Physiological Response to Cyber and Psychological Deception

Nicholas Wymbs, Maxine Major, Ryan Gabrys
Naval Information Warfare Center Pacific
nicholas.f.wymbs.civ@us.navy.mil

Kimberly Ferguson-Walter
Laboratory for Advanced Cybersecurity Research

Abstract

The complex relationship between cyber attacks and human cognition remains a critical area of investigation, as understanding the psychological and related physiological aspects of attackers can lead to significant advancements in cybersecurity. This study expands on existing data by measuring heart rate variability (HRV) and electrodermal activity (EDA) that was collected during a two-day cyber exercise involving expert participants where the experimental conditions encompassed both cyber and psychological deception. The analysis of the physiological data revealed that participants' stress responses were related to the experimental conditions involving deception (both psychological and cyber). These findings offer valuable insights into the stress levels experienced by cyber attackers and their potential impact on the success of cyber attacks. Decision analytics based off this information can be used by cyber defenders to improve cyber security tools and techniques.

Keywords: Deception, Cybersecurity, Cyber Attackers, Stress Response, Heart Rate Variability (HRV), Electrodermal Activity (EDA)

1. Introduction

Cyber deception is a rapidly expanding method within cyber defense that attempts to leverage human aspects of the cyber attacker to tip the scales in favor of the defender. While this increase in cybersecurity sophistication is potentially beneficial, in many ways we are still lacking an understanding of how deception affects the cyber attacker. This is essential, because information about how the attacker perceives and responds to defensive tactics can ultimately guide and refine future advances in cyber deception and other defensive techniques.

Common cyber deception techniques such as honeypots (Pawlick et al. (n.d.) and Stoll (1989)) and honeyfiles (Rowe (n.d.) and Saleh et al. (2021)) not only provide containment and advance warning of future attacks, but also provide a means for observing and aggregating behavioral patterns of network activity that can inform the development of future defenses. Along with these approaches there is also a push to understand and manipulate the cognitive state of the attacker (Veksler et al., 2020). Several efforts have sought to directly impact the psychological state of the cyber attacker (Cohen et al., 2001; Jafarian et al., 2016). For instance, within the Tularosa Study (Ferguson-Walter et al., 2019) attackers that encountered network-based cyber deception had reduced forward progress and were easily detected within the network even when notified that deception may be present (Ferguson-Walter et al., 2019). Further, this study Ferguson-Walter et al. (2019) along with others (Shade et al., 2020) have demonstrated that misinformation provides a powerful means of inducing feelings of frustration, confusion, and surprise in attackers, which appears to influence a biased and false perception of the network state (Ferguson-Walter et al., 2023). Indeed, it has also recently been shown that the emotional state of the adversary during network penetration can be classified using network activity data (Gabrys et al., 2023).

By instilling negative emotional stress and impairing decision-making through cyber deception, these findings suggest that a potential key to improved network defense is through the mind of the attacker (Climek et al., 2015; Ormrod, 2014). Despite this common effect, there is little direct

URI: https://hdl.handle.net/10125/106494
978-0-9981331-7-1

evidence showing how cyber deception affects the underlying physiology of the adversary. Here, we detail an approach to create a snapshot of the physiological response of offensive cyber experts as they participated in the Tularosa Study introduced above (Ferguson-Walter et al., 2019). As part of the Tularosa Study, near-continuous physiological sampling was performed using a wristband wearable that measured emotional stress state through heart rate and skin response biomarkers. It was hypothesized that changes in the physiological data would coincide with important events within the cyber attack behavior. While several papers have been published describing the results of the Tularosa Study (Ferguson-Walter, Major, Johnson, & Muhleman, 2021; Ferguson-Walter et al., 2019; Gabrys et al., 2023), few results from the second day of the experiment have been published. Moreover, this is the first paper to examine the findings of the physiological data.

Increased acute emotional stress is strongly correlated with a reaction from our central nervous system, which includes the autonomic nervous system (ANS). The ANS can be divided into two coactive branches, the excitatory sympathetic nervous system (SNS) and the inhibitory parasympathetic nervous system (PNS). When an individual is under heightened stress, the activity of the SNS becomes dominant, leading to greater physiological arousal in response.

This response can be addressed through both heart rate variability (HRV) and electrodermal activity (EDA). HRV reflects the dynamic interplay between the inhibitory PNS and the excitatory SNS, with the PNS predominantly active at a relaxed state leading to relatively high HRV. With increased emotional stress there is increased SNS activity (Balzarotti et al., 2017), which leads to an overall reduction in HRV because the timing between consecutive heartbeats becomes increasingly uniform. This decrease in HRV is further associated with negative emotional events, including states of confusion, surprise, and frustration (Kreibig, 2010). EDA is a measure of sweat gland function, such that with increased sweat gland activation there is increased EDA. Unlike HRV, which is a representation of contributions from both PNS and SNS, EDA solely represents signaling from the SNS to the eccrine sweat glands, which respond to fluctuations of emotion and mental state (Boucsein, 1989). EDA is comprised of a slowly changing signal called the skin conductance level (SCL) and a fast-changing signal that can usually occur after brief arousing events called the skin conductance response (SCR). The SCR is therefore considered an indicator of SNS responding to local arousing events and is commonly used as an index of negative emotional activation (Critchley, 2002).

In the following report, we provide an analysis of event-based physiological responding of cyber experts collected across different phases of network behavior during the two-day penetration task in the Tularosa Study (Ferguson-Walter et al., 2019). The following analysis of near-continuous HRV and EDA during the cyber task provides, for the first time, evidence consistent with an emotional stress physiological response to different cyber deception conditions experienced while performing offensive cyber activities. Further, we highlight how initial exposure to cyber deception leads to persistent activation of physiological response in cyber experts, even after cyber deception is removed from the network. Lastly, our work highlights the relationship between physiological responding during the penetration task and the cyber expert participant's own emotional self-appraisal following the penetration task.

2. Methods

2.1. Tularosa Experiment

This report details the analysis of physiological data collected during the Tularosa experiment (Ferguson-Walter et al., 2019). The objective of Tularosa was to understand the effects of real and psychological deception on cyber attackers. Over 130 professional cyber experts participated in the network penetration task. Participants were tasked with emulating an Advanced Persistent Threat (APT) by conducting reconnaissance, locating vulnerable services, and identifying misconfigurations and working exploits. The task was performed over two consecutive days for a duration of approximately 8 hours each day with regularly scheduled breaks.

Host and network traffic data were collected during the task alongside continuous acquisition of physiological data using an Empatica E4 wristband (Empatica Srl, Milan, Italy) to capture transient changes during specific behavioral events. Questionnaires were collected, including a report on aspects of their emotional state, or self-appraisal on a task-specific questionnaire (TSQ) survey. We encourage the reader to review additional details of the Tularosa experiment (Ferguson-Walter et al., 2019).

To examine the effects of cyber deception in the form of decoys, participants were given a foothold on an enterprise network, which either had only real targets (deception-absent (A)), or a mix of both real and decoy targets (deception-present (P)). To explore the effects of psychological deception, participants were either informed (I) that deception may be present on the network or were not told anything about the possibility of deception (uninformed (U)).

The cyber expert participants used Kali Linux and worked independently on their own copy of the simulated target network. The base network without decoys contained 25 Windows and 25 Linux boxes with variations in operating systems, patch levels, and services performed. Networks with decoys contained an additional 25 Windows and 25 Linux decoys. The decoys were designed to appear similar to the real targets on the network, but always returned a failure for any intrusion or exploit attempt. At the end of each day, participants completed the TSQ survey in which they performed a self-appraisal of how they felt during the task throughout the day, using a Likert scale (1 lowest – 5 highest), to score their levels of confusion, self-doubt, confidence, frustration, and surprise.

On Day 1 of the experiment, a 2x2 design was implemented to examine initial effects of deception. For each of the four condition groups, some participants performed the task on a network: (1) with decoys present and were informed (PI); (2) with decoys present but were uninformed (PU); (3) with decoys absent but were informed (AI) that decoys were present; (4) with decoys absent and uninformed (AU). Day 2 of the experiment was designed to examine if effects related to deception conditions from Day 1 would persist on Day 2 with no decoys present and uninformed of deception. Day 1 conditions with deception had no deception on Day 2, and thus the following Day 2 conditions were generated: PIAU, PUAU, and AIAU. An additional component of Day 2 was to determine if participants that did not experience Day 1 deception would respond to the new, uninformed presence of deception on Day 2 (AUPU). See Figure 1 for a schematic of the conditions and mapping from Day 1 to Day 2.

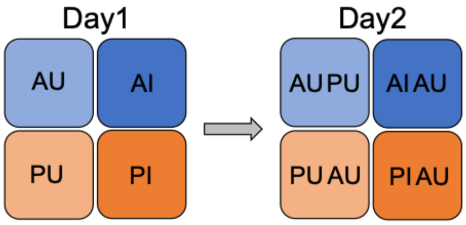

Figure 1. Tularosa Experimental Conditions. AU = deception absent, uninformed; AI = deception absent, informed; PU = deception present, uninformed; PI = deception present, informed

2.2. Tularosa Participants

The Tularosa experiment collected data from 138 professional cyber expert penetration testers, of which 126 consented to participate in the two-day human subjects research (HSR) portion of the experiment. Of these, due to technical challenges with the physiological recording device and data transfer, we began our analysis with physiological data from 95 participants on Day 1 and 100 participants on Day 2. Including only participants that also had usable cyber data, our final participant count for analysis was 93 participants for Day 1 and 95 participants for Day 2.

2.3. Network Cyber Activity

To quantify behavior during the network penetration task, we used a dataset of known network-based cyber attacks extracted from the Tularosa network traffic packet capture (PCAP) data. The cyber activity extracted from the PCAPs are labeled as one of three different types of network events: *reconnaissance* ("recon"), *exploit*, or *intrusion*. Recon behaviors are based on events that occur when an offensive cyber expert is gathering information about the network, such as searching for targets, properties of those targets, and even searching for vulnerabilities across the network or on specific hosts. Exploits are actions that exploit a vulnerability in order to gain a foothold, weaken defenses, escalate privileges, crack passwords, plant backdoors, and steal information or files that cannot be obtained by reconnaissance alone (e.g., exfiltration). Intrusions are behaviors that indicate the intention to log in, or use a foothold obtained by an exploit.

2.4. Physiological Data Collection and Preprocessing

Participants who consented to the collection of psychometric and physiologic data wore the E4 wristband on their non-dominant wrist during the two-day cyber task. HRV is derived through interbeat intervals (IBIs) that are calculated with a specific Empatica algorithm. Each IBI represents the time in milliseconds between two successive heart beats. To calculate HRV we implemented FLIRT's automatic artifact detection to remove IBIs that were outside of the 250-2000 ms range of a physiologically plausible heartbeat (Föll et al., 2021). Following this outlier detection step, the IBI data was then partitioned using a sliding window of 60 s with a step size of 30 s. To prevent inaccurate HRV window estimates, we discarded windows that contained fewer than 10% of expected IBI samples using an adaptive threshold based on the expected number of IBIs given the mean IBI for an individual window (Föll et al., 2021). Given our relatively short window size, we selected time-based features with FLIRT generating these for each consecutive 60 s window: root mean of successive differences (RMSSD), standard deviation of normal-to-normal intervals (SDNN), and proportion of normal-to-normal intervals exceeding 50 ms (pNN50).

We used a modular approach for preprocessing EDA, such that the time series was first low-pass filtered with an infinite impulse response (IIR) cutoff frequency set to 0.1 Hz, and then additional detection approaches were used to provide further noise reduction and interpolation. We identified noise artifacts using EDAexplorer (Taylor et al., 2015), which implements an SVM classifier to detect outliers with binary classification (artifact or clean) over consecutive 5 s epochs of features derived from the raw and filtered EDA, along with accelerometer and temperature E4 sensor data. Segments classified as artifacts were then removed using linear interpolation, and the resulting cleaned EDA signal were further decomposed into skin conductance response (SCR) and skin conductance level (SCL) components using Ledalab (Benedek & Kaernbach, 2010). Given the event-based approach we focus our analysis on the phasic SCR component. From here we used EDAexplorer for SCR peak detection and feature generation (Taylor et al., 2015), resulting in the following: mean SCR amplitude, mean SCR decay time (from peak), mean SCR width, and mean area under the curve (AUC) of SCR peaks.

2.5. Event-based behavior and physiology processing

Because HRV and EDA had different sampling rates, it was necessary to first align the timings of the sample windows generated during preprocessing for each subject. We performed this step by aligning the time stamps of the EDA windows to the closest HRV window tolerance set to +/- 30 s for each window, and then selected those HRV and EDA windows with aligned time stamps. For a given subject, this aligned physiology data was then cropped to the time stamp range of their corresponding PCAP behavior. On a rare occasion a participant's cyber behavior would contain very few logged results, potentially due to a technical error or the participant not performing any substantial cyber activity. We excluded participants with this issue using a set of filters that required (1) a time stamp match of at least 1% of physiological time windows, and (2) for behavioral event entropy and destination IP entropy to be greater than zero.

We then summarized cyber activity overlapping in time with each remaining physiology time window as counts for recon, intrusion, and exploit behavior events. We further specified recon events into two sub-classes: *broad* and *targeted* recon to account for different reconnaissance approaches that either monitor a wide array of machines and services (broad) or those that target a specific machine or service (targeted). Here, we define a broad recon event as any 60 s time window containing on average of two or greater attempts for any recon sub-event, and targeted recon as any 60 s time window with on average less than two attempts for any recon sub-event.

2.6. Statistics

For each cyber expert participant, we found the mean of each physiology feature and behavior event type. We next applied a transformation on these means to closely resemble a normal distribution (Box-Cox transform) and used a z-score transformation to exclude outlier means +/- 3SD from the group mean for each physiology feature. We carried out imputation for missing TSQ values using the Python library `fancyimpute` to perform multiple imputations by chained equations (MICE), a method which takes into consideration the statistical dependencies of other dependent variables present within a full dataset.

We used a mixed effects 2-way ANOVA to assess main effects and interactions of event-based physiology and conditions of network deception. We chose to implement separate tests for each day of the experiment. To measure the initial effects of deception on Day 1, the decoy deception condition had two levels (present, absent) and the psychological deception condition had two levels (informed, uninformed). We conducted a mixed linear effects (MLE) analysis using the Python `pymer4` statistics library to perform a direct comparison of Day 1 and Day 2 changes of event-based physiology. For this, the deception condition had four levels (AUPU, AIAU, PUAU, and PIAU) and the session condition had two levels (Day 1 and Day 2). For this step, we selected a MLE model because it is robust to missing repeated measures data, which was the case for us as some participants were missing entire days of event-based physiology. For ANOVA interrogation, we performed independent samples t-tests without assumption of equal variance between conditions or groups. For MLE interrogation, we used a partially overlapping samples t-test (Derrick & White, n.d.). Planned pairwise tests were handled using Bonferroni correction for multiple comparisons. Lastly, standard Pearson's tests were used to assess correlation between event-based physiology and TSQ self-assessment.

3. Results

We were interested in understanding the immediate and sustained effects of deception as measured by near-continuous sampling of HRV and EDA activity during the 2-day task and summarized these physiological measures using different behavioral elements consistent with elements of the Cyber Kill Chain (recon, intrusions, exploits).

3.1. Day 1: Early effects of deception

For HRV we found significant main effects of the psychological deception condition during both broad recon events: SDNN, F(1,77) = 5.033, p = 0.028; RMSSD, F(1,77) = 6.001, p = 0.016; pNN50, F(1,77) = 9.377, p = 0.003 (Figure 2a), and intrusion events: pNN50, F(1,82) = 5.455, p = 0.022 (Figure 2b). The effect of psychological deception was led by an overall reduction in HRV for participants that received information suggesting the potential presence of network decoys. For broad recon events, pairwise tests reflect that the absent-informed (AI) group had significantly lower HRV when compared to either absent-uninformed (AU: pNN50, T(39) = -3.156, p = 0.003; RMSSD, T(39) = -2.201, p = 0.034 (trending) or present-uninformed groups (PU: SDNN, T(35) = -2.647, p = 0.012; RMSSD, T(35) = -2.540, p = 0.016; pNN50, T(35) = -3.159, p = 0.004). Pairwise tests for intrusion events revealed a similar pattern as above, with the AI group showing relatively less HRV compared to AU (pNN50, T(41) = -2.327, p = 0.025, trending). For EDA we found a significant effect of the decoy condition during broad recon events: SCR decay time, F(1,57) = 4.403, p = 0.040 (Figure 2c). This effect was led by an overall increase in SCR decay time for the condition groups with decoys present. While both present-informed (PI) and present-uninformed (PU) show an overall increase in SCR decay time, pairwise tests indicate that PU had significantly higher decay time values compared to either AU (T(25) = 2.18, p = 0.039, trending) or AI (T(21) = 2.594, p = 0.017) groups. We did not observe any other significant main effects or interactions for HRV or EDA Day 1 event-based deception conditions.

3.2. Day 1–Day 2: effects of deception persistence

Here, we evaluated deception persistence by measuring physiological change across the two testing days and the four deception groups (AIAU, AUPU, PIAU, PUAU). For HRV, we found a significant effect of deception condition during broad recon events (SDNN, F(3,95.32) = 2.765, p = 0.046; RMSSD, F(3,94.38) = 3.089, p = 0.031; pNN50, F(3,95.02) = 3.75, p = 0.014, Figure 3a). This highlights a reduction in HRV collapsed across both days for AIAU, which was significantly lower than condition groups AUPU (SDNN, T(43) = -2.654, p = 0.014; RMSSD, T(43) = -2.779, p = 0.008; pNN50, T(43) = -3.125, p = 0.003) and PUAU (SDNN, T(42) = -2.422, p = 0.02, trending; RMSSD, T(42) = -2.252, p = 0.023, trending; pNN50, T(42) = -2.727, p = 0.009). This difference indicates a stable effect for broad recon that has persisted across both days, with the AIAU condition group showing lower HRV with respect to either condition group uninformed of deception on Day 1 (AUPU, PUAU). In addition, we observed a significant interaction for exploit events between day and deception condition group (pNN50, F(3,38.59) = 4.470, p = 0.009, Figure 3b). As depicted in Figure 3b, this interaction reflects the across-session decrease in HRV for AIAU (pNN50, T(15.32) = -2.482, p = 0.025) and increase in HRV for AUPU (pNN50, T(14) = 2.530, p = 0.024). We did not observe any event-based main effects or interactions for EDA.

3.2.1. TSQ Day 1: Early relationship between emotional state and event-based physiology of cyber experts. Cyber expert participants were asked to perform a self-appraisal of their emotional state at the end of each day in the Tularosa experiment. We tested for a relationship between early self-reflection (TSQ) at the end of the first day and cyber event-based physiology generated by participants during the cyber task. For HRV, we found significant moderate negative correlations between the cyber experts level of either confusion or surprise self-ratings and their HRV during intrusion behavioral events (confusion: RMSSD, R = -0.290 p = 0.007; pNN50, R = -0.229, p = 0.034; surprise: RMSSD, R = -0.285, p = 0.008; pNN50, R = -0.243, p = 0.024, Figure 4a), showing that those reporting higher levels of confusion or surprise had lower levels of HRV during intrusions. Similarly, we observed negative correlations between HRV and TSQ confusion during targeted recon (RMSSD, R = -0.221, p = 0.042). We also observed significant positive correlations between EDA and confusion self-rating during targeted recon (SCR AUC, R = 0.284, p = 0.024; SCR amplitude, R = 0.252, p = 0.046) and exploit events (SCR AUC, R = 0.422, p = 0.014) (Figure 4b). This result, like that for HRV and confusion described above, shows that cyber experts with higher confusion experienced greater emotional stress responding during targeted recon and exploit events on Day 1. Additionally, we found a significant moderate negative correlation with confidence during

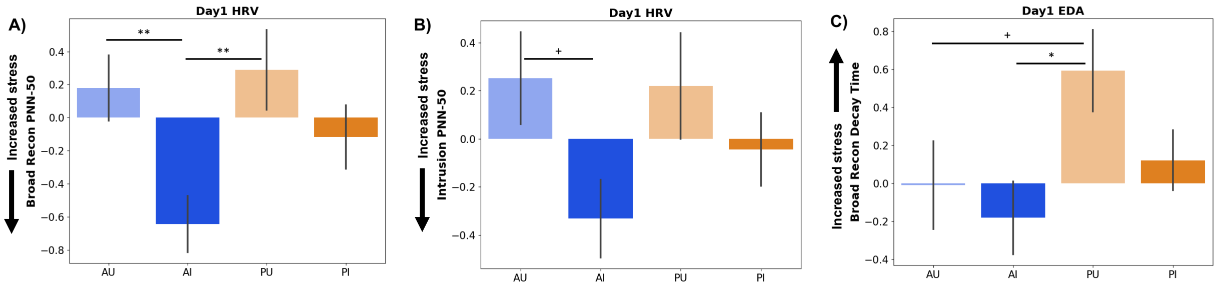

Figure 2. Early (Day1) effects of deception are consistent with increasing emotional stress (decreasing HRV, increasing EDA). We found reduced HRV during (A) broad recon and (B) intrusion events for psychological deception, as well as increased EDA for (C) broad recon decoy deception. AU = absent, uninformed; AI = absent , informed; PU = present, uninformed; PI = present, informed ; $** = p < 0.005$ **;** $* = p < 0.01$ **;** $+ = p < 0.05$

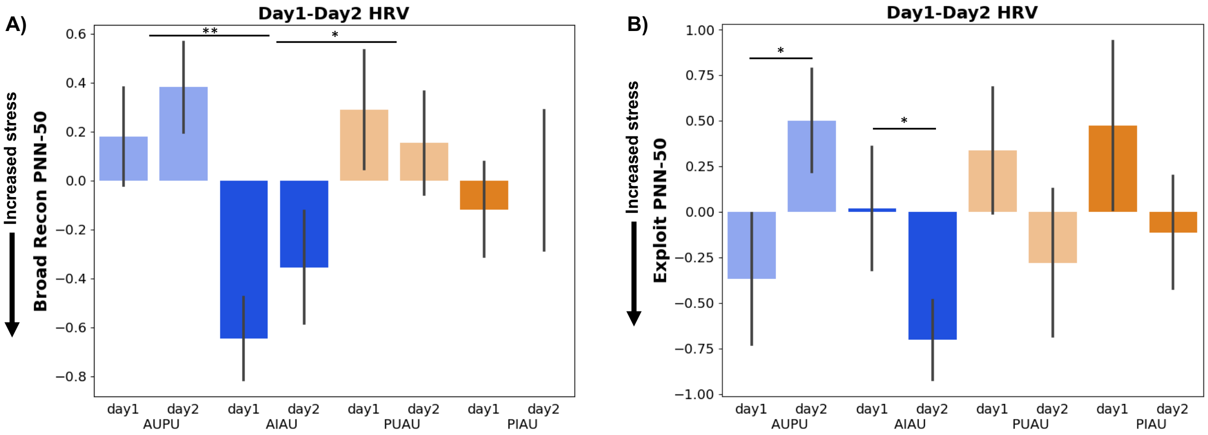

Figure 3. Effects of deception over consecutive days of penetration testing. A) There is a persistent reduction of HRV (increased stress) during broad recon for psychological deception (AIAU). B) There is a significant interaction between session and condition group during exploit events, such that exposure to deception on Day 1 leads to a reduction in HRV (increased stress) on Day 2. AU = absent, uninformed; AI = absent , informed; PU = present, uninformed; PI = present, informed ; $** = p < 0.005$ **;** $* = p < 0.01$ **;** $+ = p < 0.05$

broad recon events (SCR decay, R = -0.282, p = 0.028; SCR width, R = -0.355, p = 0.005, Figure 4c), showing that lower confidence in the task was related to greater Day 1 emotional stress responding during broad recon events.

3.2.2. TSQ Day1 - Day2: Changes in emotional state that are related to changes in event-based physiology in cyber experts. Here we describe results that capture the relationship between changes in emotional assessment of participants from Day 1 to Day 2 with corresponding changes in physiology during targeted cyber behavioral events. For HRV, we observed a significant moderate negative correlation between confusion and event-based exploit physiology (RMSSD, R = -0.420, p = 0.023; pNN50, R = -0.479, p = 0.009, Figure 5a), showing that increasing self-reported confusion was related to increasing emotional stress-based responding during exploit behavior. Further, there was a significant moderate positive correlation between confidence and exploit HRV (SDNN, R = 0.437, p = 0.018; RMSSD, R = 0.448, p = 0.015; pNN50, R = 0.482, p = 0.008, Figure 5b), indicating that increasing confidence across days is related to increasing HRV during exploit behavior across days. For EDA, we observed a significant moderate positive correlation between confidence and recon (broad recon: SCR AUC, R = 0.386, p = 0.013; SCR amplitude, R = 0.338, p = 0.031; targeted recon: SCR AUC, R = 0.416, p = 0.007; SCR amplitude, R = 0.338, p = 0.031; SCR decay time, R = 0.319, p = 0.041) and intrusion events (SCR AUC, R = 0.361, p =

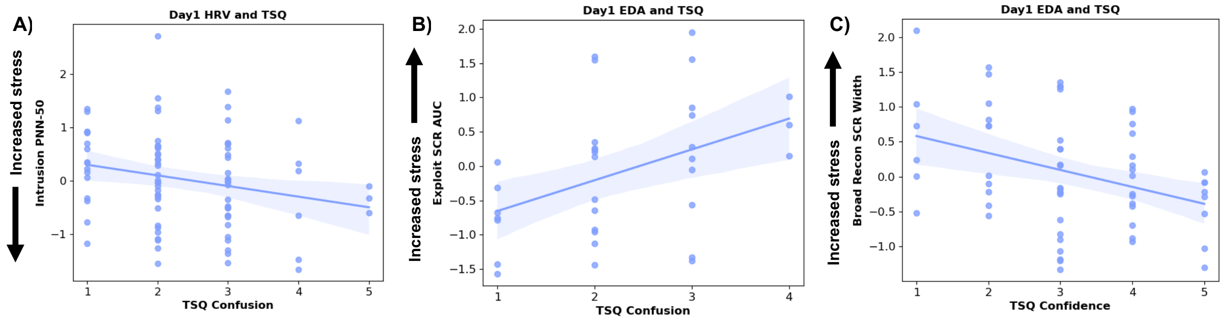

Figure 4. Relationship between early physiological responding and emotional self-appraisal following the Day 1 penetration task (TSQ). Increased confusion is related to increased stress responding (lower HRV, higher EDA) during A) intrusions and B) exploits. Similarly, in C) decreased confidence is related to increased stress responding (higher EDA) during broad recon.

0.014; SCR amplitude, R = 0.359, p = 0.017, Figure 5c). This indicates that increased confidence across days is related to increased EDA across days, which is perhaps unexpected considering that one might expect to find a reduction in stress with increasing confidence.

4. Discussion

In the reported analysis, we were interested in evaluating the relationship between cyber deception techniques and the physiological response of cyber experts during different behavioral event states common to the Cyber Kill Chain. Near-continuous physiology (HRV, EDA) was collected in Tularosa participants as they participated in a two-day network penetration test (Day1 and Day2), during which some were exposed to cyber decoys, others to psychological deception, and others to both. It was first hypothesized that cyber experts with early exposure (Day1) to deception would show a greater indication of physiological emotional stress responding during recon, intrusion, and exploit attempts. Indeed, we found physiological evidence of emotional stress responding for both deception conditions tested (decoy and psychological deception). For expert participants that were informed of the possible presence of network decoys (i.e., informed condition), we found that they exhibited lower levels of HRV during broad recon and intrusion events (Figure 2a, b), indicative of greater emotional stress response. Additionally, for participants placed on a network with cyber decoys (i.e. present condition), we found increased EDA also during broad recon events (Figure 2c). Together, these findings suggest that participants exhibited higher stress response levels when interacting with networks that contained decoys. Moreover, this indicates that the AU control condition on Day 1 exhibited less stress during broad recon relative to the other conditions that experienced deception. This further supports the claim that cyber and psychological deception both impact attacker behavior and emotional state, consistent with previous Tularosa evidence (Ferguson-Walter, Major, Johnson, & Muhleman, 2021). Furthermore, unlike previous work, the physiological data indicates a clear statistical difference between AI and AU (control) conditions, highlighting the underlying impact that psychological deception has on at offensive cyber experts' physiological response.

We were also interested in understanding if these initial effects of deception on physiology persisted over longer exposure to the network penetration task, by testing if early exposure to deception would lead to a persistent or maintained level of physiological responding on the second day of testing (Day2). Following this prediction, we found enduring effects of psychological deception during broad recon, such that participants informed of network decoys on Day 1 maintained their increased stress level (i.e., reduced level of HRV) on Day 2 (Figure 3a). Notably, this stress level was observed even though participants were not informed of any possibility of deception on Day 2 and were also briefed that they would be carrying out a penetration test on a different network than experienced on Day 1. While findings of previous Tularosa publications (Ferguson-Walter, Major, Johnson, & Muhleman, 2021) focused on the combined presence and information of deception (PI), the physiological results also begin to highlight the impact of information alone (AI). The reduced levels of HRV could in part be explained by a lack of progress through the kill chain, to perhaps, the more stressful phases of an attack. Since the present conditions (PI, PU) included decoys in addition to the other systems on the network, the number of

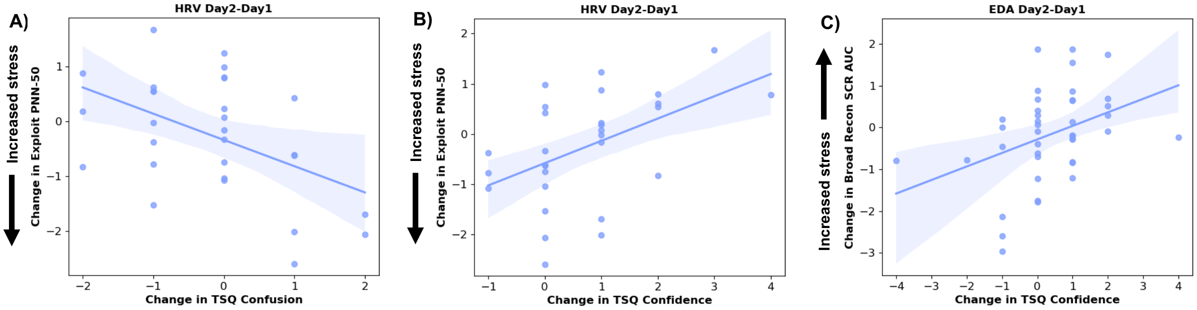

Figure 5. Relationship between changes in physiological responding (Day 2 – Day 1) and changes in emotional self-appraisal (Day 2 – Day 1 TSQ ratings). Increased confusion A) and decreased confidence B) over task days is related to increased physiological stress responding (decreased HRV) during exploits. In C), increased stress responding (increased EDA) over task days during broad recon was related to increased confidence.

potential targets was double that of the absent condition. Thus, the presence of the decoys themselves, and the extra difficulty they caused participants trying to map the network is a possible explanation for the increased stress levels.

Taking a closer look at changes between Day 1 and Day 2, we observed an interaction showing increased physiological stress on Day 2 (i.e., reduced HRV) for those groups exposed to deception on Day 1, and the opposite decreased stress on Day 2 for those not exposed to deception on Day 1 (AUPU, Figure 3b). This pattern further highlights the impact of initial deception, such that there appears to be a trend towards increased stress responding on future network penetration exercises. On the other hand, these results further reinforce the need for early deception exposure given the lack of physiological effect that technological deception has on the cyber group that was not subject to deception on Day 1 (AUPU). The early presentation of deception may shape the initial impressions that attackers have of the network defenses (i.e., difficulty or risk caused by deception), which may be an important factor on attack behavior and success that persists across a campaign, and potentially to a new network. This concept has been corroborated during simulation (Aggarwal et al., 2017; Walter et al., 2021), with authors observing that the timing or early placement of deceptive elements impacts attacker behavior and may extend further to disruption of more longstanding attacker goals.

Following the conclusion of each day, participants were given the TSQ to self-report emotional state experienced during the task. With Day 1 exposure to cyber deception, participants reporting greater confusion and surprise following Day 1 were also more likely to have increased stress (i.e. lower HRV) during events further along the Cyber Kill Chain (e.g., intrusions and exploits, Figure 4a, 4b). This is consistent with Day 1 results showing that psychological deception leads to greater stress responding during intrusion events (Figure 1b). Lastly, we anticipated that there may have been changes in emotional state over the two days of testing. Here, we found some evidence that participants who become more confused (Figure 5a) and less confident (Figure 5b) from Day 1 to Day 2 tended to be more stressed on Day 2 exploits. This complements our observation that physiological stress increased during exploit events for participants exposed to deception on Day 1 (Figure 3b).

While our results show the impact of cyber deception on physiology there are limitations that should be noted. For instance, our current analysis is limited at explaining differences at the level of deception condition or group. We were unable to formulate tests at the level of the asset (real or decoy) or whether an attack was successful or not. It is unknown how interaction with asset type or attack efficacy may differentially affect physiology. Further, it would be helpful to determine the relationship between behavioral performance and physiology. There is value in understanding how certain behavioral characteristics of attack (e.g., time to first recon or first exploit) may relate to fluctuations in physiology and deception condition. We also experienced limitations in statistical power. Due to physiology data quality and a less frequent occurrence of behavior events further along the kill chain, some comparisons have a reduced number of data points. Related to this limitation, we were unable to provide in-depth condition-specific correlations with physiology and instead collapsed across conditions to expand inferential power.

We report physiological results that stem from two distinct but complementary measures, EDA and HRV. While it is common to carry out physiological inquiry using both measures, it is worth noting that we sometimes observed patterns that were not always complementary. For instance, on Day 1, with HRV we found greater stress responding of broad recon during psychological deception conditions, but not EDA. Instead, we observed greater stress responding during decoy deception with EDA. These differences are not unexpected given that each metric is sampling a different aspect of the ANS. Whereas the temporal features for HRV measure a balance between the SNS and PNS branches of the ANS, the EDA metric largely represents the output of the SNS. Given the complexities of the Tularosa experiment, it is possible that our results may be explained by an additional third variable, such as additional load on cognitive processes (Ayres et al., 2021) or demand on emotional self-regulation (Thayer & Lane, 2007). This alternative interpretation may explain the occasional counter-intuitive result. For instance, we found a positive correlation between increasing EDA and increasing confidence during recon and intrusion events (Figure 4c, broad recon shown). An alternative interpretation may be that increasing arousal is related to greater confidence during these events. Further, our quantification of emotional self-appraisal is limited to the ratings derived from a simple Likert scale within the TSQ survey. Of note, previous research on the free response items of the Tularosa TSQ suggest that certain cyber behaviors may contribute to greater negative emotional self-appraisal, which may serve to inform future cyber defense strategies (Ferguson-Walter, Gutzwiller, et al., 2021).

5. Conclusions

A key motivation in cyber defense is to level the playing field between attacker and defender by understanding an attacker's psychological state. It is therefore important to consider objective measures related to an attacker's physiology that also correspond to psychological constructs. Here, we utilized two complementary measures of emotional stress response (HRV, EDA) to understand how the underlying psychological state of an attacker fluctuates in response to different conditions of cyber deception during a two-day penetration task. We were able to show that early exposure to deception conditions led to differences in emotional stress responding early in the kill chain, during recon, and that these initial changes were evident for attackers that faced misinformation within the network (i.e., being informed of decoys when no decoys were present). This unbalancing effect of psychological deception has yet to be described from Tularosa, highlighting the sensitivity of physiological metrics. While no immediate behavioral differences were previously observed (Ferguson-Walter, Major, Johnson, & Muhleman, 2021), the persistent effect of psychological deception may cause future consequences to cyber attack behaviors that impact future campaigns.

Related to this sustained influence, we show that early exposure to psychological deception leads to persistent physiological changes over extended attack phases. This highlights the potential impact that early deception exposure may bring to downstream activity, impacting processes that may take longer to develop within a campaign. In this regard, attackers that were exposed to deception early in their campaign happened to show greater emotional stress responding during exploit behaviors on the second and final day of the campaign, despite no longer being exposed to deception of any sort. Overall, these physiological correlates suggest that early exposure to deception shifts an attacker's overall schema and approach for a campaign.

References

Aggarwal, P., Gonzalez, C., & Dutt, V. (2017). Modeling the effects of amount and timing of deception in simulated network scenarios. *International Conference On Cyber Situational Awareness, Data Analytics And Assessment*, 1–7.

Ayres, P., Lee, J. Y., Paas, F., & van Merriënboer, J. J. G. (2021). The validity of physiological measures to identify differences in intrinsic cognitive load. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.702538

Balzarotti, S., Biassoni, F., Colombo, B., & Ciceri, M. R. (2017). Cardiac vagal control as a marker of emotion regulation in healthy adults: A review. *Biological Psychology*, *130*, 54–66.

Benedek, M., & Kaernbach, C. (2010). A continuous measure of phasic electrodermal activity. *Journal of Neuroscience Methods*, *190*, 80–91.

Boucsein, W. (1989). *Electrodermal activity*. Springer.

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

Cohen, F., Marin, I., Sappington, J., Stewart, C., & Thomas, E. (2001). Red teaming experiments with deception technologies. *IA Newsletter*.

Critchley, H. D. (2002). Electrodermal responses: What happens in the brain. *Neuroscientist*, *8*.

Derrick, B., & White, P. (n.d.). Review of the partially overlapping samples framework: Paired observations and independent observations in two samples. *The Quantitative Methods for Psychology*, *18*.

Ferguson-Walter, K. J., Gutzwiller, R. S., Scott, D. D., & Johnson, C. J. (2021). Oppositional Human Factors in Cybersecurity: A Preliminary Analysis of Affective States. *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops*, 153–158. https : / / doi . org / 10 . 1109 / ASEW52652.2021.00040

Ferguson-Walter, K. J., Major, M. M., Johnson, C. K., Johnson, C. J., Scott, D. D., Gutzwiller, R. S., & Shade, T. (2023). Cyber expert feedback: Experiences, expectations, and opinions about cyber deception. *Computers & Security*, *130*(103268), 15–23.

Ferguson-Walter, K. J., Major, M. M., Johnson, C. K., & Muhleman, D. H. (2021). Examining the efficacy of decoy-based and psychological cyber deception. *30th USENIX Security Symposium*, 1127–1144.

Ferguson-Walter, K. J., Shade, T. B., Rogers, A. V., Niedbala, E. M., Trumbo, M. C., Nauer, K., Divis, K., Jones, A. P., Combs, A., & Abbott, R. G. (2019). The Tularosa Study: An experimental design and implementation to quantify the effectiveness of cyber deception. *Hawaii International Conference on System Sciences*.

Föll, S., Maritsch, M., Spinola, F., Mishra, V., Barata, F., Kowatsch, T., Fleisch, E., & Wortmann, F. (2021). Flirt: A feature generation toolkit for wearable data. *Computer Methods and Programs in Biomedicine*, *212*(106461).

Gabrys, R., Venkatesh, A., Silva, D., Bilinski, M., Major, M., Mauger, J., Muhleman, D., & Ferguson-Walter, K. (2023). Emotional state classification and related behaviors among cyber attackers. *Proceedings of the 56th Hawaii International Conference on System Sciences*.

Jafarian, J. H., Niakanlahiji, A., Al-Shaer, E., & Duan, Q. (2016). Multi-dimensional host identity anonymization for defeating skilled attackers. *In Proceedings of the 2016 ACM Workshop on Moving Target Defense*, 47–58.

Kreibig, S. D. (2010). Autonomic nervous system activity in emotion: A review. *Biological Psychology*, *84*, 394–421.

Ormrod, D. (2014). The doordination of cyber and kinetic deception for operational effect: Attacking the c4isr interface. *IEEE Military Communications Conference*, 117–122.

Pawlick, J., Colbert, E., & Zhu, Q. (n.d.). A game-theoretic taxonomy and survey of defensive deception for dybersecurity and privacy. *ACM Computing Surveys*, *52*.

Rowe, N. C. (n.d.). A model of deception during cyber-attacks on information systems. *IEEE First Symposium on Multi-Agent Security and Survivability*, 21–30.

Saleh, A. R., Al-Nemera, G., Al-Otaibi, S., Tahir, R., & Alkhatib, M. (2021). Making honey files sweeter: Sentryfs - a service-oriented smart ransomware solution. *17th European Dependable Computing Conference*.

Shade, T., Rogers, A., Ferguson-Walter, K., Elson, S., Fayette, D., & Heckman, K. (2020). The moonraker study: An experimental evaluation of host-based deception. https://doi.org/10.24251/HICSS.2020.231

Stoll, C. (1989). *The cuckoo's egg: Tracking a spy through the maze of computer espionage*. Doubleday.

Taylor, S., Jaques, N., Chen, W., Fedor, S., Sano, A., & Picard, R. (2015). Automatic identification of artifacts in electrodermal activity data. *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 1934–1937.

Thayer, J. F., & Lane, R. D. (2007). The role of vagal function in the risk for cardiovascular disease and mortality [Special Issue of Biological Psychology on Cardiac Vagal Control, Emotion, Psychopathology, and Health.]. *Biological Psychology*, *74*(2), 224–242. https://doi.org/https://doi.org/10.1016/j.biopsycho.2005.11.013

Veksler, V. D., Buchler, N., LaFleur, C. G., Yu, M. S., Lebiere, C., & Gonzalez, C. (2020). Cognitive models in cybersecurity: Learning from expert analysts and predicting attacker behavior. *Frontiers in Psychology*, *11*(1049).

Walter, E., Ferguson-Walter, K., & Ridley, A. (2021). Incorporating deception into cyberbattlesim for autonomous defense. *IJCAI-21: 1st International Workshop on Adaptive Cyber Defense*.