The World of CTF is Not Enough Data: Lessons Learned from a Cyber Deception Experiment

Conference Paper · December 2019

DOI: 10.1109/CIC48465.2019.00048

Kimberly Ferguson-Walter
Department of Defense

Sunny Fugate
Naval Information Warfare Center

Robert S. Gutzwiller
Arizona State University

The World of CTF is Not Enough Data: Lessons Learned from a Cyber Deception Experiment

Kimberly J. Ferguson-Walter
Laboratory for Advanced Cybersecurity Research

Maxine M. Major
Naval Information Warfare Center - Pacific

Dirk C. van Bruggen
Laboratory for Advanced Cybersecurity Research

Sunny J. Fugate
Naval Information Warfare Center - Pacific

Robert S. Gutzwiller
Arizona State University

Abstract

The human side of cyber is fundamentally important to understanding and improving cyber operations. With the exception of Capture the Flag (CTF) exercises, cyber testing and experimentation tends to ignore the human attacker. While traditional CTF events include a deeply rooted human component, they rarely aim to measure human performance, cognition, or psychology. We argue that CTF is not sufficient for measuring these aspects of the human; instead, we examine the value in performing red team behavioral and cognitive testing in a large-scale, controlled human-subject experiment. In this paper we describe the pros and cons of performing this type of experimentation and provide detailed exposition of the data collection and experimental controls used during the Tularosa Study [10]. Finally, we will discuss lessons learned and how our experiences can inform best practices in future cyber operations studies of human behavior and cognition.

1 Introduction

Adversary emulation (i.e., red teaming) and capture the flag (CTF) events have become a steady part of computer security culture and education. CTF events are being held annually across universities (e.g., National Collegiate Cyber Defense Competition), military academies (e.g., OverTheWire: Wargames), and conferences (e.g., DEF CON Hacking Conference CTF Competition). CTFs provide a platform for competition, training, assessment, and recruitment, and are therefore valuable exercises. Research on how to construct more useable, repeatable, and realistic CTF environments does exist [7]. Yet, for measurement, these events are often designed to assess system performance, security, and defensibility, serving the purpose of training or supporting system security with few attempts to focus on measuring human performance [3, 17]. While it is useful to have CTFs which are focused on solving puzzles, the data collected (if any) is often opportunistic, sparse, and lacking in experimental controls or validity. As a result these events have provided few contributions to understanding the people behind the attacks.

Studying red teamer behavior and cognition can lead to understanding how to gain improvements in adversary emulation performance and efficiency. These insights can consequentially improve the security and defensibility of systems and networks. By better understanding the human performance and human factors aspects of experts who are trained in adversary emulation, we also better understand the adversary, and improve our defenses against them. The human component of the cyber attack deserves more investigation.

The Human Factors community has only recently begun to address cognitive and social issues surrounding the humans involved in cyber defense [5, 14, 20]. Research focused on the human component of the cyber *attack* is even less mature, but is of growing interest to the community [8]. Oppositional Human Factors (OHF) [13] was introduced into the human factors community in 2018 and is described as the theory and practice of inverting human factor guidelines to degrade and disrupt the usability and progress of a cyber attacker for defensive purposes. OHF advocates for measuring the performance, cognition, and psychology of attackers by applying thoroughly studied and well understood experimental paradigms drawn from decades of psychology and human factors research.

Collecting the necessary data to understand attacker cognition and behavior from current CTF exercises is not realistic. As designed, information about individual participant actions, perceptions, and feelings are not isolated or collected. These are not intended to be rigorous controlled studies and are not required to complete Institutional Review Board (IRB) approval for human subjects research (HSR). They fundamentally serve a different purpose.

Large-scale HSR studies on cyber attacker behavior are required to study human behavior and cognition. This knowledge is crucial to cybersecurity research focused on resilient, reliable, and adaptive defenses. Red teamer behavior differs from unauthorized hacking, but research focused on better understanding adversary emulation can still fill a critical gap in knowledge.

Cyber Deception is an active area of cybersecurity research that fits within the OHF paradigm as deception is often focused on effecting the human behind the attack to disrupt, delay and deter their progress. Research has sought to address effective deception for many years [15], and deception has been a primary focus of several human subject experiments [1, 4]. These studies often use non-expert participants, are uncontrolled with limited data collection, and/or lack realism due to use of simplistic games or scenarios.

In order to provide concrete examples, this paper expands upon the experimental design, motivation, and limitations of the Tularosa Study [10], which was a large-scale HSR study of 138[1] professional red teamers. The principal purpose of this experiment was to measure the efficacy of both cyber deception and psychological deception on expert participants given the task of attacking and exploiting computing systems on a controlled test network. Participants in the study were split into four different conditions, where cyber deception and/or psychological deception were present or absent (details can be found in the original paper).

In this paper, we will provide additional detailed discussions of the Tularosa experimental design as it attempted to fill a role that CTF can not, including data collection and processing methods used on the test network. We will discuss design decisions which led to a test network and methodology which differs substantially from a standard CTF. Finally, we will outline lessons learned and trade-offs from multiple experimental stages including design, execution, and data collection, and highlight these findings with insights drawn from participants' self-reports. Our conclusions will address the inadequacies of CTFs for studying humans behavior and provide guidance for designing future experiments.

2 Why CTFs are Not Enough

While CTF environments range widely in complexity, appearance and technical specification, conceptually they have many commonalities that draw them apart from other types of environments and data sources.

2.1 Research Environment Limitations

Cybersecurity research has a spectrum of options available, each with pros and cons.

Datasets: Extensive work has gone into creating many different cybersecurity datasets with Zheng et al. finding that from 2012-2016, there were over 900 datasets created [22]. However, the datasets that were identified as "attacker-related" primarily capture data about attacks, vulnerabilities, exploits, and cyber-crime activities instead of the attackers themselves. These datasets are useful for analyzing attacker techniques and evaluating detection methods but lack the features necessary to deeply understand attacker responses to various defensive mitigations (such as biases or other OHF techniques [12]). Finally, the nature of static datasets makes it difficult to assess changes in behavior when attackers encounter mitigations.

CTFs: CTFs have been proposed as great sources of data for research [7, 19] however they vary in the level of realism used. CTF events commonly pit red teamers against a static environment with the objective of breaking into systems and collecting flags. More realistic attack/defend CTFs pit red (offensive) and blue (defensive) teamers against one another to emulate real-world cyber scenarios, albeit at a potentially accelerated pace. These events are great training environments because they offer a safe way to learn [21]. However, given the team-based nature and competitive aspects, they are difficult to use for controlled experiments on human subjects. Additionally, pitting red against blue introduces variance in behavior, with different blue teams reacting (defending) differently, which makes comparison across red teams difficult. With or without the presence of a blue team, data collected is insufficient for understanding attacker behavior and cognition. This is expanded upon in Sections 3.3 and 3.2.

Real World Trials: Bug bounties are a popular method for evaluating and protecting products [6]. These programs define scope and rules of engagement for products and then offer to pay anyone able to find specific classes of bugs within a program (e.g. bugs that lead to remote code execution). While beneficial for security researchers and companies alike, it is difficult to make conclusions about the effects of a system's defenses based only on the outcome of bugs being found or not. Once again, as described above, the variability of humans means one would want to know just how many bug bounty hunters are working on the problem (and how good they are) before any conclusions could be drawn.

Levesque et al. [16] proposed increasing the ecological validity (real-world applicability) through the use of clinical trials. Additionally, Forget et al. also conducted studies which instrumented real-world users to observe and analyze behavior "in the wild" [11]. While these approaches drastically improve the external validity of a study, they trade-off the benefits of controlling conditions surrounding attackers.

Experiments: The DETER [2] community has focused on being able to create repeatable experiments in a controlled environment. While the majority of work has focused on testing technical solutions, not effects on attackers, work such as DEXTAR [18] which incorporated use of DETER resources, did focus on human cyber defense operators. The Tularosa Study starts to address cyber attacker behavior and cognition.

2.2 Participant Motivation

The placement and collection of flags in CTFs does little to promote or account for intrinsic motivation of cyber attackers. For research questions focused on cognition and behavior, it

[1] Since publication an additional session was completed raising the number of participants.

is important to understand attacker motivation. This could vary by sophistication of adversary, goals, resources, etc. To make an experiment more realistic, it is important to try and replicate this motivation in the subject population. CTFs (and experiments) are time bounded which can increase the pressure on a participant to find and exploit vulnerabilities to achieve the goal of collecting more flags and "winning" the competition. Bug bounties make use of extrinsic rewards (e.g. cash, equipment) as a primary form of motivation. It is important to consider whether or not such methods of motivation align to real world scenarios and what the impact of that motivation may have on the experiment. For example, does the method of motivation impact a strategy an attacker chooses to use? How might that affect the conclusions of the study? What methods could be used to mitigate such an impact?

3 Why Tularosa is Different

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

3.1 Experimental Validity

In contrast to the the bring-your-own-device policy of most CTFs, human subjects studies must control for internal validity by ensuring each subject has the same tools and the same target environment, which ensures an equal chance to succeed. What we cannot control are the participants themselves and the methods they choose to use. These uncontrolled variables make the process of evaluating human performance in a more realistic scenario much more challenging than scoring CTFs or evaluating bug bounty success.

Examples of factors over which we have little control include: 1) Participants have the same skill ("red teamer"), but their specialized experience and subset of skills varied drastically (from 0.25 years to 20 years of experience in network penetration); 2) Methods of compromise were typical of red teamers - the use of the same tools (e.g. nmap, hping3, Wireshark, Metasploit) - but the use of tools and techniques varied based on subject's background and level of familiarity; 3) Indicators of compromise and ease of detection within the data set varied based on the participants' method of attack. Even the same attack, executed with a different tool, generated different data in order to provide proof of the subject's success.

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

3.2 Human Subjects Research

Human subjects research requires IRB approval and voluntary consent of the participants. This is different from how CTFs operate. CTF data collected does not focus on human behavior or cognition, but rather on network activity and flag capture. CTFs are limited because they do not collect information on the humans–their expertise, experiences, thoughts, or feelings.

3.2.1 HSR Data

Studies in both computer and behavioral science have investigated different methods for soliciting details on human cognition, emotion, and decision making during studies. Having a trained expert observer being able to observe and question the participant might provide the most detailed information about their thoughts and feelings, but the trade-off is evident if there are any timing research questions being pursued, as these questions can take participants off task. On the other end of the spectrum, open ended reporting, during and after the study can be less disruptive, but the freedom it allows leads to a wide variance of useful information reported.

This was the method used in the Tularosa Study. Reporting in real time was needed in order to correlate self-reported participant cognition and emotion with the time-stamped cyber and physiological data. Additional reporting and questionnaires were provided at the end of the task. The variance in real-time reporting ranged from zero to 304 chat messages in a day with 17% providing zero for at least one day and under 5% skipping at least one of the end-of-task red team reports.

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

While the data collected from the Tularosa Study will no doubt be used to answer various research questions across the community, the study was designed to answer specific research questions on the effects of cyber and psychological deception on the success of a red teamer when performing a network penetration task. In order to properly proctor and monitor the execution of the study, sessions were limited to at most 10 participants. Precautions were taken to reduce the number of participants at one time, group similar conditions together when possible, provide written instruction, and reduce the chance of participants discussing or interacting with each other during the study to reduce possible bias in the results. These reasons, in addition to the desire to increase the number of participants per condition, influenced the decision not to have participants work in teams. However, these decisions forced proctors to run sessions over a broad period of time (14 months) which can effect internal validity. Because the cyber world is fast paced, with new exploits and techniques discovered to attack old vulnerabilities, even if the target network stays static, the knowledge of the participants can change over time.

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

In order to elicit self-reporting of progress by each participant, and mimic the more realistic team scenario, the task description also included instructions to report any findings to an external "team" via a separate, internet-connected laptop with a chat interface. The participants were also able to use this laptop to perform internet searches. Participant browsing activities were collected throughout the study.

3.2.2 Red Team Population

The Tularosa Study is novel among current human subjects studies due to the inclusion of professional red-team participants. CTF challenges recruit participants from all backgrounds and demographics, from curious hobbyists and professional penetration testing experts, to hackers who specialize in unsanctioned cyber activities. While these unauthorized hackers are the population from which we hope to defend our networks, it is not feasible to assume they will volunteer for a controlled research study. A hacker's greatest asset is a treasure trove of skills and techniques that are unknown or difficult to detect by most targets, and they would not want to use these tricks and tools in a fully-monitored environment.

Additionally, by collecting human subjects data, the Tularosa Study was able to exclude data from participants who self-reported information that would have further disqualified them from the initial recruitment (Example: excluded data from one subject with 0 years of experience).

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

3.3 Environment Design

Teams. Conventional CTF environments tend to be team events which allow for fairly free-form creative exploration of security deficiencies and techniques for maneuver, exploitation, and exfiltration. In many cases individuals can use their own uniquely developed capabilities and tools. This introduces a problem for experimentation. If the research question relies on comparing an experiment treatment versus a control condition (such as adding deception or other defenses to measure effects on performance), then other factors (e.g., unique individual capabilities) have to be controlled. Otherwise, the results may be due to those unique tools, and not the cognitive or behavioral effects of the experimental treatment. Additionally, as mentioned above, participants often engage with each other for help; this introduces several sources of variance that can disrupt the goals of the experiment. Therefore, in addition to restricting the uniqueness of tools (participants could still request publicly available tools for their environment), and by isolating each participant to their own test environment, we intentionally limited the potential for interference from unique tools, external network factors or human interactions.

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

Tasking. The test network was configured to represent an isolated enterprise business network consisting of dozens of servers and desktop computing systems running realistic services and software applications.

Participant tasking was designed to emulate the offensive activities of an unanticipated attacker, rather than a penetration test or formal red team assessment. Participants were provided with a high-level description of their task which included instructions to perform reconnaissance, system exploitation, and data exfiltration. The only difference in task description between conditions was a single sentence statement to subjects who were in the *informed* conditions about the possible presence of deception on the network. The psychological deception component of the experiment was shaped by the mention of possible deception not always corresponding with the true presence of decoys.

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

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

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

Even in studies where deception doesn't play a role, hunting for flags biases the behavior of the participant, which may or may not mimic their actual strategy in a more realistic attack scenario. Do they collect all the easy, low-point flags first? This may just be their strategy to win the game, and not how they would normally perform. In the Tularosa Study, many participants self-reported a desire to remain undetected, despite the single-day time constraint per target network.

Many CTF competitions also give direct feedback to participants on if the flag is valid and its point value. This cycle of feedback helps keep CTF challengers interested and engaged in the competition, but does not represent a realistic cycle of feedback in a real penetration testing scenario, particularly where false documents and decoy systems are a realistic threat to the perceived successes of an adversary.

Furthermore, there are often leaderboards that inform participants in real-time how well they are performing compared to other teams. Leaderboards and the direct competitive aspects with teams competing against each other is another aspect of CTFs that could unnecessarily change a participant's behavior, which provides further evidence of how using CTF to capture human behavior might provide an inaccurate picture of cognitive choices and human performance.

Data Collection.

CTF competitions collect some cyber data that can provide insight into the red-teamers' actions, however most CTF data is limited, such as packet captures and exfiltrated flags, only the latter of which are typically used to evaluate participant success. Conversely, the Tularosa Study collected additional types of human and digital data, including cognitive surveys, physiological data, raw process and log data, and subjects' self-reported cyber strategies and working notes.

The Tularosa Study permitted participants to each attack their own network, without interference from any other subject. As a consequence, 138 different networks were launched throughout the scenario, each utilizing one of three possible configurations (held constant across conditions). Individual host data were not collected from these environments, but the amount of data created by each subject having their own private target was still significant, as there was no log data or network activity overlap between participants.

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

Data Coding. Large, heterogeneous data collections can be difficult to analyze. After experimental design and execution the immense challenge of labeling and analyzing the data begins. In contrast to CTF flag scoring, Tularosa data analysis requires iteratively cleaning and processing each data source.

Raw Data, existing in the same format it was created, is often unusable. *Processed/Extracted Data* curates the raw data files into simplified, standardized formats. *Queried Data* imports the formatted data into tools such R and MySQL, and searches for patterns, statistics, and outliers, then adds contextual metadata to the dataset. Subject matter experts generate *Labeled Data* by applying professional evaluations of the subject's cognitive state and activities. For example, cognitive psychologists review the participants' self-reported activity data for indicators of confusion, frustration, biases, and confidence in success. Finally, *Expert Analysis* requires technical expertise to extract ground-truth data to evaluate the accuracy of a subject's self-reported cyber efforts, such as which attacks were actually successful, and reveals cyber activities that the subject did not report.

4 Trade-Offs and Lessons Learned

While design decisions and trade-offs were discussed throughout the paper, this section summarizes the lessons learned that could be helpful to the community who may wish to conduct related experiments. While many trade-offs discussed are well known within the experimental design community, in our experience, the cyber community does not often perform human subjects experimentation, and so these lessons build up the potential for CTF improvement in this area.

4.1 Realism vs. Repeatability

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

While a real operational network would have provided the most realistic environment, as discussed previously, the study was more interested in controlling across conditions and participants to ensure internal validity. Realism was not the top concern. Based on the initial[2] pilot study [9], it is believed that both Cyber and Psychological Deception may be more effective on an operational network, where the natural messiness of a large network improves the plausibility of the deceptive effects, and provides additional confusion through real complexities and anomalies.

4.2 Managing Red Teamers

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

Professional penetration testers were intentionally selected because they are more likely to closely emulate adversarial hacking activity and demonstrate subject matter expertise than most other subject populations. CTFs do allow for easier recruiting and testing of expert subjects, and the score is easily evaluated with flag collection used as a proxy for expertise. However, other data (e.g., demographics and expertise questionnaires) are typically not collected to determine if any of the data should be excluded from analysis.

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

4.3 Timestamp Correlation

Each participant activity data source in the Tularosa Study must be timestamp correlated with other activities. However, because multiple components in the experiment design were unable to be networked together, correlating timestamps necessitated a workaround.

Timestamped resources different time sources include: 1) The subject's Kali laptop with a NTP Server (off limits as a target) on a sandboxed target network 2) the internet-connected reference machine, connected to NIST internet time servers, 3) the Empatica E4 physiological monitoring devices, not networked during the experiment, and 4) an iPod with a data plan, used by the proctor to log the start and stop times of the cyber task as well as individual breaks and lunches. Because the Tularosa infrastructure did not allow for automating timestamp correlation during the experiment, the iPod also served as our workaround for the timestamp puzzle by using an app called "Timestamp Camera Basic", which recorded

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

the screen clock on each device, once before and after the experiment, and in conjuction with other scripts was used to semi-manually synchronize each data source.

4.4 Cognitive Considerations

There are fundamental differences in the design of experiments that collect human subject performance and cognitive data, and CTFs that collect computer and network data with little probing into the human behind the machine. The Tularosa Study encountered three main cognitive considerations which are critical to the design of current and future studies investigating the human behind adversarial cyber activity.

First, measurement of human behavior and cognition requires IRB approval while measurement of network and computer activity does not. Extra precautions must be taken to conduct HSR studies and protect participants, but they must provide necessary insight and understanding about the motivations, perceptions, emotions, and decisions of the people being studied. CTF and other events focus only on the network activity and successful collection of flags, have less difficulty recruiting expert participants, and do not need to instrument host machines for data collection, thus relieving privacy concerns of the participants.

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

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

5 Conclusions

It should be clear that the current world of CTF is not enough to provide us with the robust experimental design required to provide definitive answers to our most pressing cyber challenges. That said, while CTF events were never intended to answer scientific questions, this does not mean they could not be designed to provide scientific insights. To do so, CTF events should increase focus on the human component in addition to technical activity and the impact on target systems. CTFs may be in a unique position to understand behavior when there are less constraints on performance, even as they impose their own restrictions.

Traditional experimental paradigms such as those used in the Tularosa Study make significant trade-offs to achieve control over experimental conditions, sometimes eschewing ecological validity for the purposes of answering current hypotheses. While there is unlikely to be a large shift in the overall approach to CTF or to cyber-focused human subject research, the community would benefit by a transfer of approaches and mentalities – a good thing for both sides. CTF has remained outside the purview of psychologists and institutional review boards; perhaps it is time for this to change. The support and encouragement of the broader research community could bring experimental methodologies into the CTF event planning and creation process. If every CTF event were also measuring human performance, this would be a game-changing transition for our understanding of human-machine systems and for cyber defense.

The overall design and scientific instrumentation of such events would need to be far more tightly controlled – which could be counter to hacker and CTF culture, where a no-holds-barred approach is often used. For CTFs to generate rigorous experimental results, measurement methodologies and experimental controls would need to be understood by event organizers, which requires more interaction with the scientific community and would thus impose additional effort. Some limitations may need to be imposed and followed by participants as well, which changes the competitive experience. As attempts at CTF experimentation become more common these approaches may become easier to implement. To this end, we clearly need a wealth of experimental paradigms — as well as scientific theory and measurement – from which to draw insights from and to inspire future CTF events.

It is conceivable that we are already deep in the realm of human-machine experimentation but with no controls or measurements. Even the occasional benign software update may appear to emulate poorly run experiments on the human-machine system. When we modify a system or add an additional security control we are affecting both the human's interactions and choices in addition to the machine. When we do so without control or instrumentation of both human and machine we lose an opportunity to learn. It is time that we refine the effort of CTF events by examining the human-machine system through the lens of the scientific method.

References

[1] P. Aggarwal, V. Dutt, and C. Gonzalez. Cyber-Security: Role of Deception in Cyber-Attack Detection. In *Advances in Human Factors in Cybersecurity*, pages 85–96. Springer, July 2016.

[2] T. Benzel. The science of cyber security experimentation: the DETER project. In *Proceedings of the 27th Annual Computer Security Applications Conference*, pages 137–148. ACM, 2011.

[3] N. Buchler, P. Rajivan, L.R. Marusich, L. Lightner, and C. Gonzalez. Sociometrics and observational assess-

ment of teaming and leadership in a cyber security defense competition. *Computers & Security*, 73:114–136, March 2018.

[4] F. Cohen and D. Koike. Feature: Leading Attackers Through Attack Graphs with Deceptions. *Computers & Security*, 22(5):402–411, July 2003.

[5] A. D'Amico and K. Whitley. The Real Work of Computer Network Defense Analysts. In J.R. Goodall, G. Conti, and K. Ma, editors, *VizSEC 2007: Proceedings of the Workshop on Visualization for Computer Security*, Mathematics and Visualization, pages 19–37. Springer, 2008.

[6] E. Dans. You May Never Have Heard Of Bug Bounty Programs... — Forbes.com. https://www.forbes.com/sites/enriquedans/2019/02/11/you-may-never-have-heard-of-bug-bounty-programs, February 2019. [Online; accessed 06-May-2019].

[7] A. Doupé, M. Egele, B. Caillat, G. Stringhini, G. Yakin, A. Zand, L. Cavedon, and G. Vigna. Hit 'em where it hurts: A live security exercise on cyber situational awareness. In *Proceedings of the 27th Annual Computer Security Applications Conference on - ACSAC '11*, page 51. ACM Press, 2011.

[8] J. Dykstra and C. L. Paul. Cyber Operations Stress Survey (COSS): Studying fatigue, frustration, and cognitive workload in cybersecurity operations. In *11th USENIX Workshop on Cyber Security Experimentation and Test (CSET 18)*, 2018.

[9] K.J. Ferguson-Walter, D.S. LaFon, and T.B. Shade. Friend or faux: deception for cyber defense. *Journal of Information Warfare*, 16(2):28–42, 2017.

[10] K.J. Ferguson-Walter, T.B. Shade, A.V. Rogers, E.M. Niedbala, M.C. Trumbo, K. Nauer, K.M. Divis, A.P. Jones, Angela Combs, and R.G. Abbott. The Tularosa Study: An Experimental Design and Implementation to Quantify the Effectiveness of Cyber Deception. In *Proceedings of the 52nd Hawaii International Conference on System Sciences*, 2019.

[11] A. Forget, S. Pearman, J. Thomas, A. Acquisti, N. Christin, L.F. Cranor, S. Egelman, M. Harbach, and R. Telang. Do or do not, there is no try: User engagement may not improve security outcomes. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*, pages 97–111. USENIX Association, 2016.

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

[13] R.S. Gutzwiller, K.J. Ferguson-Walter, S.J. Fugate, and A.V. Rogers. "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, 2018.

[14] R.S. Gutzwiller, S.J. Fugate, B.D. Sawyer, and P.A. Hancock. The human factors of cyber network defense. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 59, pages 322–326. SAGE publications, 2015.

[15] K.E. Heckman, F.J. Stech, R.K. Thomas, B. Schmoker, and A.W. Tsow. *Cyber Denial, Deception and Counter Deception: A Framework for Supporting Active Cyber Defense*. Advances in Information Security. Springer, 2015.

[16] F.L. Levesque and J.M. Fernandez. Computer Security Clinical Trials: Lessons Learned from a 4-month Pilot Study. In *7th Workshop on Cyber Security Experimentation and Test (CSET 14)*. USENIX Association, 2014.

[17] A. Malviya, G.A. Fink, L. Sego, and B. Endicott-Popovsky. Situational Awareness as a Measure of Performance in Cyber Security Collaborative Work. In *2011 Eighth International Conference on Information Technology: New Generations*, pages 937–942. IEEE, April 2011.

[18] S.M. Shope. Effective cyber situation awareness (csa) assessment and training. US Army Final Report #W911NF-13-C-0060, 2013.

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

[20] S. Trent, R.R. Hoffman, D. Merritt, and S. Smith. Modelling the Cognitive Work of Cyber Protection Teams. *The Cyber Defense Review*, page 11, April 2019.

[21] E. Trickel, F. Disperati, E. Gustafson, F. Kalantari, M. Mabey, N. Tiwari, Y. Safaei, A. Doupé, and G. Vigna. Shell we play a game? ctf-as-a-service for security education. In *2017 USENIX Workshop on Advances in Security Education (ASE 17)*. USENIX Association, 2017.

[22] M. Zheng, H. Robbins, Z. Chai, P. Thapa, and T. Moore. Cybersecurity research datasets: taxonomy and empirical analysis. In *11th USENIX Workshop on Cyber Security Experimentation and Test (CSET 18)*, 2018.