Emotional State Classification and Related Behaviors Among Cyber Attackers

Ryan Gabrys, Anu Venkatesh, Daniel Silva,
Mark Bilinski, Maxine Major, Justin Mauger,
Dan Muhleman
NIWC Pacific

Kimberly Ferguson-Walter
Laboratory for Advanced Cybersecurity Research

Abstract

Cyber deception is a strategy that defenders can leverage to gain an advantage over cyber attackers. The effects of deception on the attacker however, are not yet well understood. Quantifying the tangible and emotional effects of deception on the attacker's performance, beliefs, and emotional state are critical to deploying effective, targeted cyber deception. Our work uses data from a human-subjects experiment measuring the impact of cyber and psychological deception on over 100 professional red-teamers. These results demonstrate that an attacker's cognitive and emotional state can often be inferred from data already observed and collected by cyber defenders world-wide. Future work will leverage this observed data-set to formulate more informed defensive strategies.

1. Introduction

In the cyber domain, since there is no flawless defense, persistent attackers typically have the advantage. However, it is thought that deception could provide a means of flipping that dynamic in favor of the defender. There has been a good amount of theoretical and simulation research done to this end (Al-Shaer E., 2019). Deception strategies, such as the use of honeypots and decoys help bolster the defense of information systems (Cohen, 2006) by focusing on human elements of the attacker, and using those attributes to benefit defenders (Gutzwiller et al., 2018, 2019). Several steps will be necessary for full effect. First, identify the emotional/cognitive state of an attacker using data typically available to defenders. Next, alter this state using on-network interactions such as deception. Finally, create a mapping between change in state and desired change in attacker behavior (i.e., decreased confidence impedes reaching the goal). In this work we focus on the first step, laying the groundwork for potentially new impactful, customized defenses.

Typically, a defender has no way of directly assessing an attacker's emotional state; so, the idea is to instead infer it from network behavior (which the defender can observe). This paper explores if such an inference is possible. The results presented here, which to our knowledge are the first of their kind, were derived using data from a controlled human-subjects research (HSR) experiment designed to understand how defensive deception affects attackers. We define the ***emotional state*** of an attacker to be the levels of confusion, self-doubt, confidence, frustration, and surprise that an attacker reported experiencing throughout the day. Using the data from this experiment, we develop a classifier that can determine high/low levels of confusion and self-doubt with accuracy of roughly 0.68.

The contributions of this work are the following:

- We propose a simple, yet meaningful, model that captures attacker behavioral patterns.
- Using data collected from a real-world experiment involving cyber attack experts, we show that it may be possible in some cases to accurately infer some aspects of the emotional state of an attacker.
- We demonstrate that there appears to be a strong correlation between emotional state and the frequency an attacker performs reconnaissance and intrusion actions.
- We analyze the behavior of attackers who report extreme levels of confusion and self-doubt according to the proposed model.
- We evaluate the functional consequence of knowing emotional states by examining relationships to maximum threat potential.

For contribution 5), our main result is to show that seven independent variables account for 28.7% of the variance in threat potential, with attacker's emotional state accounting for a majority of it. These results together imply that not only may it be possible to infer an attacker's emotional state, but also that this

URI: https://hdl.handle.net/10125/102735
978-0-9981331-6-4

information is associated with attacker performance on the network.

2. Related Work

As cyber attacks have become increasingly more sophisticated, defending networks against such attacks has become even more difficult. In some instances, researchers have claimed that network defenses have "reached the limits of what traditional defenses... can do" (Heckman et al., 2015). Consequently, recent research into cyber defense has investigated new innovative strategies that leverage additional information such as deception as well as the attacker's cognitive/emotional state (Ferguson-Walter et al., 2021b; Veksler et al., 2020)

Researchers have measured the efficacy of deception through the use of decoys and the result such defensive tactics can have on attacker performance (Ferguson-Walter et al., 2021b; Fraunholz et al., 2018; Michael, 2002). However, a fuller understanding of using cyber deception for maximal impact is still needed. For example, the work in (Ferguson-Walter et al., 2021b) experimentally demonstrated that knowledge of deception may play a critical role in increasing the impact of the actual cyber deception that is present. (Heuer Jr, 1981) and (Yuill et al., 2006) also claim that the knowledge of deception heavily influences the decision-making process of an attacker. It has been proposed that in order to maximize the impact of any deceptive strategy, the cognitive/emotional state of the attacker should be taken into account (Cranford et al., 2020, 2021).

In (Veksler et al., 2020), the authors used symbolic deep learning to construct cognitive models of expert behavior. One of the goals was to develop models of attacker decision bias to reduce the risk of successful attacks. The goal of (Cranford et al., 2020) was to characterize cognitive state using the Instance-Based-Learning (IBL) model, to then later use this information to improve network defenses.

In this work, we explore the connection between the emotional state of a cyber attacker and thier observed network behavior. Previous studies on the subject (Al-Nafjan et al., 2017; Kim et al., 2013; Kotowski et al., 2020; Soroush et al., 2017) have leveraged machine learning algorithms to differentiate between emotional states provided a set of derived electroencephalography (EEG) signals as input. Accuracies of up to 94% have were reported for two-class discrete problems and up to 82% for four-class discrete problems. In (Ghosh et al., 2017) and (Trojahn et al., 2013), the authors considered the related problem of performing emotional recognition using keystrokes rather than EEG signals. For this setup, accuracies ranging between 77-88% were reported for two-class discrete problems and 84% for four-class. While the accuracy is high, the data used to achieve it is not something that can be realistically collected from malicious cyber intruders.

This work, which to the best of the authors' knowledge is the first of its kind, departs from previous research in the following ways: (1) We consider the problem of predicting an individual's emotional state within an ***adversarial cyber environment*** using HSR data from a real exercise. (2) Unlike previous work that usually aims to differentiate between two or four emotional states, we attempt to differentiate high/low levels of ***each attribute*** of an attacker's emotional state. As people can feel multiple emotions simultaneously, this is a more general approach and provides a much richer possibility of outputs that can potentially be leveraged by a defender in making decisions. (3) The input to each of our classifiers ***does not*** require signals from wearable devices or keystroke records, which are often impossible to obtain from a cyber attacker. Instead, the input to each of our classifiers is a derived set of features that can be obtained from an attacker's packet capture (PCAP) data, which a defender could obtain.

3. Tularosa Study

The Tularosa study was designed to understand the effects of deception, both real and psychological, on cyber attackers. During 17 sessions, over 130 experts participated in two days of network penetration testing. To measure the effect of deception, participants were either presented a network with decoys or no decoys, and were either told deception may be present or left uninformed of the possibility of deception. In addition to host and network traffic, data collected from each participant included cognitive surveys: fluid intelligence, working memory, personality, decision-making, task-specific questionnaires (TSQ), and qualitative self-reports. More details can be found in the Tularosa Study (Ferguson-Walter et al., 2018), its online Appendix, and results showing efficacy of deception (Ferguson-Walter et al., 2021b).

Participants worked independently and were given an initial foothold on their individual copy of the simulated target network. The network for participants with no deception included 25 real Windows and 25 real Linux machines representing a variety of operating systems, patch levels, and services. Networks with deception present included an additional 25 Windows decoys and 25 Linux decoys. Each decoy responded to

scans similarly to their real counterparts; however, the decoys did not respond the same as real machines to intrusions or exploits. Logon attempts on decoys always failed.

Participants were motivated to emulate an Advanced Persistent Threat (APT) and asked to: *"conduct recon on the network and locate vulnerable services, misconfigurations, and working exploits..."* using Kali Linux. In addition to the cyber task, the participants logged their findings and strategies in a variety of ways, including real time self-reporting, end-of-day briefings, and in the TSQ, which provides the participants' own evaluation of their emotional state for that day.

For the results discussed in Section 5, we made use of two sources of data: (a) A dataset consisting of known network-based cyber attacks identified from the raw network traffic (PCAP) for each participant, and (b) the TSQ survey. We included data from the 108 test subjects for whom we had both network attack data and a completed TSQ survey. For the TSQ survey, the participants were asked to assign scores between 1-5 indicating their levels of confusion, self-doubt, confidence, frustration, and surprise throughout the day. For shorthand, we refer to these as ***aspects of emotional state***. See (Ferguson-Walter et al., 2021a) for previously published results on a thematic analysis of the TSQ data.

In addition to TSQ scores, Section 6 also incorporates threat potential ratings and measures of personality and decision-making style. Threat potential ratings were derived based on a consensus between three cyber experts. A rating between $1-5$ was assigned for each of the 22 unique PCAP events, including success and failure outcomes for each event. Reconnaissance events (e.g., web requests) were coded as "1" whereas successful intrusions were coded as "5" (e.g., successful Server Message Block [SMB] logon). The middle scores ranging from 2-4 included more involved reconnaissance events (e.g., SMB scan), intrusion failures (e.g., SMB logon failure) and exploit failures (e.g., failed Eternalblue exploit), in order of increasing threat. Importantly, the threat potential ratings considered both the lethality of the action as well as the outcome (failures and successes). The highest threat potential rating for each subject was then extracted and labeled as ***maximum threat***, and refers to the highest level of potential damage each subject was capable of causing to the network. We also incorporated implicit traits including personality (Big Five Inventory; BFI, (John et al., 1999)) and decision-making style (General Decision-Making Style Inventory; GDMSI, (Scott and Bruce, 1995)) to determine the relative strength of the association between emotional state and behavior, compared to implicit traits. The same 108 subjects from Section 5 were included, however one subject was dropped due to a lack of BFI data.

The Tularosa Study occurred over 4 years ago and was conducted for a different purpose. Thus, while we extracted additional value from its data, there are notable limitations for our purposes. In the original experiment...While the TSQ data did capture aspects of emotional state, they did so only at the end of each day. Better resolution and/or other means of measuring emotional state would of course be helpful. Further, while our analysis establishes a number of interesting correlations, very specific subsequent experiments are needed to tease out causation.

4. Model

In this section, we first review the model used to characterize attacker behavior. Afterwards, we present the procedure used to generate the classifiers discussed in Section 5 with initial observations from Tularosa data.

Recall from the previous section, that in order to design our emotional classifiers, we leveraged both PCAP data and information from a TSQ survey that was collected at the end of of the first day. Each network event recorded in the PCAP data was assigned one of the following three labels: (1) recon, (2) intrusion, and (3) exploit. Events identified as recon occur when the attacker is gathering information about the environment. Intrusion events are those used to gain access to a system whereas exploit events compromise systems by causing them to enter into an insecure state. A more detailed description of this process can be found in Appendix A.

As a starting point, we defined the state of an attacker at any given time to be the label of the last recorded network event for that attacker. Since there are three states, and it is possible to go from any state to any other state, we characterized the behavior of an attacker over a fixed time interval according to a first-order discrete Markov chain, which can be represented as a set of 9 probabilities: $\left\{p_{1|1}, p_{2|1}, p_{3|1}, p_{1|2}, \ldots, p_{3|3}\right\}$, where $p_{i|j}$ denotes the probability of going to state i provided we are in state j. For shorthand, we refer to this model as the ***first-order behavioral Markov model***.

Despite its simplicity, the information contained in this model is correlated with the scores provided by participants during the TSQ survey. For example, we consider two extreme groups with respect to their reported frustration levels on the TSQ survey. The left and right state diagrams in Figure 1 respectively show the average transition probabilities of attackers. The attackers who reported a "5" on the TSQ for high frustration are shown on the left, whereas the attackers

who reported a "1" for low frustration are shown on the right. For both types of attackers, they each spend most of their time in recon and intrusion states. Furthermore, both types of attackers behave similarly given that they are in the exploit state. However, for the case where the attackers are in the intrusion state, the high frustration attacker is more likely to remain in the intrusion state than the low-frustration attacker. This could indicate frustrated attackers are more likely to continue pursuing a fruitless intrusion endeavor or that pursuing fruitless endeavors can increase frustration. In general, the high frustration attacker is less likely to launch an exploit regardless of its previous state. This further motivates creating defenses that impart frustration upon attackers.

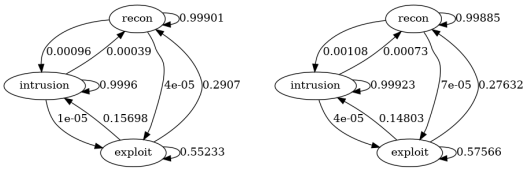


Figure 1. Behavior of high/low frustration attacker

Figure 2 displays from left to right the average transition probabilities for attackers who reported a high (4 or 5) confusion score along with attackers that reported a low (1) confusion score. Given that an attacker is in the recon state, the low-confusion attacker is more likely to trigger an intrusion or exploit event than a high-confusion attacker. This further motivates defensive strategies, like deception, which cause confusion to attackers.

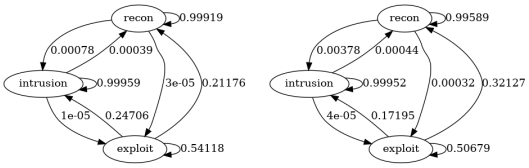


Figure 2. Behavior of high/low confusion attacker

In order to further explore this behavioral model and to develop a more accurate classifier, we constructed a set of 5 binary prototype classifiers (Wellek, 2002) (one for each of the 5 aspects of emotional state) that attempt to predict the emotional state based upon the following features:

1. Transition probabilities (9) from the first-order behavioral Markov model,
2. Number of seconds the attacker is idle,
3. Number of successful exploits,
4. Ratio of successful versus total exploits.

In order to compute feature 2) above, we first computed the number of idle seconds (rounded down) between the start time of every event (recon, intrusion, exploit) that originated with the attacker, then summed these terms.

For shorthand, we refer to the classifier that predicts confusion, self-doubt, confidence, frustration, or surprise as the confusion classifier, self-doubt classifier, etc. As will be described in more details in the next few paragraphs, the inputs to each of our classifiers will be a set of data points that represent some subset of the features (1-4) at various points in time.

The output of each of our classifiers is a binary integer where, for some particular emotional aspect, the output 0 or 1 indicate a low or high level, respectively. For instance, an output of 1 from the frustration classifier indicates that the attacker is experiencing high levels of frustration. The prototype classifiers presented here use the Euclidean distance in the feature space to determine distance, and only the data from individuals who reported high/low levels of each of the emotional aspects was used in this work.

In order to generate a set of data points that can be used for training and validation, we first partition each participant's data into non-overlapping intervals of 30 minutes. For each of these intervals, we compute the nine transition probabilities for the first-order behavioral Markov model along with the values of the other three features. These 9 transition probabilities along with number of idle seconds, successful exploits, and ratio of successful exploits comprise the set of 12 features that will be used for training. Altogether, this procedure resulted in 1336 data points that were used for both testing and validation. For more information regarding the generation of our data, see Appendix B.

Procedure 1 highlights the 3 steps of our methodology for developing binary prototype classifiers. Consider the classifier for emotional aspect A where $A \in \{\texttt{confusion}, \texttt{self-doubt}, \texttt{confidence}, \texttt{frustration}, \texttt{surprise}\}$. Our procedure takes as input the emotional aspect A along with the number of cluster centers that are labeled 1, denoted k_h, and the number of cluster centers labeled 0, denoted k_l. Using Elkan's algorithm, we run k-means clustering on the data labeled 1 given k_h centroids [1], and then we run k-means clustering on the data labeled 0 given k_l centroids. Finally, the classifier is evaluated using the shuffled leave-p-groups-out cross-validation (LPGOCV) method for 20 iterations each time considering 30% of the data as the validation set. The empirical results presented in Section 5 were generated using the RandomForestClassifier, StandardScalar, and KMeans libraries from from `scikit-learn`.

[1] The clustering algorithm partitions the data into groups of similar points (clusters), according some distance metric. The centroid represents the center of a particular cluster.

Procedure 1 ClassifyAspect

Input: $A \in \{\texttt{confusion}, \texttt{self-doubt}, \texttt{confidence}, \texttt{frustration}, \texttt{surprise}\}$,

Parameters: k_h, k_l

Training Stage:

1: Perform feature selection using random forest selection. For each of the 12 possible features, a feature is selected if its Gini Importance exceeds 0.02.

2: Scale the training data so that the resulting set of features have zero mean and unit variance.

3: Using only data labeled 1, perform k-means clustering provided a target of k_h centroids. Afterwards, using only data labeled 0, perform k-means clustering provided a target of k_l centroids.

Validation Stage:

Validate the training set using shuffled leave-p-groups-out for 20 iterations considering each time 30% of the data as the validation set, and performing steps 1)-3) in the training stage.

In order to get an indication of how well individual participants can be categorized into different behavioral groups, we ran k-means clustering on all 12 features and measured the inertia[2] as a function of the number of clusters using the data from all the participants (regardless of their TSQ scores). The inertia dropped from over 12000 to roughly 1500 when the number of clusters was increased from 1 to 20. Increasing the number of clusters beyond 20 resulted in a less significant decrease and the resulting inertia for 30 clusters for instance was around 1000. Motivated by this observation, the prototype classifiers considered in the next section will have no more than 20 centroids.

5. Results

For each of the five emotional aspects, we ran the `ClassifyAspect` approach outlined in Procedure 1 where we allowed k_l and k_h to each range independently from 2 to 10. Figure 3 displays the result of selecting for each emotional aspect, the parameters k_l and k_h that maximize the accuracy of the resulting classifier along with the total number of centroids used for each one.

As can be seen from Figure 3, each of the resulting classifiers have an accuracy of between 59% and 69%. From our data set, we were able to infer high/low levels of confidence and self-doubt with the highest accuracy whereas we were only able to determine frustration accurately around 59% of the time. There does not appear to be a strong correlation between the number of centroids and the accuracy of the resulting classifier. For instance, the confidence and frustration classifiers have 8 and 7 centroids, respectively, yet the accuracy of the frustration classifier is much lower than confidence.

We found that certain behaviors (under our first-order behavioral Markov model) are more

[2]The inertia is the sum of the Euclidean distances between each data point and its centroid.

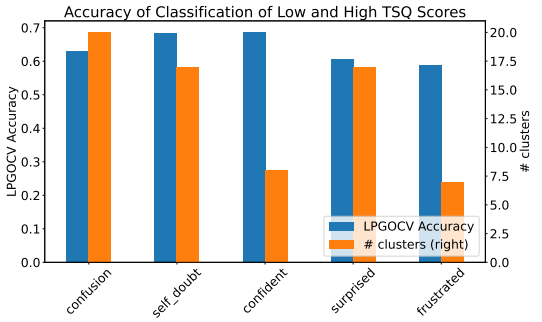


Figure 3. Emotional Aspects Classifiers

correlated with high/low levels of various emotional aspects than others. Table 1 displays the values of k_h, k_l that achieved the maximum average accuracy. The average accuracy of the classifier appears in the third column[3]. The fourth column shows the average number of centroids found during the validation step described in Procedure 1 that had precision exceeding 70%. For example, the fifth column, second row of indicates that on average there was a set, say S of 8.9 centroids with the property that when the input data point was closest in Euclidean distance to one of the centroids in S, then the output of the classifier (which indicates a high/low level of confusion) is correct at least 70% of the time.

	k_l	k_h	accuracy	# precise
confusion	10	10	0.63	8.9
self-doubt	8	9	0.68	8.75
confidence	5	3	0.69	3.7
surprise	7	10	0.61	7.05
frustrated	3	4	0.59	1.95

Table 1. Centroid count and accuracies

In Figure 4, we show the 5 most important features using the Gini Importance measure for each of our 5 classifiers. For example, the first set of 5 bars show the 5 features with the largest Gini Importance for confusion. The feature which counts the number of idle seconds is denoted as "idle" on the x-axis. The remaining features are related to transition probabilities in our first-order behavioral Markov model. We abbreviated these transitions using the notation "state 1 -> state 2" where we denote the states by their first letters only. For example, "R -> I" represents recon to intrusion. [4]

From Figure 4, notice that the set of top 5 features are the same for each of our classifiers, and the number of idle seconds consistently had the highest Gini Importance score. However, the relative importance of the remaining 4 (of the top 5 features) varied

[3]Note that for our setup, we are considering the average accuracy over each of the validation sets.

[4]The data shown below for Gini Importance was derived using the RandomForestClassifier available from `scikit-learn` (Scikit-Learn, 2022).

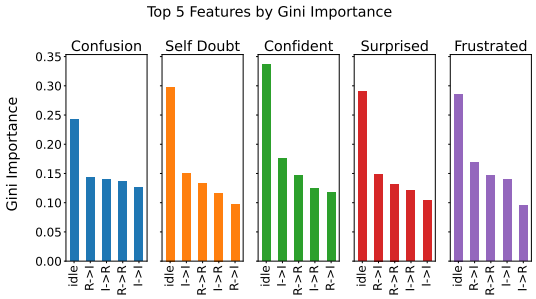


Figure 4. Top Features According to Gini Importance

depending on the particular classifier. For example, the recon to intrusion transition probability was the second highest feature for confusion, surprise, and frustration whereas the transition probability from intrusion to intrusion was the second highest scoring feature for self-doubt and confidence. Another interesting trend from the observed data is that the transition probabilities involving intrusion and recon seem more important than transition probabilities involving exploit. None of the features involving the exploit state appear in Figure 4.

In order to better capture behaviors that are correlated with high/low levels of various emotional aspects, we first perform the same procedure as described in Procedure 1 except for two key differences: (1) In the training stage, we do not perform feature selection (step 1)), and (2) Rather than perform leave-p-groups out cross-validation, we set aside 30% of the data set (randomly chosen) for validation and train on the remaining 70%. For shorthand, we will refer to the resulting procedure as `ClassifyAspect2`.

We ran `ClassifyAspect2` for confidence with $k_l = 5$ and $k_h = 3$. Note that this number of clusters gave the highest accuracy with respect to confidence according to Table 1. Figure 5 depicts the centroid with the label 0 that had the highest precision (from among the 5 possible centroids) and the centroid with the label 1 with the highest precision (from among the 3 possible). For instance, the feature corresponding to the transition probability from recon to intrusion for the centroid on the right has value 0.3333 and the feature corresponding to the transition probability from recon to recon for the same centroid has value 0.66667. We display the features associated with the transition probabilities graphically to better convey these two centroids where the high confidence centroid is displayed on the left and the low confidence on the right in Figure 5. Note that since these plots are representing centroids (and are not actual conditional probabilities), the sum of the arrows going out from any fixed state need not sum to one.

Perhaps not surprisingly, a low confidence attacker is more likely to remain in the recon state than a high confidence attacker. Furthermore, as expected, high confidence individuals are more likely to launch an intrusion than low confidence individuals. For the centroids depicted, we can see that individuals associated with the high confidence centroid are more likely to go from the recon state to the intrusion state.

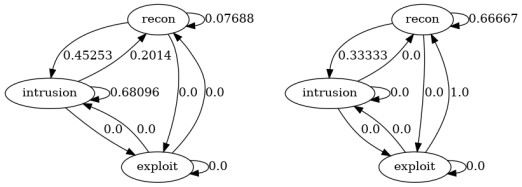


Figure 5. High/Low Confidence Centroids

Running `ClassifyAspect2` for self-doubt with $k_l = 8$ and $k_h = 9$ gives Figure 6 where the image on the left and right represent the transition probability features for the highest precision centroids labeled 1 and 0, respectively. We see a similar pattern here as in Figure 5 where a low self-doubt attacker is more likely to to launch an intrusion provided they are already in the intrusion state, and a high self-doubt attacker is more likely to remain in the recon stage.

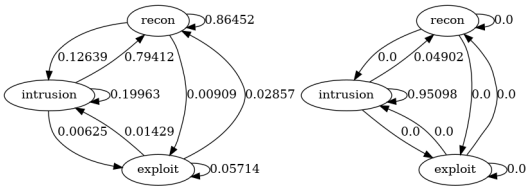


Figure 6. High/Low Self-Doubt Centroids

6. Relationship to Threat Potential

Recall that the TSQ scores were self-reported at the end of the day, and not broken down into 30 minute windows, as the PCAP data was in the classifier model. This temporal difference in measurement between emotional state and behavior may have impacted our ability to get highest possible classifier accuracy. To help address this, the relationship between emotional state and overall behavior was assessed another way using maximum threat potential ratings, which have better temporal concordance with the TSQ scores. To test the relative strength of the association between explicit emotional state (TSQ) and behavior, we also included implicit measures (BFI, GDMSI) of attacker cognition for comparison.

107 subjects were used to analyze the relationships between cognition (TSQ, BFI, GDMSI) and maximum threat potential. In this section, relationships between psychological traits and maximum threat ratings were

assessed using all the TSQ scores rather than just the extremes (as in Section 5) to allow for linear modeling of the variance. First, separate individual correlations were run between attacker traits (TSQ, BFI and GDMSI measures) and maximum damage ratings, to assess the underlying relationships between the variables. Next, a step-wise linear regression with backwards selection was run in R (R-Project, 2022) using the stepAIC function. The outcome variable was the maximum threat ratings, and the entered candidate variables were emotional states (TSQ), personality (BFI) and decision-making styles (GDMSI). The top predictors were chosen based on smallest akaike information criterion (AIC) for the model. The output took the form of a linear model, $\hat{y_i} = \hat{B_0} + \hat{B_1}x_1 + \hat{\epsilon}$, where $\hat{\epsilon}$ is sampled from $N(0, \sigma^2)$. $\hat{B_0}$ is the intercept, $\hat{B_1}$ is the estimate coefficient for the first independent variable in the model (followed by other independent variables if more than one was included), and $\hat{\epsilon}$ is calculated using the residual standard error (σ). For step-wise regression, the final model was: $\hat{y_i} = 6.64 - 0.26x_1 + 0.14x_2 - 0.02x_3 - 0.02x_4 - 0.03x_5 - 0.04x_6 + 0.04x_7 + \hat{\epsilon}$ with Adjusted $R^2 = 0.29$, $F(7, 99) = 7.08, p < 0.001$. The top independent variables were confusion (TSQ), confidence (TSQ), extraversion (BFI), agreeableness (BFI), openness (BFI), avoidant (GDMSI) and spontaneous decision-making styles (GDMSI), and correspond in order to x_1 through x_7, see Figure 8. Together, these explain 28.7% of variance in maximum threat potential, and the strongest associations to maximum threat potential were the emotional states.

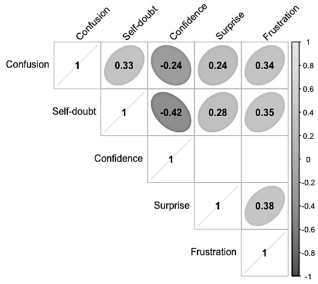
	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	6.64	0.71	9.41	0.00
Confusion	-0.26	0.07	-3.94	0.00
Confidence	0.14	0.06	2.35	0.02
Extraversion	-0.02	0.01	-2.29	0.02
Agreeableness	-0.02	0.01	-1.49	0.14
Openness	-0.03	0.01	-2.34	0.02
Avoidant	-0.04	0.01	-3.10	0.00
Spontaneous	0.04	0.02	2.37	0.02

Outcome variable: Maximum threat potential

Table 2. Final Step-Wise Regression Model

Since pre-existing relationships between independent variables (i.e., multi-collinearity) could have affected the final linear model, a correlation matrix was used to quantify the amount of overlap within TSQ measures. Most TSQ measures were significantly ($p < 0.0125$, Bonferroni corrected) correlated with each other, except confidence, which was not significantly correlated to surprise or frustration ($p > 0.05$), see Figure 8. The direction of the associations were as expected (e.g. lower self-doubt was associated with higher confidence; Spearman's Rho = -0.42), and the amount of overlap between the TSQ measures ranged from 17% to 5.8% (Spearman's Rho^2), with confidence and confusion showing the lowest overlap, see Figure 7.

Building on this, variance inflation factors (VIF) were calculated to quantify the amount of collinearity between all independent variables in the step-wise regression. VIFs were calculated ($1/(1-R^2)$) using the VIF function in R (Naimi et al., 2014). Variables with VIF > 10 should be dropped from the model because they display a strong relationship to other independent variables (typically demographic variables such as age, sex). For our results, VIFs were less than 2.8 for all variables in the final model, suggesting that the influence of collinearity was minimal (Welton et al., 2020). Although there was some degree of correlation between some TSQ measures (Figure 7, left), total collinearity between independent variables in the final model was small.



Variable	VIF
Confusion	1.45
Confidence	1.39
Extraversion	1.18
Agreeableness	1.79
Openness	1.40
Avoidant	2.71
Spontaneous	2.57

Figure 7. TSQ Correlation Matrix

7. Conclusions, Limitations, Future Work

One of the primary challenges of cyber defense is overcoming the asymmetry between attacker and defender: where an attacker only needs to find a single weakness in the network, while the defender must keep the whole network safe. Overcoming this asymmetry by allocating ever more resources to defense is untenable, so gaining special insight into attacker psychology could help reverse the asymmetry and provide defenders another tool for protecting information systems.

The current study takes steps towards that goal by demonstrating that an attacker's emotional state is closely associated with their behavior on a network. Specifically, in Figure 6 we demonstrate that emotions such as confidence (low vs. high) are associated with an attacker staying in the recon stage rather than proceeding to the intrusion stage. Preliminary analysis across multiple days has been promising as well, suggesting that the trends discussed in

Sections 4 and 5 may hold in more general settings. Table 2 demonstrates that explicit emotional states are unique compared to implicit measures (BFI or GDMSI) by explaining the bulk of the variance in maximum threat potential; higher confidence and spontaneous decision-making, and lower confusion and agreeableness was associated with greater threat potential. More work is needed to fine-tune defensive responses to maximize negative impact on attacker activity, especially more cyber-relevant experiments built upon the emotional state of attackers (Johnson, 2022).

This study builds on previous work (Ferguson-Walter et al., 2018, 2021b) to demonstrate that the psychology of a cyber adversary is associated with their actions on a network. As next steps, it would be valuable to investigate associations between certain psychological features and behavior at different temporal stages of an attack, since Section 5 demonstrated that emotional state during early stages of the kill chain is correlated with behavior at later stages. It would be interesting to investigate how behavior can be manipulated through coordinated deployment of deception to keep attackers in the recon stage by affecting psychological features like confidence and self-doubt. Regarding limitations, the ability to collect certain features likely affected our results. PCAP data was collected every 30 minutes, whereas TSQ data was only collected at the end of the day, thereby reducing our ability to match varying cognitive states with performance throughout the day. Thus, cognitive measures are more likely to reflect overall features of attackers rather than precise changes in psychology as they performed the task. Despite this limitation, our results still indicated a significant correlation between PCAP data and TSQ measures. This suggests that even generalized measures of cognitive state can be captured using PCAP data, at least to some degree. Future experiments with more temporal precision would be informative. Another limitation was that we were not able to fully examine the effects of psychological deception and decoy presence due to a high attrition rate in participants. For example, although 130 participants were initially recruited only 107 had intact PCAP and TSQ data sets, this reduction across deception conditions limited the current study's statistical power. Future studies which specifically examine this impact of deception would be highly valuable.

The ability to identify and measure a cyber attacker's emotional state sets the stage for shifting the advantage to defenders. Recent work on understanding decision-making biases of cyber attackers further demonstrates the utility of understanding and manipulating an attacker's mental state for defensive advantage (Johnson et al., 2021). Exploitation of the Sunk Cost Fallacy is one notable example. HSR using the CYPHER game indicates that this decision-making bias exists in cyber-relevant progress decisions, and can be induced by presenting certain kinds of scenarios to the cyber attacker (Johnson, 2022). While CYPHER investigated the role of uncertainty, results were inconclusive, and other emotional states were not explored. As the linkage between explicit emotional traits and cyber attack behavior become clearer, our work on emotional state classification can augment this kind of HSR by providing additional indicators of when the bias elicitation is effective, or when its use is most warranted.

Adaptation of deceptive elements like honeypots, decoys, and honeythings, are quickly moving from manual to automated (Al-Shaer et al., 2019). However, how and when these deceptive elements should be adapted is still an open question. If it can be shown that emotional state has a causal effect on an attacker's success, then learning how to influence the attacker's emotional state could be a first step towards preventing network penetration.

References

Al-Nafjan, A., Hosny, M., Al-Ohali, Y., and Al-Wabil, A. (2017). Review and classification of emotion recognition based on eeg brain-computer interface system research: a systematic review. *Applied Sciences*, 7(12):1239.

Al-Shaer, E., Wei, J., Kevin, W., and Wang, C. (2019). Autonomous cyber deception. *Springer*.

Cohen, F. (2006). The use of deception techniques: Honeypots and decoys. *Handbook of Information Security*, 3(1):646–655.

Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., and Lebiere, C. (2020). Toward personalized deceptive signaling for cyber defense using cognitive models. *Topics in Cognitive Science*, 12(3):992–1011.

Cranford, E. A., Gonzalez, C., Aggarwal, P., Tambe, M., Cooney, S., and Lebiere, C. (2021). Towards a cognitive theory of cyber deception. *Cognitive Science*, 45(7):e13013.

Ferguson-Walter, K., Shade, T., Rogers, A., Trumbo, M. C. S., Nauer, K. S., Divis, K. M., Jones, A., Combs, A., and Abbott, R. G. (2018). The tularosa study: An experimental design and implementation to quantify the effectiveness of cyber deception. Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Ferguson-Walter, K. J., Gutzwiller, R. S., Scott, D. D., and Johnson, C. J. (2021a). Oppositional human factors in cybersecurity: A preliminary analysis of affective states. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, pages 153–158. IEEE.

Ferguson-Walter, K. J., Major, M. M., Johnson, C. K., and Muhleman, D. H. (2021b). Examining the efficacy of decoy-based and psychological cyber deception. In *30th*

USENIX Security Symposium (USENIX Security 21), pages 1127–1144.

Fraunholz, D., Anton, S. D., Lipps, C., Reti, D., Krohmer, D., Pohl, F., Tammen, M., and Schotten, H. D. (2018). Demystifying deception technology: A survey. *arXiv preprint arXiv:1804.06196*.

Ghosh, S., Ganguly, N., Mitra, B., and De, P. (2017). Evaluating effectiveness of smartphone typing as an indicator of user emotion. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 146–151. IEEE.

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

Gutzwiller, R., Ferguson-Walter, K., Fugate, S., and Rogers, A. (2018). "oh, look, a butterfly!" a framework for distracting attackers to improve cyber defense. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 272–276. SAGE Publications Sage CA: Los Angeles, CA.

Heckman, K. E., Stech, F. J., Thomas, R. K., Schmoker, B., and Tsow, A. W. (2015). Cyber denial, deception and counter deception. *Advances in Information Security*, 64.

Heuer Jr, R. J. (1981). Cognitive factors in deception and counterdeception. In *Strategic military deception*, pages 31–69. Elsevier.

John, O. P., Srivastava, S., et al. (1999). The big-five trait taxonomy: History, measurement, and theoretical perspectives.

Johnson, C. K. (2022). *Decision-Making Biases in Cybersecurity: Measuring the Impact of the Sunk Cost Fallacy to Delay and Disrupt Attacker Behavior*. Dissertation, Arizona State University, Mesa, AZ.

Johnson, C. K., Gutzwiller, R. S., Gervais, J., and Ferguson-Walter, K. J. (2021). Decision-making biases and cyber attackers. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, pages 140–144. IEEE.

Kim, M.-K., Kim, M., Oh, E., and Kim, S.-P. (2013). A review on the computational methods for emotional state estimation from the human eeg. *Computational and mathematical methods in medicine*, 2013.

Kotowski, K., Fabian, P., and Stapor, K. (2020). 2 machine learning approach to automatic recognition of emotions based on bioelectrical brain activity. In *Simulations in Medicine*, pages 15–34. De Gruyter.

Michael, J. B. (2002). On the response policy of software decoys: Conducting software-based deception in the cyber battlespace. In *Proceedings 26th Annual International Computer Software and Applications*, pages 957–962. IEEE.

Naimi, B., Hamm, N. A., Groen, T. A., Skidmore, A. K., and Toxopeus, A. G. (2014). Where is positional uncertainty a problem for species distribution modelling? *Ecography*, 37(2):191–203.

R-Project (2022). https://r-project.org.

Scikit-Learn (2022). https://scikit-learn.org/stable/modules/clustering.html.

Scott, S. G. and Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and psychological measurement*, 55(5):818–831.

Soroush, M. Z., Maghooli, K., Setarehdan, S. K., and Nasrabadi, A. M. (2017). A review on eeg signals based emotion recognition. *International Clinical Neuroscience Journal*, 4(4):118.

Trojahn, M., Arndt, F., Weinmann, M., and Ortmeier, F. (2013). Emotion recognition through keystroke dynamics on touchscreen keyboards. In *ICEIS (3)*, pages 31–37.

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

Wellek, S. (2002). *Testing statistical hypotheses of equivalence*. Chapman and Hall/CRC.

Welton, J. M., Walker, C., Riney, K., Ng, A., Todd, L., and D'Souza, W. J. (2020). Quality of life and its association with comorbidities and adverse events from antiepileptic medications: Online survey of patients with epilepsy in australia. *Epilepsy & Behavior*, 104:106856.

Yuill, J., Denning, D., and Feer, F. (2006). Using deception to hide things from hackers: Processes, principles, and techniques. *Journal of Information Warfare*, 5(3):26–40.

A. Data Curation Process

Data used for this analysis were extracted from the Tularosa network traffic packet captures (PCAPs). Data are categorized into one of three types of events: *reconnaissance* (recon), *exploits*, and *intrusions*.

To generate this dataset, a cyber expert reviewed screen recordings for participants with high levels of activity and successes such as gaining access to the domain controller or compromising multiple Windows boxes. The expert selected actions which were frequently observed among the participants, examined the cyber event in the network data using tools such as Wireshark and developed python scripts to automatically identify and extract these events from the PCAP data. This list of events was validated by the data analysis team by informally comparing those actions to other cyber data sources such as the participant's self-report logs, keystrokes, and IDS alerts. The cyber events used for this analysis are described in Table 3.

There are a few limitations to this dataset. The cyber activity was only evaluated by a single cyber expert, and while the most impactful and most prevalent events were prioritized, it is not exhaustive of all events. Additionally, the data points identified here are not comprehensive—many smaller attacks used by a subset of participants, or the same attacks executed in a unique way—will not be identified by the script looking for specific indicators of that attack. However, these events are prolific among the entire participant pool and demonstrate nearly universal paths to success on the cyber task. Data collected from this selection is robust enough to apply, learn, and test multiple hypotheses.

B. Generation of Feature Data Points

For the purposes of this study, we only included data from individuals who rated high or low on the TSQ survey provided at the end of the first day. Table 4 summarizes how we determined high/low levels of the various emotional aspects along with the number of attackers that were subsequently characterized as possessing high/low levels of certain emotional aspects accordingly. For example, the second row conveys that individuals who reported a score of 1 for confusion on the TSQ survey are labeled as low-confusion while those who reported a score of 4 or 5 are labeled

Recon Events	*Description*
Port/Service Scan	Repeat `RST/ACK` flags indicating "port closed" messages.
SMB Scan (enum4linux, etc.)	SMB auth attempt with username "`NULL`".
DNS Query	Standard DNS Query
NMap (SMB) Script Scan	Hard-coded values of `\\winreg` and `\\router` used during nmap SMB script scanning.
Nmap HTTP Scan	Web scans with "Nmap" in `user_agent` string.
Nikto HTTP Scan	Web scans with "Nikto" in `user_agent` string.
Web Request	Web scan with other `user_agent`.

Exploit Events	*Description*
MS-08_010: Eternalblue Attempt	Metasploit payload includes the distinct memory offset of a string of 'A's.
Eternalblue 117 Byte Negotiate	Initial Eternalblue packet, useful to detect attacks on decoys, invalid targets.
vsFTPd v. 2.3.4 Backdoor Attempts	Attempt to log into FTP with a "`:)`" username.
SMB MS09-050 Attempt	Metasploit attack on Win. Vista & Win. 8, despite targets absent from network.

Intrusion Events	*Description*
SMB Logon Attempt	SMB connection from host to target with `messagetype == NTLMSSP_AUTH`. Successful if subsequent packet stream matches login, and frame higher than login.
SMB Logon Failure	SMB response packet contains failure code `0xc000006d`.
Reverse Shell (Successes)	Successful TCP sessions initiating from target back to attack machine on high ports (`>1000`). Most subjects used default port `4444`.
VNC Connection	A VNC authorization sent to connect to a target.

Table 3. Cyber Activities Identified in PCAP Data

high-confusion. According to this labeling, there are 17 high-confusion individuals and 10 low-confusion individuals. The confusion classifier was trained and validated using data from these 27 individuals.

Although different numbers of individuals were used to generate high/low data points, the data was normalized so that (for each aspect) the number of datapoints labeled high was equal to the number of datapoints equal to 0. Because we partitioned each participant's data into non-overlapping intervals of 30 minutes, each participant could contribute up to 18 data points[5]. Our implicit assumption in this approach was that even when the emotional state is known, the attacker behavior is not necessarily homogeneous.

C. Additional Results and Observations

Table 5 includes additional details not discussed in Section 5. In particular, this table shows the total number of features used on average by the respective classifiers along with the average number of clusters that had precision at least 65%, which appears as the fourth column, along with the average number of clusters with precision at least 70%, which is displayed in the fifth column. Notice that the information in the fifth column of Table 5 is the same information contained in the last column (fifth column) of Table 1. From Table 5, it can be observed that the number of features was fairly consistent amongst the five different classifiers where the average number of features ranged from 7.2 for confidence to 9.85 for surprise. Overall, there doesn't seem to be a strong relationship between the number of features and accuracy of the classifier. Both the most and least accurate classifiers had the smallest number of features and the second most accurate classifier (self-doubt) had a rather large number of features.

[5] 9 hours $\times$ 2 data samples per hour

	Low	High	# High	# Low
confusion	1	[4, 5]	17	10
self-doubt	1	[4, 5]	10	18
confidence	1	5	6	11
surprise	1	[4, 5]	31	26
frustrated	1	5	17	12

Table 4. High/Low Data Points

	accuracy	features	≥ 65	≥ 70
confusion	0.63	9.05	10.3	8.9
self-doubt	0.68	9.15	9.85	8.75
confidence	0.69	7.2	4.6	3.7
surprise	0.61	9.85	9.15	7.05
frustrated	0.59	8.6	2.3	1.95

Table 5. Centroid count and precise centers

The recall, precision and F1 scores for each of our five classifiers is displayed in Figure 8. The performance of our classifiers with respect to these metrics varies significantly. Both the confidence and self-doubt classifiers have precision above 70%. Recall for self-doubt is 0.72. However, these scores are much lower for our other classifiers – recall and precision for frustration is only 0.48 and 0.63 respectively.

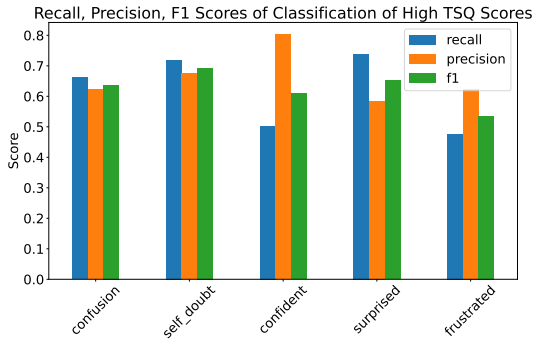


Figure 8. Additional Metrics