of anchoring) in at least one of the three data sources analyzed. Interestingly only 13% of those exhibiting a bias were in the control condition. In limited self-reports, of which not all participants completed, we observed that over a third of the red teamers exhibited at least one of the small subset of biases we selected. Furthermore, 87% of those participants were in a condition with cyber or psychological deception employed, suggesting that the biases can be induced by deception and other methods.

This research clearly demonstrates that cognitive biases are effecting red teamer behavior and that methods, such as cyber deception, that induce these biases in cyber attackers, can change attacker behavior and improve defensive posture. Future work will include creating custom experiments designed to specifically examine the effect of cognitive biases in cyber and the ability for defenders to induce them to mitigate attacks.

CHAPTER 5

DISCUSSION ON HUMAN ASPECTS OF CYBER SECURITY EVALUATIONS

Cyber security is becoming a universal requirement with millions of information technology users invested in its effectiveness. As such, rigorous evaluation of cyber security tools and techniques becomes a critical and growing part of the research community. It is something that clients should require from cyber security vendors, and it is something end users should be able to assume occurs in an non biased manner. Part of this evaluation must include human-subjects experiments. Effort is needed to understand what research from cognitive, behavioral, and social science apply in cyber space and how to apply these findings to improve cyber security.

Studying red teamer behavior and cognition can inform defensive techniques to increase attacker cost on the network and lead to understanding how to gain improvements in adversary emulation performance and efficiency. These insights can improve the security and defensibility of systems and networks. By better understanding the human performance and human factors aspects of experts who are trained in adversary emulation, we also better understand the adversary, and improve our defenses against them. Our work is based on the assertion that the human component of the cyber attack deserves more investigation.

Collecting the necessary data to understand attacker cognition and behavior from existing cyber exercises, such as Capture-the-Flags (CTFs), presents many difficulties. As designed, information about individual participant's actions, perceptions, and feelings are not isolated or collected. These are not intended to be rigorous controlled

studies and are not required to complete Institutional Review Board (IRB) approval for human subjects research (HSR). While CTFs have been proposed as great sources of data for research Doupé et al. [2011], Sommestad and Hallberg [2012], they often fundamentally serve a different purpose. Large-scale HSR studies on cyber attacker behavior are required to study human behavior and cognition. This knowledge is crucial to cybersecurity research focused on resilient, reliable, and adaptive defenses. Red teamer behavior differs from unauthorized hacking, but research focused on better understanding adversary emulation can fill a critical gap in knowledge.

5.1 Lessons Learned

In this section[1], we will discuss design decisions which led to a test network and methodology for the Tularosa Study which differs substantially from a standard CTF. We will also outline lessons learned and trade-offs from multiple experimental stages including design, execution, and data collection and highlight these findings with insights drawn from participants' self-reports. Our conclusions will address the inadequacies of CTFs for studying human's behavior and provide guidance for designing future experiments Ferguson-Walter et al. [2019b].

5.1.1 Participant Motivation.

The placement and collection of flags in CTFs does little to promote or account for intrinsic motivation of cyber attackers. For research questions focused on cognition and behavior, it is important to understand attacker motivation. This could vary by sophistication of adversary, goals, resources, etc. To make an experiment more realistic, it is important to try and replicate this motivation in the participant population. CTFs (and experiments) are time bounded which can increase the pressure

[1]This section is based on submitted work: K. J. Ferguson-Walter, M. M. Major, D. C. van Bruggen, S. J. Fugate, R. S. Gutzwiller, *The World (of CTF) is not Enough*, IEEE Humans and Cyber Security Workshop, 2019 (To appear).

on a participant to find and exploit vulnerabilities to achieve the goal of collecting more flags and "winning" the competition. It is important to consider whether or not such methods of motivation align to real world scenarios and what the impact of that difference in motivation may have on the experimental results. For example, does the method of motivation impact which strategy an attacker chooses to use? How might that effect the conclusions of the study? What methods could be used to mitigate such an impact?

The overall design of the Tularosa Study was intended to capture the real-world effects of deception on the performance of human attackers. Much of the experimental design was intended to decrease the potential for confounding factors (e.g., isolating participants so that they do not interact on the problem set, which is very different from a CTF environment). While more traditional CTF environments were considered, several key issues discussed below drove our overall design decisions.

5.1.2 Experimental Validity

Based on the limitations of CTFs, some factors to carefully consider include controlling for internal validity like ensuring each subject has the same tools and the same target environment which ensures an equal chance to succeed. What we cannot control are the participants themselves, and the methods they choose to use. These uncontrolled variables make the process of evaluating human performance in a more realistic scenario very challenging, and much harder than evaluating CTF success.

Examples of factors over which we have little control include: 1) Participants have the same skill (i.e., "red teamer"), but their specialized experience and subset of skills varied drastically (from 0.25 years to 20 years of experience in network penetration); 2) Methods of compromise were typical of red teamers — the use of the same tools (e.g. nmap, hping3, Wireshark) — but the use of tools and techniques varied based on subject's background and level of familiarity; 3) Indicators of compromise and

ease of detection within the data set varied based on the participants' attacks. Even the same attack, executed with a different tool, generated different data to provide proof of the subject's success. The effect of these differences across participants on the analysis results should be minimized by random assignment of participants to each condition.

While internal validity is important, external validity is also key. Yet, when evaluating real-world cyber attacks, the limitation is in discovering all possible avenues of compromise and methods used. However, if we can discover even some indicators in this rigorous setting, this fills a gap in knowledge of how cyber attackers operate. Furthermore, if we correlate these successful attacks to the cognition, experience, characteristics, and strategies of the human behind the attack, this provides a wealth of new information useful to cyber defenders, cybersecurity researchers, and tool developers alike.

5.1.3 Human Subjects Research

Human subjects research requires IRB approval and voluntary consent of the participants. This is different from how CTFs operate. CTF data collected does not focus on human behavior or cognition, but rather on network activity and flag capture. Use of CTF data is limited to a subset of research questions because it does not include information on the humans—their expertise, experiences, thoughts, or feelings.

5.1.3.1 HSR Data.

Studies in both computer and behavioral science have investigated different methods for soliciting details on human cognition, emotion, and decision making. Having a trained expert observer being able to observe and question the participant might provide the most detailed information about their thoughts and feelings, but the trade-off is evident if there are any timing research questions being pursued, as these

questions can take participants off task. On the other end of the spectrum, open ended reporting, during and after the study can be less disruptive, but the freedom it allows leads to a wide variance of useful information reported.

While our Pilot Studies had an observer for each participants, the Tularosa Study relied on reports from the participants. Reporting in real time was needed to correlate self-reported participant cognition and emotion with the time-stamped cyber and physiological data. Additional reporting and questionnaires were provided at the end of the task. The variance in real-time reporting ranged from zero to 304 chat messages in a day with 17% providing zero for at least one day and under five percent skipping at least one of the overall end-of-task red team reports.

Semi-structured questionnaires can lead the participants to provide details that are specifically related to the research questions, but there are still other trade-offs to consider. Answering a questionnaire at the end of the day, is notoriously less accurate since human memory is faulty and biased, however stopping to answer questions in the middle of the task can effect timing metrics. To waylay timing concerns, specific breakpoints can be set to ask all participants to answer questions for the same time period, however, this forces an unnatural break in their task and can cause extra confusion and delay by taking them off task at inopportune times.

While the data collected from the Tularosa Study will no doubt be used to answer various research questions across the community, the study was designed to answer specific research questions on the effects of cyber and psychological deception on the success of a red teamer when performing a network penetration task. To properly proctor and monitor the execution of the study, sessions were limited to at most ten participants. Precautions were taken to reduce the number of participants at one time, group similar conditions together when possible, provide written and verbal instruction, and reduce the chance of participants discussing or interacting with each other during the study to reduce possible bias in the results. These reasons, in ad-

dition to the desire to increase the number of participants per condition, influenced the decision not to have participants work in teams. However, these decisions forced proctors to run sessions over a broad period of time (14 months) which can effect internal validity. Because the cyber world is fast paced, with new exploits and techniques discovered to attack old vulnerabilities, even if the target network stays static, the knowledge of the participants can change over time.

Studies need a much larger sample size when teams and competitions are involved. This is an age-old problem in team research and one that is more challenging in cybersecurity where the participant pool is restricted. The existing participants in the Tularosa Study would have only created 46 teams when put into small three person groups. Given four major conditions, this would have reduced comparisons to 11 or 12 data points, a statistically non-viable tradeoff for initial research. Teams would also make it harder to isolate and measure the data (i.e., number of successful commands), and cognitive effects (i.e., feelings of confusion and frustration).

To elicit self-reporting of progress by each participant, and mimic the more realistic team scenario, the task description also included instructions to report any findings to an external "team" via a separate, internet-connected laptop with a Mattermost chat interface. While this instruction was designed as a motivator for detailed red team reporting, some participants inferred that their task was only to perform reconnaissance and that the external team would be responsible for exploit and exfil tasks. The participants were also able to use this laptop to perform internet searches. Participant browsing activities were collected throughout the study.

5.1.3.2 Red Team Population.

The Tularosa Study is novel among current human subjects studies to date due to the inclusion of professional red-team participants. CTF challenges recruit participants from all backgrounds and demographics, from curious hobbyists and profes-

sional penetration testing experts, to hackers who specialize in unsanctioned cyber activities. While these unauthorized hackers are the population from which we hope to defend our networks, it is not feasible to assume they will volunteer for a controlled research study. A hacker's greatest asset is a treasure trove of skills and techniques that are unknown or difficult to detect by most targets, and would not want to use these tricks and tools in a fully-monitored environment.

Additionally, by collecting human subjects data, the Tularosa Study was able to exclude data from participants who self-reported information that would have further disqualified them from the initial recruitment (e.g., one subject with zero years of experience; giving us 138 useable participants.)

Despite the professional backgrounds of the participants recruited for the Tularosa Study, several exhibited their natural desire to break things and challenge the resource-constrained environment. A subject on the very first day of the experiment modified the attack laptop to connect to a wireless hotspot so they could download additional tools. Because of this, future daily briefings were modified to specifically ban WiFi.

5.1.4 Environment Design

5.1.4.1 Teams.

Conventional CTF environments tend to be team events which allow for fairly free-form creative exploration of security deficiencies and techniques for maneuver, exploitation, and exfiltration. In many cases individuals can use their own uniquely developed capabilities and tools. This introduces a problem for experimentation. If the research question relies on comparing an experiment treatment versus a control condition (e.g., adding deception or other defenses to measure effects on performance), then other factors (e.g., unique individual capabilities) have to be controlled. Otherwise, the results may be due to those unique tools, and not the cognitive or behavioral

effects of the experimental treatment. Additionally, as mentioned above, CTF participants often engage with each other for help; this introduces several sources of variance that can disrupt the goals of an experiment. Therefore, in addition to restricting the uniqueness of tools (participants could still request publicly available tools for their environment), and by isolating each participant to their own test environment, we intentionally limited the potential for interference from unique tools, external network factors or human interactions.

5.1.4.2 Network.

In general then, the network and tasking needed to be as realistic as possible without compromising internal validity. A simulated network was required to ensure that all participants were presented with the same network assets, topology, and vulnerabilities. While a real network would have provided more *external* validity, the *internal* validity would have suffered as there would be no way to ensure that each participant started with a network of the same level of difficulty and compromiseablility. Furthermore, real users on a network, while providing extra realism, can dramatically change the experiment in many unexpected ways (including providing different footholds and attack vectors to one participant versus others).

5.1.4.3 Tasking.

The test network was configured to represent an isolated enterprise business network consisting of dozens of servers and desktop computing systems running realistic services and software applications. Participant tasking was designed to encompass offensive activities of an unanticipated attacker, rather than a penetration test or red team assessment. Participants were provided with a high-level description of their task which included instructions to perform reconnaissance, system exploitation, and data exfiltration. The only difference in task description between conditions was a single sentence statement about the possible presence of deception on the network

provided to the *informed* conditions. The psychological deception condition was shaped by the information of the possible presence of deception being unaligned with the true presence of decoy machines. All primary activities by the participants were performed on a single laptop computer with a complete copy of Kali Linux which was instrumented to perform various measurements of participant interaction. This laptop was connected to the isolated test network with all external connections and wireless connections disabled. The subject was not allowed to connect other devices to the test laptop or to provide their own data, attack tools, or hardware. The isolated nature of the test network is one aspect of the design that represents a significant deviation from conventional CTF environments.

5.1.4.4 Metrics of Success.

A CTF challenge is heavily weighted toward scoring the simulated successes of the participants. Flags are planted on target machines, and proof that the participant successfully gained access to the flag is reported by the participant via the checksum or hash of the flag. This is a single point of success that can be easily and accurately measured. However, CTF challenges do not usually score a myriad of other success metrics that should matter to an attacker, such as the ability to remain undetected, to gain persistence, to pivot through a network, and to gain meaningful information to infiltrate other related devices at a later date. CTFs often carefully design and place flags to mimic the vulnerabilities in a real network by nesting flags in various locations and designing different difficulty levels for collection of the flags. However, because of the other aspects of attacker behavior that are omitted from this score, we feel that this significantly changes the goal and motivation of the attacker. This leads to a significant change in attacker cognition and behaviors, which influenced the decision to not include flags in the Tularosa Study. Researchers may be interested in a wide range of measures of success including: level of compromise and methods

used to gain compromise, efficient use of tools, complexity of exploits, efficiency of exploits, noisiness, persistence gained, backdoors established, etc.

Flags are commonly used in CTFs as a proxy for score or success, and if a study does not use flags, a different method for determining success must be devised. Counting the number of flags each participant collects is a quick way to measure which participants made the most progress. There are different strategies and methods for developing flags, with some tied to key terrain and the network, and others just amusing puzzles. However, especially for a study focused on deception, flags become more complicated. Should flags be associated with only real things on the network? If so, this is an easy "tell" for determining real from decoy. Should we include false flags? This could mislead a participant in an unfair way, biasing the results too heavily in favor of deception.

Even in studies where deception doesn't play a role, hunting for flags biases the behavior of the participant, which may or may not mimic their actual strategy in a more realistic attack scenario. Do they collect all the easy, low-point flags first? This may just be their strategy to win the game, and not how they would normally perform. In the Tularosa Study, many participants self-reported a desire to remain undetected, despite the single-day time constraint. Many CTF competitions also give direct feedback to participants on if the flag is valid and its point value. This cycle of feedback helps keep CTF challengers interested and engaged in the competition, but does not represent a realistic cycle of feedback in a real network penetration scenario, particularly where false documents and decoy systems are a realistic threat to the perceived successes of an attacker.

Furthermore, CTFs often have leaderboards that inform participants how well they are performing compared to other teams. Leaderboards and the direct competitive aspects with teams competing against each other could unnecessarily change a participant‘s behavior, which provides further evidence of how using CTF to capture

human behavior might provide an inaccurate picture of cognitive choices and human performance.

5.1.4.5 Data Collection.

CTF competitions collect some cyber data that can provide insight into the red-teamers' actions to recon and infiltrate the target network, however most CTFs collect basic content, such as packet captures and exfiltrated flags, only the latter of which are actively used to evaluate participant success within the competition. Conversely, the Tularosa Study collected additional types of data, including cognitive surveys, physiological data, raw process and log data, and subject self-reported cyber strategies.

The Tularosa Study permitted participants to each attack an identical copy of the same network, without interference from any other subject. As a consequence, 139 different networks were launched throughout the scenario, each utilizing one of three possible configurations (held constant across conditions). Individual host data were not collected from these environments, but the amount of data created by each subject having their own private target was still significant, as there was no log data or network activity overlap between participants.

In addition to process log data, each event-based data point in the Tularosa dataset includes a referential timestamp, through which a timeline of events for an entire subject's activities could theoretically be constructed. Timeline assessments can help answer research questions that examine correlations between the presence of decoys and time spent on various actions such as recon, and provides further insight linking cognitive state to cyber success.

5.1.4.6 Timestamp Correlation.

The collection of timestamps for each data point related to cyber activity is critical. However, not all of the subject's activities could be synchronized to the same time server and no optimal solution was readily available.

Timestamped resources include: 1) The subject's Kali laptop, on the target network with a local NTP Server (explicitly off-limits for participants to attack), 2) the internet-connected reference and reporting machine, which pulled its date/time from NIST internet time servers, 3) the Empatica E4 physiological monitoring wristband devices, which were reasonably synchronized when connected to a proctor setup laptop, not connected to the subject's environment, and 4) the proctor's cell phone, which was used to log the minute that the participants started the cyber task and took breaks and lunches. The connecting piece of the timestamp puzzle was 5) an iPhone, which used the "Timestamp Camera Basic" app to timestamp a video which visually recorded the clock on the Kali screen, the clock on the internet-connected machine, and the timestamps logged from a script running on that machine that was then used to synchronize with the physiological devices.

5.1.4.7 Data Coding.

Large and heterogenous data collections can be difficult to analyze. After experimental design and execution the immense challenge of labeling and analyzing the data begins. The first step is reducing each data source to the key features through which the data can be surveyed to answer research questions.

Raw Data, existing in the same format it was created, is often unusable. *Processed/Extracted Data* takes the raw data files and converts it to smaller, standardized formats. *Queried Data* imports the formatted data into tools such R and MySQL, and searches for patterns, statistics, and outliers. New contextual metadata are added to the dataset. Subject matter experts generate *Labeled Data* by applying professional

evaluations of the subject's cognitive state and activities. For example, cognitive psychologists review the participants' self-reported log data and label content which contains indicators of confusion, frustration, biases, and confidence in success. Finally, *Expert Analysis* leverages technical expertise to extract difficult ground-truth data to evaluate the accuracy of a subject's self-reported cyber efforts, in addition to cyber activities that the subject did not report.

5.1.5 Realism versus Repeatability

An enduring problem for security research is obtaining ground truth information. Methods for obtaining ground truth data or proxies for ground truth data vary greatly, and can be a difficult task. When designing an approach, it is important to consider the validity of information collected, both external validity (i.e., how well does the study generalize to different circumstances) and internal validity (i.e., how well designed is the study to support making causal inferences). Further, external validity can consider questions regarding ecological validity (i.e.,how well does the experiment map to real world situations?), population validity (i.e., how well do the participants align with the larger population?), and historical validity (i.e., how well do these results hold up over time?). When designing a study, multiple situations call for making a trade-off between different types of validity.

While a real operational network would have provided the most realistic environment, as discussed previously, the study was more interested in controlling across conditions and participants to ensure internal validity. Realism, while desirable, was not the top concern. Based on the initial[2] Pilot Study Ferguson-Walter et al. [2017] performed, it is believed that both cyber and psychological deception may be more effective on an operational network, where the natural messiness of a large network with

[2]Additional pilots were implemented to test changes in the design, procedure, environment, and data collection. Pilot studies are highly recommended.

real users improves the plausibility of the deceptive effects, and provides additional confusion through real complexities and anomalies.

5.1.6 Managing Red Teamers

The Tularosa experiment sampled red teamers from diverse skill sets and backgrounds, and managed several individuals with the desire to challenge the purpose of the experiment rather than to overcome challenges inherent to the network. One subject stated early on Day 2, *"I have shifted my perspective to more creative attacks not likely considered by those that set up the lab"*. The mildly disobedient behavior of the participants could not be easily controlled. Not all participants felt that the challenge was worthy of their skills. Some participants neglected to pay attention during the proctor briefing, and other participants egregiously defied lunch, break times, and end times. These are challenges inherent to managing human subjects, and definitely should not be unexpected in experiments involving cyber adversarial behavior.

Using professional penetration testers to perform adversary emulation and provide subject matter expertise will more closely match activity performed by unauthorized hackers than many other populations. While many participants in CTFs are also experts, little data (e.g., demographics and expertise questionnaires) are collected to determine if any of the data should be excluded from analysis. This allows for easier recruiting, with the score earned by flag collection used as a proxy for expertise.

While Red Teamers are often accustomed to working in teams, individual participants increase the power of the results and allow for easier isolation and measurement of behaviors. However, the lack of teams increased frustration and effectiveness of some participants and some research questions involving team dynamics cannot be addressed without them. Future work is needed to address research questions regarding teams.

5.1.7 Cognitive Considerations

While there are many design decisions necessary for any study, we will detail three main cognitive considerations critical to future study design. First, measurement of human behavior and cognition requires IRB approval while measurement of network and computer activity does not. Extra precautions must be taken to conduct HSR studies and protect participants, but they provide necessary insight and understanding about the motivations, perceptions, emotions, and decisions of the people being studied. CTF and other events focused on the network activity and successful collection of flags have less difficulty recruiting expert participants and do not need to instrument host machines for data collection, thus relieving privacy concerns of the participants.

Second, when participant goals are not specific, intrinsic motivation guides the red teamer behavior making it possible to see more varied tactics, techniques and procedures. Allowing participants to decide what they deem reportable reveals what they perceive as important. Collection of flags, while increasing speed and ease of judging success, may provide faulty extrinsic motivation and skew the resulting human behavior which we desire to study.

Lastly, verbal explanation of thoughts and decisions in real time provides detailed information vital to understanding cyber attackers. However, while the quantity and quality of useful information in questionnaires greatly varies, they can provide some insight into the thought process of the participant without requiring complete isolation (e.g., only running one participant at a time).

5.2 Concluding Remarks

Our work is unique among the cyber defense community and is one example of research that fills a critical gap in computer security research. Achieving more formal scientific underpinnings of computer security requires the use of the scientific method.

More rigorous experimentation to better understand cyber attackers and defenders is needed. CTFs do not provide us with the robust experimental design required to provide definitive answers to our most pressing cyber challenges. That said, CTF events were never intended to answer scientific questions. This does not mean they could not provide scientific insights. To do so, CTF events should focus more on measuring human effects in addition to system impacts and those measures which have more direct value to a competitive environment.

Traditional experimental paradigms such as those used in the Tularosa Study make significant trade-offs to achieve control over experimental conditions, sometimes eschewing ecological validity for the purposes of answering current hypotheses. Research on cyber defenders is a growing area on interest, but research on cyber attackers has been a slower effort. Understanding attacker cognition and behavior in cyberspace can be a critical, but often overlooked component of improving cybersecurity. These research finding can help more accurately model attacker behavior for testing of systems and techniques and possibly aid in increasing the realism of attacker emulation and improve red teamer training. They can also help understand how to create new techniques that focus on decisions made by a human attacker, rather than just blocking their movement on the network.

As cyber security becomes increasingly a mainstream concern, we see a rise in discussions on the adoption of automated and autonomous systems. Incorporation of deception into adaptive defensive systems are inevitable. Experiments like the Tularosa Study are necessary, not only, to know how to design these systems to best adapt their deceptive defenses to changing attacker behavior, but to measure the effectiveness of the systems themselves.

5.3 Summary of Findings

Our research provides new contributions in the area of cyber deception — a growing and promising area of research in the computer security community. Our work contributes to the understanding, measurement, and implementation of cyber deception to improve cyber defense. We detailed our pilot studies and Tularosa experimental design and execution, including trade-offs and lessons learned. We performed data analysis to examine the effectiveness of cyber deception, with consideration of if the attacker is aware of the deception, and discussed results indicating that a combination of the presence of deception and the information that deception is being used for defense can impede attacker forward progress, increase detectability, and boost attacker confusion and surprise. We also provided examples of evidence of cognitive biases exhibited by the red teamers during cyber attacks in the pilot studies and Tularosa data, providing corroboration to our theory of oppositional human factors for cyber defense.

Human-decision making is a critical but often overlooked component of cyber security. While the elusive hacker community will likely remain difficult to study, we believe there is vital research that can be done on a similar population—the authorized hacker, also known as a white hat hacker, red team, or purple team. Initial characterization of the red team population in the Tularosa Study determined that they tend to have a more rational, less avoidant, and less spontaneous decision-making style that the general population. They have a faster reaction time and higher need for cognition and tend to pursue difficult problems and enjoy the process of thinking. They tend to be more decisive, have a higher predilection towards trust and compliance, a higher level of efficiency and organization, and are less neurotic. It seems that some of these traits are specific to the white hat variety, since many hacker communities are well known for their lack of trust and compliance.

While much future work remains to learn how to improve the use of cyber deception for cyber defense, this research makes some initial contributions addressing the following hypotheses:

- Hypothesis $H1$: Defensive cyber, and psychological, deception tools impede attackers who seek to penetrate computer systems and infiltrate information. To address this hypothesis we compared performance on the cyber task between control and experimental conditions. We found that participants in the Present conditions had statistically significantly less keystrokes as well as fewer commands containing real IP addresses, indicating fewer real machines targeted. Participants in the Present conditions correctly identified the Domain Controller as a high value target less often than those in the Absent condition; we also noted a trend of fewer reported exploit successes in the Present condition. The participants in the Present conditions also triggered statistically more snort alerts than those in the Absent condition, increasing the risk of exposing themselves to defenders. These results are all consistent with a delay in forward progress and support the hypothesis that cyber deception tools impede attackers.

- Hypothesis $H2$: Defensive deception tools are effective even if an attacker is aware of their use. To address this hypothesis we compared performance on the cyber task between conditions where participants were informed about deception to where they were not informed. We found that participants in the Present-Informed condition reported significantly more confusion and were most easily detected (based on Snort alert count). The participants in the Present-Informed condition had significantly more decoy alerts overall and triggered the first decoy alert faster than those in the Present-Uninformed condition, indicating more aggressive initial behavior. However, participants in the Present-Informed condition had statistically fewer probe alerts and intrusion

alerts which would be triggered later in the kill chain, indicating less forward progress. In general, we found that the Present-Informed condition had the most effected behavior across many measurements consistent with the idea that a combination of information about and presence of deception will provide the best defense. This is counter to common thinking that deception tactics must remain covert to be effective. Participants in the Present-Informed condition also reported less failures, likely due to attribution of the failures more on the deception than on themselves.

- Hypothesis $H3$: Defensive deception is effective even if the attacker merely believes that a tool may be in use, even when it is not. To address this hypothesis we will compared performance on the cyber task between control condition and psychological deception condition. There were no statistically significant findings in this data to support this hypothesis. There was supporting evidence in the self-reports of the Absent-Information condition i.e., blaming failures on non-existent deception. We assert that additional experiments with more real-world network, user, and system details, to better match the natural messiness cyber space are needed to address this hypothesis.

- Hypothesis $H4$: Defensive cyber, and psychological, deception causes increased frustration, confusion, and self-doubt in the attacker, which impacts performance in cyber penetration tasks. To address this hypothesis we compared the levels of cognitive effects reported between control group and experimental conditions and then compare level of cognitive effects across all conditions. The participants in the Informed conditions reported significantly more surprise than those in the Uninformed conditions. Similar to our $H2$ findings, this further supports our theory that informing attackers of deceptive techniques when they are in use can benefit defenders. We also noted that frustration significantly

increased on the second day for participant moving from an Absent condition to a Present condition. The participants in the Present-Informed condition also reported a significantly higher suspicion of deception. These results, considered with the support for $H1$ described above, are consistent with the idea that exacerbating feelings of confusion and surprise impact cyber performance.

- Hypothesis $H5$: Cognitive biases are prevalent in cyber attacker behaviors and can be intensified to disrupt cyber attacks. To address this hypothesis we cataloged the types of cognitive biases observed in the cyber deception experiments, providing corroboration to our theory of oppositional human factors aiding cyber defense. Similar to our $H2$ findings, evidence of confirmation bias and framing effects is most prevalent in the Present-Informed condition.

Additionally, our empirical assessment of cyber deception demonstrated the technical utility of decoy systems in the following ways:

- For conditions where decoys were present, every participant triggered a decoy alert prior to any successful exploitation of real machines.
- For conditions where decoys were present, 35% of the packets sent targeted decoy IPs.
- For conditions where decoys were present, more IDS alerts are on decoys than real machines.
- For conditions where decoys were present, the number of Snort alerts on real machines were reduced by about half, when compared to the Absent conditions.
- The participants in the Present conditions were not easily able to identify the decoys and misidentified a total of 254 assets.

In summary, our research design and data analysis provides empirical evidence that not only is cyber deception an effective technique for impeding cyber attacks, but it may actually be more effective if the attacker is aware of the presence of deception.

5.4 Future Work

The initial Tularosa data analysis results are consistent with the theory that suspicion by an attacker that deceptive defenses are in place can increase its effect on cyber attack behavior and improve defensive posture. However, future work is still needed. Security best-practices and security hygiene behavior will always be a critical, but not sufficient, component of cyber security. The amount of detailed information provided, the method and the timing with which that information about the deceptive defenses is given is bound to be important, and requires further examination; we are planning future experiments to examine these questions. What is the *best* amount of information to provide? It is likely that providing too many details such as which commercial decoy system is deployed, on which subnets, and what configuration each decoy is using will make the systems ineffective. Even without providing this detailed information, it is likely that some Advanced Persistent Threats (APTs) will still be able to devise a method for differentiating and avoiding decoys on networks of interests. Cyber security is an arms race, and cyber deception does not change that. However, there is extra attacker time and resources that these techniques force to be spent, and even if one APT finds a work around, these defenses can still help protect the network from other attackers. However, network defenders and owners must always remain vigilant.

Future work will take us in many directions. Further analysis of the Tularosa data will examine Day 2 data to assess the persistence of effects of cyber and psychological deception over time, as well as the physiological and cognitive data matched

with more complex measures of success indicated by the cyber data. We intend to design and execute new experiments focused on measuring and intensifying cognitive biases—carefully selecting biases relevant to cyber operations. As we continue to focus on artificial intelligence for adaptive decoy systems, we intend to use these and future data analysis results to inform our utility scores, reward functions, and models Bilinski et al. [2019], Ferguson-Walter et al. [2019a], Fugate and Ferguson-Walter [2019]. We will continue to work with experts in cyber operations to improve our understanding of attacker and defender decision-making and work to improve reasoning and decision-making models to better account for realistic human-behavior. Finally, we are planning to work with several large CTF-style events to determine how to better to leverage these events to better collect useful data to help fuel the research community.

APPENDIX A

INDIVIDUAL MEASURES

A.1 Red Team Briefing

Following the network penetration task each day, participants were asked to spend 15 minutes responding to an open-ended question about their experience. The following language was used to prompt participants, with the day updated to "ONE" or "TWO" and the underlined portion only displayed to participants who were in an informed condition that day:

Please take 15 minutes to brief us on your experience during the cyber task on DAY ONE (today). Please share any information you think is relevant or important for a briefing. Specific questions to consider include: major vulnerabilities found, flaws in the network, success in exfiltrating assets, strategies you used, aspects of the network that were particularly frustrating and/or confusing, and nature of deception on network, if found.

A.2 Overall Briefing

Following the Day 2 Red Team Briefing, participants were to respond the following open-ended question about their experience:

Please take 15 minutes to brief us on your overall experience during the cyber tasks across BOTH DAYS (today and yesterday). Please share any information you think

is relevant or important for a briefing. Specific questions to consider include: information not included in either daily briefing, changes in strategy or approach between the days, differences noted between the days, suspicions about the networks, etc.

They were then asked to answer the following questions:

How much do you rely on each source of information/reference material during a typical engagement on a scale from 1 to 5? (With "1" indicating not at all and "5" indicating frequently).

- Public Internet (website/forums)
- Corporate forums (e.g., internal wiki)
- Professional network (friends/colleagues)
- Private forums (e.g., restricted IRC channel)
- Personal resources (e.g., code repositories, notes)
- Books/printed materials

How would you rate the tools available to you a scale from 1 to 5? (With "1" indicating none of the tools you needed were available and "5" indicating you had every tool you needed).

Were there any tools you would normally rely on that we didn?t give you? If so, which ones?

Before coming to participate in this exercise, did you do any research on the project beyond the information provided in the recruitment message? If so, please describe.

Did you discuss the cyber task with other red teamers (e.g., at lunch or between Day 1 and Day 2)? If so, what did you talk about?

A.3 Cyber Task Questionnaire

On each day, participants were asked about the psychological and cognitive effects of their experience during the network penetration task. The following language was used to prompt participants, with the day updated to "ONE" or "TWO" and the underlined portion only displayed to participants on Day 2:

While working on the cyber task on DAY ONE:

1. On a scale from 1-5, how much **confusion** did you experience throughout the task? (With "1" indicating you were never confused and "5" indicating you were always confused). What caused your confusion?

2. On a scale from 1-5, how much **self-doubt** did you experience throughout the task? (With "1" indicating you never doubted yourself and "5" indicating you were always doubting yourself). What caused your self-doubt?

3. On a scale from 1-5, how **confident** did you feel throughout your attack? (With "1" indicating not confident at all and "5" indicating very confident).

4. On a scale from 1-5, how **surprised** were you during the task by unexpected aspects of the network? (With "1" indicating not at all surprised and "5" indicating very surprised). What surprised you?

5. On a scale from 1-5, how **frustrated** were you during the task by unexpected aspects of the network? (With "1" indicating not at all frustrated and "5" indicating very frustrated). What frustrated you?

6. Please describe your planned, attempted, successfully executed, and/or unsuccessfully executed **strategies**.

7. Do you believe **deception** was present on the network on either Day 1 or Day 2? If so, what do you believe the deception entailed? On which day or days was it present?

A.4 Demographics Questionnaire

Participants were asked to answer the following questions:

What is your gender?

- Male
- Female
- Other

What is your age range?

- Less than 35 years
- 35-50 years
- Over 50 years

What is the highest level of education you've completed?

- High School
- Associates/Technical School
- Bachelors
- Masters

- PhD

Is English your primary language?

- English is primary language
- English is secondary language

A.5 Experience Questionnaire

Participants were asked the following questions about their red teaming experience:

For each of the following areas, please rate your level of expertise on a scale of 1 to 5 (1 = novice, 5 = expert):

- Cyber security
- Network penetration
- Host penetration
- Network reconnaissance
- Incidence response
- Generalized defense practice
- Network protocol reverse engineering
- Binary reverse engineering

How involved are you in each phase of an engagement, on a scale of 1 to 5 (1 = least, 5 = most)? (Phases from Lockheed Martin "Cyber Kill Chain").

- Reconnaissance (e.g., harvesting email addresses)

- Weaponization (coupling exploit with backdoor into deliverable payload)
- Delivery of weaponized bundle via email, web, USB, etc.
- Exploitation (execute code on victim?s system)
- Installation of malware on the asset
- Command and control channel for remote manipulation of the victim
- Actions on objectives/accomplishment of goals

How well do each of these objectives describe a typical engagement you are involved with, on a scale of 1 to 5 (1 = least, 5 = most)?

- Compliance testing (e.g., HPPA)
- Blue team training
- Demonstrate needs for increased security investments
- Whiteboarding / gaming / tabletop exercises
- Post-attack remediation effort
- Vulnerability analysis (e.g., source code / reverse engineering)
- Security architecture review
- Persistent adversary (APT) emulation

Please indicate how many years of experience you have in each of the following areas:

- Cyber security
- Network penetration

- Host penetration
- Network reconnaissance
- Incidence response
- Generalized defense practice
- Network protocol reverse engineering
- Binary reverse engineering

Which operating system do you use the most (Linux, Windows, or Other)? If "Other" please specify.

What is the context in which you generally work? Please answer each of the following:

- Size of the team you normally work in (Individually, 2-3 people, or 4 or more people)
- What is the total duration of a typical engagement (1-2 days, 3 days-1 week, 1-2 weeks, 2 weeks to one month, or over one month)?
- Types of expertise on the team (place an X next to each category, as applies to the core team):
  - Network penetration
  - Host penetration
  - Network reconnaissance
  - Incidence response
  - Generalized defense practice

  - Network protocol reverse engineering
  - Binary reverse engineering
  - Other (Please Specify)

- Expertise of other people you have easy access to, if needed (place an X next to each that applies):
  - Network penetration
  - Host penetration
  - Network reconnaissance
  - Incidence response
  - Generalized defense practice
  - Network protocol reverse engineering
  - Binary reverse engineering
  - Other (Please Specify)

A.6 Deception Questionnaire

Participants were asked the following open-ended questions:

- What makes you suspicious?
- When you experience something as suspicious, what do you interpret it as?
- When attacking a system, would you be likely to you think that the system has deception mechanisms in place?
- When attacking a system, do you first look for signs for deception?
- How do you respond when you suspect deception is in the system?

- How do you respond when you confirm the system is utilizing deception?
- If you attacked a system where deception was used, how likely are you to think deception will be present the next time you attack it?
- If you attacked a system where deception was used, how likely is it that you will attack the system again?
- If you attacked a system where deception was used, do you think that a Blue Team is also operating as part of the defense?
- If the system explicitly warned you that deception is present, how likely are you to believe the message?
- If we wanted to convince attackers that deception is present, what should we do?
- If we wanted to convince attackers that no deception is present, what should we do?

APPENDIX B

TASK BRIEFING

See below for the exact wording used in the task briefing at the start of the day. The underlined sections were only shown to participants in the informed condition.

Scenario

You represent an APT group attempting to gather information from the company Demokratika Petroleum (abbreviated as DP). You have achieved an initial foothold on the DP company network, and now must discover as much as you can about potentially valuable targets on the network. You will conduct recon on the network and locate vulnerable services, misconfigurations, and working exploits. Specifically, your task is to provide actionable intelligence about the company network which can be used by the follow-on team over the next 3-6 months. Your objective is to collect as much relevant information about the target network as you can in the allotted time without compromising future network operations.

<u>There may be deception on the network.</u>

Procedures

1. You will access the DP network using a dedicated laptop which has a Kali Linux operating system to use for reconnaissance and system exploitation (user: root password: toor). There is a Kali repository installed on the computer and you may install additional tools as needed during your activities.

2. You will also have access to a second laptop which is connected to the internet for research and technical assistance (user: recoilforce password: f0r3ns1c).

However, you may not electronically transfer information from this internet connected laptop to the attack laptop (or vice versa); you must manually enter all commands, reporting, etc.

3. When you learn potentially useful information about target systems on this network you will immediately report this information to your team via your internet connected laptop using the Mattermost website at mattermost-dev.recoilforce.net using the following format:

    - The last 2 octets of the IP address
    - Why you believe the host is interesting
    - How you obtained this information
    - Estimate its value to future operations

    You don't need to be sure about a host to file a report; you can make multiple reports on the same host. Normally you will not receive a reply to these reports, but **they are your primary deliverable**.

4. Additional notes, commands, etc (that are not sent in a Mattermost report) should be kept in the file /root/notes

5. We will be monitoring your progress, and taking into account how noisy your activities are. Prioritize obtaining as much actionable intelligence about target systems as possible without compromising future operations on the target network.

6. If you experience any technical difficulties, you can reach technical support using Mattermost at
   mattermost-dev.recoilforce.net, which is the homepage in Firefox.

7. A proctor will be present for general questions, including help contacting technical support. The proctors and tech support are not role-players in the simulation and may not be consulted for help in performing tasks on the network; they are here to facilitate your independent effort.

8. If you need to reboot either laptop for any reason, ask a proctor for assistance so that we can ensure it is collecting the data for this exercise. (For example, the attack laptop is running screen capture and keyboard capture programs).

Ground Rules:

- Limit your recon/attacks to the simulation network, 192.168.5.0/24. Within this network, do not perform attacks against the NTP server, located at 192.168.5.2 (it provides accurate time for data collection purposes and is not relevant to the task). The DP infrastructure is virtualized. You may not attack the virtual infrastructure (the hypervisor). You may not perform physical attacks on the system or social engineering attacks.

- Do not stop the recording programs running on our laptops (e.g. screen and keyboard capture). The information collected is important to the exercise we have hired you to support and will not be linked to your identity. Please help us protect your privacy by NOT entering any personally-identifying information (such as using your name in your notes or Mattermost reports, or logging into Facebook) on either laptop.

- You may not make copies of information (including software) from any of our computer systems to any storage device or computer system except the ones we have provided. Do not enable the WiFi on the attack client computer or connect it to any network other than the simulation network provided.

- Do not disclose your observations about the network simulation, its vulnerabilities, or defenses encountered. This includes not discussing your observations with other participants present at this event or with individuals that might be participating in future sessions; each individual's performance must be independent. This is important to the scientific validity of our results.

- You are expected to utilize your cyber-security subject matter expertise and perform to the best of your ability, however you are not required to utilize knowledge or techniques deemed proprietary by your employer.

APPENDIX C

SCHEDULE

Day 1

8:30 A.M. to 9:00 A.M. (Introduction and Set-Up): Participants were introduced to the study and assigned a work station. Participants who opted in to the HSR portion also had the Empatica E4 set up and filled out the Experience Questionnaire. All participants worked through an electronic task briefing to orient themselves with the red teaming scenario (see Appendix B). Those in the informed condition were also verbally informed that deception may be present on the network.

9:00 A.M. to 11:30 A.M. (Cyber Task, Part 1): Participants started on the network penetration task. Proctors noted the timing of breaks and any extreme behaviors (e.g., slamming mouse down in frustration) in the HSR subjects.

11:30 A.M. to 12:00 P.M. (Lunch Break): Participants were given a lunch break and reminded not to discuss the details of the cyber task, as per the nondisclosure agreement.

12:00 P.M. to 4:00 P.M. (Cyber Task, Part 2): Participants continued the network penetration task. Proctors continued to note the timing of breaks and any extreme behaviors in HSR subjects.

4:00 P.M. to 4:15 P.M. (Briefing): All activity on the attack laptops was halted and participants filled out the Day 1 Red Team Briefing (see Appendix A)

4:15 P.M. to 5:15 P.M. (Task Battery or Report Writing): Participants who opted out of the HSR portion continued to write a report on the cyber task (continuing the red team briefing). Participants who opted into the HSR portion complete the following tasks in order: Shipley-2 (hard copy), Day 1 Cyber Task Questionnaire (hard copy), Demographics Questionnaire (computer), Big Five Inventory (computer), General Decision-Making Style Inventory (computer), Indecisiveness Scale (computer), Sandia Matrices (computer), Over-Claiming Questionnaire (computer), and Sleep Quality Questionnaire (computer).

5:15 P.M. to 5:30 P.M. (Wrap-Up): Participants were reminded what to expect the next day and not to discuss the task with others. Proctors collected the Empatica E4 devices from participants who participated in the HSR portion of the study.

Day 2

8:30 A.M. to 9:00 A.M. (Introduction and Set-Up): Participants were reminded of the rules of engagement and told they would be working on a separate network on Day 2 (compared to Day 1). Participants who opted into the HSR portion of the study also had the Empatica E4 devices set up. All participants were given a hard copy of the task briefing document (see Appendix B); those in the informed condition were verbally told that deception may be present on the network.

9:00 A.M. to 11:30 A.M. (Cyber Task, Part 1): Participants started on the network penetration task. Proctors noted the timing of breaks and any extreme behaviors (e.g., slamming mouse down in frustration) in the HSR subjects.

11:30 A.M. to 12:00 P.M. (Lunch Break): Participants were given a lunch break and reminded not to discuss the details of the cyber task, as per the nondisclosure agreement.

12:00 P.M. to 4:00 P.M. (Cyber Task, Part 2): Participants continued the network penetration task. Proctors continued to note the timing of breaks and any extreme behaviors in HSR subjects.

4:00 P.M. to 4:30 P.M. (Briefing): All activity on the attack laptops was halted and participants filled out the Day 2 Red Team Briefing followed by the Overall Briefing (see Appendix A).

4:30 P.M. to 5:15 P.M. (Task Battery or Report Writing): Participants who opted out of the HSR portion continued to write a report on the cyber task (continuing the red team briefing). Participants who opted into the HSR portion complete the following tasks in order: Day 2 Cyber Task Questionnaire (hard copy), Deception Questionnaire (hard copy), Operation Span (computer), Need for Cognition (computer), Remote Associates Task (computer), Sandia Matrices (computer), Insight/Analytical Problem Solving (computer), and Sleep Quality Questionnaire (computer).

5:15 P.M. to 5:30 P.M. (Wrap-Up): Participants were debriefed and reminded not to discuss the task with others. Proctors handed out gift cards and collected the Empatica E4 devices from participants who participated in the HSR portion of the study.

BIBLIOGRAPHY

Acalvio. Shadowplex, 2019. URL `https://www.acalvio.com/product/`. Accessed = 2019-10-11.

Palvi Aggarwal, Varun Dutt, and Cleotilde Gonzalez. Cyber-Security: Role of Deception in Cyber-Attack Detection. In *Advances in Human Factors in Cybersecurity*, pages 85–96. Springer, July 2016. ISBN 978-3-319-41932-9. doi: 10.1007/978-3-319-41932-9_8.

Palvi Aggarwal, Aksh Gautam, Vaibhav Agarwal, Cleotilde Gonzalez, and Varun Dutt. Hackit: A human-in-the-loop simulation tool for realistic cyber deception experiments. In Tareq Ahram and Waldemar Karwowski, editors, *Advances in Human Factors in Cybersecurity*, pages 109–121, Cham, 2020. Springer International Publishing. ISBN 978-3-030-20488-4.

Torbjörn Åkerstedt, Ken Hume, David Minors, and Jim Waterhouse. The Subjective Meaning of Good Sleep, An Intraindividual Approach Using the Karolinska Sleep Diary. *Perceptual and Motor Skills*, 79(1):287–296, August 1994. ISSN 0031-5125. doi: 10.2466/pms.1994.79.1.287.

Ehab Al-Shaer, Jinpeng Wei, Kevin W. Hamlen, and Cliff Wang. *Autonomous Cyber Deception: Reasoning, Adaptive Planning, and Evaluation of HoneyThings*. Springer, Cham, 2019. ISBN 978-3-030-02109-2.

Mohammed H. Almeshekah and Eugene H. Spafford. Planning and Integrating Deception into Computer Security Defenses. In *Proceedings of the 2014 New Security Paradigms Workshop (NSPW)*, NSPW '14, pages 127–138, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-3062-6. doi: 10.1145/2683467.2683482.

Frederico Araujo, Kevin W. Hamlen, Sebastian Biedermann, and Stefan Katzenbeisser. From patches to honey-patches: Lightweight attacker misdirection, deception, and disinformation. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, CCS '14, pages 942–953, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2957-6. doi: 10.1145/2660267.2660329.

H. Arkes and C. Blumer. The psychology of sunk cost. *Organizational Behavior and Human Decision Processes,*, 35:124–140, 1985.

Catherine M. Arrington and Gordon D. Logan. Voluntary task switching: chasing the elusive homunculus. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, 31(4):683–702, July 2005. ISSN 0278-7393. doi: 10.1037/0278-7393.31.4.683.

Gbadebo Ayoade, Frederico Araujo, Khaled Al-Naami, Ahmad M. Mustafa, Yang Gao, Kevin W. Hamlen, and Latifur Khan. Automating cyberdeception evaluation with deep learning. In *Proceedings of the 53rd Hawaii International Conference on System Sciences (HICSS)*, Grand Wailea, Maui, January 2020.

A. Basak, J. Cerny, M. Gutierrez, S. Curtis, C. Kamhoua, D. Jones, B. Bosansky, and C. Kiekintveld. An initial study of targeted personality models in the flipit game. *Conference on Decision and Game Theory for Security*, October 2018.

Mark Bilinski, Kimberly Ferguson-Walter, Sunny Fugate, Ryan Gabrys, Justin Mauger, and Brian Souza. You only lie twice: A multi-round cyber deception game of questionable veracity. *Conference on Decision and Game Theory for Security (GameSec)*, October 2019.

Edward M. Bowden and Mark Jung-Beeman. Normative data for 144 compound remote associate problems. *Behavior Research Methods, Instruments, & Computers*, 35(4):634–639, November 2003. ISSN 0743-3808, 1532-5970. doi: 10.3758/BF03195543.

Brian M. Bowen, Shlomo Hershkop, Angelos D. Keromytis, and Salvatore J. Stolfo. Baiting Inside Attackers Using Decoy Documents. In Yan Chen, Tassos D. Dimitriou, and Jianying Zhou, editors, *Security and Privacy in Communication Networks*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pages 51–70. Springer Berlin Heidelberg, 2009. ISBN 978-3-642-05284-2.

Matthew L. Bringer, Christopher Chelmecki, and Hiroshi Fujinoki. A Survey: Recent Advances and Future Trends in Honeypot Research. *International Journal of Computer Network and Information Security*, 4, September 2012. doi: 10.5815/ijcnis.2012.10.07.

Albert Brzeczko, A. Selcuk Uluagac, Raheem Beyah, and John Copeland. Active deception model for securing cloud infrastructure. In *Computer Communications Workshops (INFOCOM WKSHPS), 2014 IEEE Conference on*, pages 535–540. IEEE, 2014.

Anna Buczak and Erhan Guven. A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection - IEEE Journals & Magazine. *IEEE Communications Surveys and Tutorials*, 18(2):1153–1176, 2016.

John T. Cacioppo, Richard E. Petty, and Chuan Feng Kao. The Efficient Assessment of Need for Cognition. *Journal of Personality Assessment*, 48(3):306–307, June 1984. ISSN 0022-3891. doi: 10.1207/s15327752jpa4803_13.

R. M. Campbell, K. Padayachee, and T. Masombuka. A survey of honeypot research: Trends and opportunities. In *2015 10th International Conference for Internet Technology and Secured Transactions (ICITST)*, pages 208–212, Dec 2015a. doi: 10.1109/ICITST.2015.7412090.

Susan G. Campbell, Polly O'Rourke, and Michael F. Bunting. Identifying Dimensions of Cyber Aptitude: The Design of the Cyber Aptitude and Talent Assessment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1): 721–725, September 2015b. ISSN 1541-9312. doi: 10.1177/1541931215591170.

Davide Canali and Davide Balzarotti. Behind the Scenes of Online Attacks: an Analysis of Exploitation Behaviors on the Web. In *Network and Distributed System Security Symposium (NDSS)*, page 18, 2013.

L. J. Chapman. Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior*, 6(1):151–155, 1967.

Vincent Clark. Darpa ram replay. private communication, 2014. DARPA/ARO Contract W911NF-16-C-0018.

D Climek, A Macera, and W Tirenin. Cyber Deception. *Cyber Security & Information Systems Information Analysis Center*, 4(1):14–17, 2015.

Fred Cohen and Deanna Koike. Feature: Leading Attackers Through Attack Graphs with Deceptions. *Comput. Secur.*, 22(5):402–411, July 2003. ISSN 0167-4048. doi: 10.1016/S0167-4048(03)00506-6.

Fred Cohen, Irwin Marin, Jeanne Sappington, Corbin Stewart, and Eric Thomas. Red teaming experiments with deception technologies. *IA Newsletter*, 2001.

Arthur Cropley. In Praise of Convergent Thinking. *Creativity Research Journal*, 18: 391–404, July 2006. doi: 10.1207/s15326934crj1803_13.

A. D'Amico and K. Whitley. The Real Work of Computer Network Defense Analysts. In John R. Goodall, Gregory Conti, and Kwan-Liu Ma, editors, *VizSEC 2007: Proceedings of the Workshop on Visualization for Computer Security*, Mathematics and Visualization, pages 19–37. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-78243-8. doi: 10.1007/978-3-540-78243-8_2.

Anita D'Amico, Kirsten Whitley, Daniel Tesone, Brianne O'Brien, and Emilie Roth. Achieving Cyber Defense Situational Awareness: A Cognitive Task Analysis of Information Assurance Analysts. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(3):229–233, September 2005. ISSN 1541-9312. doi: 10.1177/154193120504900304.

Dorothy E Denning. Framework and principles for active cyber defense. *Computers and Security*, 40:108–113, 2014.

A. Doupé, Be. Egele, M.and Caillat, G. Stringhini, G. Yakin, A. Zand, L. Cavedon, and G. Vigna. Hit'em where it hurts: a live security exercise on cyber situational awareness. In *Proceedings of the 27th Annual Computer Security Applications Conference*, pages 51–61. ACM, 2011.

Josiah Dykstra and Celeste Lyn Paul. Cyber Operations Stress Survey (COSS): Studying fatigue, frustration, and cognitive workload in cybersecurity operations. *USENIX Workshop on Cyber Security Experimentation and Test (CSET)*, page 8, 2018.

Serge Egelman and Eyal Peer. Predicting privacy and security attitudes. *ACM SIGCAS Computers and Society*, 45(1):22–28, 2015.

Joyce Ehrlinger, Thomas Gilovich, and Lee Ross. Peering into the bias blind spot: people's assessments of bias in themselves and others. *Personality & Social Psychology Bulletin*, 31(5):680–692, May 2005. ISSN 0146-1672. doi: 10.1177/0146167204271570.

Hillel J. Einhorn and Robin M. Hogarth. Confidence in judgment: Persistence of the illusion of validity. *Psychological Review*, pages 395–416, 1978.

Kimberly Ferguson-Walter, Sunny Fugate, Justin Mauger, and Maxine Major. Game theory for adaptive defensive cyber security. *ACM Hot Topics in the Science of Security Symposium (HotSoS)*, March 2019a.

Kimberly J. Ferguson-Walter, Dana S. LaFon, and Temmie B. Shade. Friend or Faux: Deception for Cyber Defense. *Journal of Information Warfare*, 16(2):28–42, 2017.

Kimberly J. Ferguson-Walter, Maxine M. Major, Dirk C. Vgartner an Bruggen, Sunny J. Fugate, and Robert S. Gutzwiller. The world of CTF is not enough data: Lessons learning from a cyber deception experiment. *In Proceedings of First IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA) - Workshop on Human Aspects of Cyber Security (HACS)*, December 2019b.

Kimberly J. Ferguson-Walter, Temmie B. Shade, Andrew V. Rogers, Elizabeth M. Niedbala, Michael C. Trumbo, Kevin Nauer, Kristin M. Divis, Aaron P. Jones, Angela Combs, and Robert G. Abbott. The Tularosa Study: An Experimental Design and Implementation to Quantify the Effectiveness of Cyber Deception. In *Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii, January 2019c.

Fidelis Cybersecurity. Fidelis Deception®, 2019. URL `https://www.fidelissecurity.com/products/deception/`. Accessed = 2019-10-11.

Erwin E. Frederick, Neil C. Rowe, and Albert B. G. Wong. Testing Deception Tactics in Response to Cyberattacks. In *Proceedings of the National Symposium on Moving Target Research*. Monterey, California. Naval Postgraduate School, June 2012.

James Friedrich. On Seeing Oneself as Less Self-Serving than Others: The Ultimate Self-Serving Bias? *Teaching of Psychology*, 23(2):107–109, April 1996. ISSN 0098-6283. doi: 10.1207/s15328023top2302_9.

Sunny Fugate and Kimberly Ferguson-Walter. Artificial intelligence and game theory models for defending critical networks with cyber deception. *AI Magazine*, 40(1): 49–62, Mar 2019. doi: 10.1609/aimag.v40i1.2849.

Galois. CyberChaff^{TM}, 2019. URL `https://galois.com/project/cyberchaff/`. Accessed = 2019-10-11.

Gartner Report. Emerging Technology Analysis: Deception Techniques and Technologies Create Security Technology Business Opportunities. Technical report, Gartner, Inc., 2015. URL `https://www.gartner.com/doc/3096017/emerging-technology-analysis-deception-techniques`.

R. S. Gutzwiller, K. J. Ferguson-Walter, and S. J. Fugate. Are cyber attackers thinking fast and slow? Evidence for cognitive biases in red teamers reveals a method for disruption. *Proceedings of the Human Factors and Ergonomics Society (HFES) Annual Meeting*, 2019.

Robert Gutzwiller, Kimberly J. Ferguson-Walter, Sunny Fugate, and Andrew Rogers. 'Oh, Look, A butterfly!' A framework for distracting attackers to improve cyber defense. In *Human Factors and Ergonomics Society (HFES)*, Philadelphia, Pennsylvania, October 2018.

Robert S. Gutzwiller, Sunny Fugate, Benjamin D. Sawyer, and P. A. Hancock. The Human Factors of Cyber Network Defense. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1):322–326, September 2015. ISSN 1541-9312. doi: 10.1177/1541931215591067.

Peter A. Hancock, Aaron A. Pepe, and Lauren L. Murphy. Hedonomics: The Power of Positive and Pleasurable Ergonomics. *Ergonomics in Design*, 13(1):8–14, January 2005. ISSN 1064-8046. doi: 10.1177/106480460501300104.

Kristin Heckman, Michael Walsh, Frank Stech, Todd A. O'Boyle, and Stephen R. DiCato. Active cyber defense with denial and deception: A cyber-wargame experiment. *Computers & Security*, 37:72–77, September 2013. doi: 10.1016/j.cose.2013.03.015.

Kristin E. Heckman, Frank J. Stech, Roshan K. Thomas, Ben Schmoker, and Alexander W. Tsow. *Cyber Denial, Deception and Counter Deception: A Framework for Supporting Active Cyber Defense*. Advances in Information Security. Springer International Publishing, 2015. ISBN 978-3-319-25131-8.

Illusive Networks. Illusive platform, 2019. URL `https://www.illusivenetworks.com/technology/platform/`. Accessed = 2019-10-11.

Oliver P. John and S. Srivastava. The Big-Five trait taxonomy: History, measurement, and theoretical perspectives. In *Handbook of Personality: Theory and Research*, volume 2, pages 102–138. Guilford Press, New York, NY, USA, l. a. pervin & o. p. john edition, 1999.

Eric J. Johnson and Daniel G. Goldstein. Do defaults save lives? *Science*, 302: 1338–1339, 2003.

Aaron P. Jones and Michael C. Trumbo. Personal Communication, November 2019. Sandia National Laboratories.

Andrea Kiesel, Marco Steinhauser, Mike Wendt, Michael Falkenstein, Kerstin Jost, Andrea M. Philipp, and Iring Koch. Control and interference in task switching–a review. *Psychological Bulletin*, 136(5):849–874, September 2010. ISSN 1939-1455. doi: 10.1037/a0019842.

Ellen J. Langer and Jane Roth. Heads i win, tails it's chance: The illusion of control as a function of the sequence of outcomes in a purely chance task. *Journal of Personality and Social Psychology*, page 191, 1975.

J. Liao and N. Moray. A Simulation Study of Human Performance Deterioration and Mental Workload. *Le Travail Humain*, 56(4):321–344, 1993. ISSN 0041-1868.

Lim, Sze Li Harry. *Assessing the effects oh honeypots on cyber-attackers*. PhD thesis, Naval Postgraduate School, 2006.

David Lubinski. Scientific and Social Significance of Assessing Individual Differences: "Sinking Shafts at a Few Critical Points". *Annual Review of Psychology*, 51(1): 405–444, 2000. doi: 10.1146/annurev.psych.51.1.405.

Samuel Mahoney, Emilie Roth, Kristin Steinke, Jonathan Pfautz, Curt Wu, and Mike Farry. A Cognitive Task Analysis for Cyber Situational Awareness. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 54(4):279–283, September 2010. ISSN 1541-9312. doi: 10.1177/154193121005400403.

Vincent F. Mancuso, Gregory J. Funke, Adam J. Strang, and Monica B. Eckold. Capturing Performance in Cyber Human Supervisory Control. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 59(1):317–321, September 2015. ISSN 1541-9312. doi: 10.1177/1541931215591066.

Laura E. Matzen, Zachary O. Benz, Kevin R. Dixon, Jamie Posey, James K. Kroger, and Ann E. Speed. Recreating Raven's: software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods*, 42(2):525–541, May 2010. ISSN 1554-3528. doi: 10.3758/BRM.42.2.525.

James Bret Michael, Neil C. Rowe, Hy S. Rothstein, Thomas C. Wingfield, Mikhail Auguston, and Doron Drusinsky. Phase II report on intelligent software decoys: intelligent software decoy tools for cyber counterintelligence and security countermeasures. *Technical Report NPS-CS-04-001*, 2004.

Chirag Modi, Dhiren Patel, Bhavesh Borisaniya, Hiren Patel, Avi Patel, and Muttukrishnan Rajarajan. A survey of intrusion detection techniques in Cloud. *Journal of Network and Computer Applications*, 36(1):42–57, January 2013. ISSN 1084-8045. doi: 10.1016/j.jnca.2012.05.003.

Stephen Monsell. Task switching. *Trends in Cognitive Sciences*, 7(3):134–140, March 2003. ISSN 1879-307X.

Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998. doi: 10.1037/1089-2680.2.2.175.

Vincent Nicomette, Mohamed Kaâniche, Eric Alata, and Matthieu Herrb. Set-up and deployment of a high-interaction honeypot: experiment and lessons learned. *Journal in Computer Virology*, 7(2):143–157, May 2011. doi: 10.1007/s11416-010-0144-2.

Nodder, Chris. *Evil by Design: Interaction Design to Lead Us into Temptation*. Wiley Publishing, 2013.

David Ormrod. The coordination of cyber and kinetic deception for operational effect: attacking the C4isr interface. In *IEEE Military Communications Conference*, pages 117–122, October 2014.

Delroy Paulhus, Peter Harms, M Nadine Bruce, and Daria C Lysy. The Over-Claiming Technique: Measuring Self-Enhancement Independent of Ability. *Journal of personality and social psychology*, 84:890–904, May 2003. doi: 10.1037/0022-3514.84.4.890.

Lawrence Pingree. Emerging Technology Analysis: Deception Techniques and Technologies Create Security Technology Business Opportunities. *Gartner, Inc*, 2015.

Emily Pronin, Daniel Y. Lin, and Lee Ross. The Bias Blind Spot: Perceptions of Bias in Self Versus Others. *Personality and Social Psychology Bulletin*, 28(3):369–381, March 2002. ISSN 0146-1672, 1552-7433. doi: 10.1177/0146167202286008.

Niels Provos. A Virtual Honeypot Framework. In *Proceedings of the 13th Conference on USENIX Security Symposium - Volume 13*, SSYM'04, pages 1–1, Berkeley, CA, USA, 2004. USENIX Association.

Eric Rassin, Peter Muris, Ingmar Franken, Maartje Smit, and Maggie Wong. Measuring General Indecisiveness. *Journal of Psychopathology and Behavioral Assessment*, 29(1):60–67, March 2007. ISSN 0882-2689, 1573-3505. doi: 10.1007/s10862-006-9023-z.

Thomas S. Redick, James M. Broadway, Matt E. Meier, Princy S. Kuriakose, Nash Unsworth, Michael J. Kane, and Randall W. Engle. Measuring Working Memory Capacity With Automated Complex Span Tasks. *European Journal of Psychological Assessment*, 28(3):164–171, January 2012. ISSN 1015-5759, 2151-2426. doi: 10.1027/1015-5759/a000123.

Malcolm James Ree and James A. Earles. Intelligence Is the Best Predictor of Job Performance. *Current Directions in Psychological Science*, 1(3):86–89, June 1992. ISSN 0963-7214. doi: 10.1111/1467-8721.ep10768746.

Neil Rowe and Julian Rrushi. *Introduction to Cyberdeception*. Springer International Publishing, 2016. ISBN 978-3-319-41185-9.

Neil C. Rowe, E. John Custy, and Binh T. Duong. Defending cyberspace with fake honeypots. *Journal of Computers*, 2(2):25–36, April 2007.

Sankardas Roy, Charles Ellis, Sajjan Shiva, Dipankar Dasgupta, Vivek Shandilya, and Qishi Wu. A Survey of Game Theory as Applied to Network Security. In *Hawaii International Conference on System Sciences (HICSS)*, pages 1–10. IEEE, 2010. ISBN 978-1-4244-5509-6. doi: 10.1109/HICSS.2010.35.

D Salvucci and N. A. Taatgen. *The Multitasking Mind*. Oxford University Press, Inc., New York, NY, USA, 2011.

Frank L. Schmidt. The Role of General Cognitive Ability and Job Performance: Why There Cannot Be a Debate. *Human Performance*, 15(1-2):187–210, April 2002. ISSN 0895-9285. doi: 10.1080/08959285.2002.9668091.

Frank L. Schmidt and John E. Hunter. The validity and utility of selection methods in personnel psychology: Practical and theoretical implications of 85 years of research findings. *Psychological Bulletin*, pages 262–274, 1998.

Susanne G. Scott and Reginald A. Bruce. Decision-Making Style: The Development and Assessment of a New Measure. *Educational and Psychological Measurement*, 55 (5):818–831, October 1995. ISSN 0013-1644. doi: 10.1177/0013164495055005017.

Temmie B. Shade, Andrew V. Rogers, Kimberly J. Ferguson-Walter, Sara Beth Elsen, Daniel. Fayette, and Kristin E. Heckman. The Moonraker Study: An Experimental Evaluation of Host-Based Deception. In *Hawaii International Conference on System Sciences (HICSS)*, Maui, Hawaii, January 2020.

Adrienne Shaw, Kate Kenski, Jennifer Stromer-Galley, Rosa Mikeal Martey, Benjamin A. Clegg, Joanna E. Lewis, James E. Folkestad, and Tomek Strzalkowski. Serious efforts at bias reduction: The effects of digital games and avatar customization on three cognitive biases. *Journal of Media Psychology: Theories, Methods, and Applications*, 30(1):16–28, 2018. ISSN 2151-2388(Electronic),1864-1105(Print). doi: 10.1027/1864-1105/a000174.

W. C. Shipley, C. P. Gruber, T. A. Martin, and A. M. Klein. *Shipley-2 Manual*, volume 30. Western Psychological Services, Los Angeles, CA, 2009.

T. Sommestad and J. Hallberg. Cyber security exercises and competitions as a platform for cyber security experiments. In *Nordic Conference on Secure IT Systems*, pages 47–60. Springer, 2012.

Sanjay Srivastava, Oliver P. John, Samuel D. Gosling, and Jeff Potter. Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology*, 84(5):1041–1053, 2003. ISSN 1939-1315, 0022-3514. doi: 10.1037/0022-3514.84.5.1041.

Timothy C. Summers, Kalle J. Lyytinen, Tony Lingham, and Eugene A. Pierce. How hackers think: A study of cybersecurity experts and their mental models. *Social Science Research Network (SSRN) Electronic Journal*, 2013.

Sun-tzu and Samuel B Griffith. *Sun Tzu: the art of war*. Oxford University Press, New York, NY, USA, 1963.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1st edition, 1998.

Thinkst Applied Research. Thinkst canary, 2019. URL `https://canary.tools`. Accessed = 2019-10-11.

TrapX Security. DeceptionGrid^{TM}, 2019. URL `https://trapx.com/product/`. Accessed = 2019-10-11.

S.T. Trassare, R. Beverly, and D Alderson. A technique for network topology deception. In *IEEE Military Communications Conference*, pages 1795–1800, 2013.

A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.

A. Tversky and D. Kahneman. The framing of decisions and the psychology of choice. *Science*, 211(4481):435–458, 1981.

Nash Unsworth, Richard P. Heitz, Josef C. Schrock, and Randall W. Engle. An automated version of the operation span task. *Behavior Research Methods*, 37(3): 498–505, August 2005. ISSN 1554-351X, 1554-3528. doi: 10.3758/BF03192720.

T. Vollmer and M. Manic. Cyber-Physical System Security With Deceptive Virtual Hosts for Industrial Control Networks. *IEEE Transactions on Industrial Informatics*, 10(2):1337–1347, May 2014. ISSN 1551-3203. doi: 10.1109/TII.2014.2304633.

Gérard Wagener, Radu State, Alexandre Dulaunoy, and Thomas Engel. Self Adaptive High Interaction Honeypots Driven by Game Theory. In *Proceedings of the 11th International Symposium on Stabilization, Safety, and Security of Distributed Systems*, SSS '09, pages 741–755, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 978-3-642-05117-3. doi: 10.1007/978-3-642-05118-0_51.

Richard F. West, Russell J. Meserve, and Keith E. Stanovich. Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology*, 103(3):506–519, 2012. ISSN 1939-1315, 0022-3514. doi: 10.1037/a0028857.

Barton Whaley. *Stratagem: Deception and Surprise in War*. Artech House, Boston, 1 edition edition, January 2007. ISBN 978-1-59693-198-5.

Christopher Wickens. Multiple resources and mental workload. *Human Factors*, 50 (3):449–455, June 2008. ISSN 0018-7208. doi: 10.1518/001872008X288394.

Mareike Wieth and Bruce D. Burns. Incentives improve performance on both incremental and insight problem solving. *Quarterly Journal of Experimental Psychology (2006)*, 59(8):1378–1394, August 2006. ISSN 1747-0218. doi: 10.1080/17470210500234026.

Gerald Willard. Understanding the Co-Evolution of Cyber Defenses and Attacks to Achieve Enhanced Cybersecurity. *Journal Of Information Warfare*, 14(2), April 2014.