Measuring Expertise and Bias in Cyber Security Using Cognitive and Neuroscience Approaches

Daniel Krawczyk, James Bartlett, Murat Kantarcioglu, Kevin Hamlen and Bhavani Thuraisingham
The University of Texas at Dallas
Richardson, Texas, USA

***Abstract*—Toward the ultimate goal of enhancing human performance in cyber security, we attempt to understand the cognitive components of cyber security expertise. Our initial focus is on cyber security attackers – often called "hackers". Our first aim is to develop behavioral measures of accuracy and response time to examine the cognitive processes of pattern-recognition, reasoning and decision-making that underlie the detection and exploitation of security vulnerabilities. Understanding these processes at a cognitive level will lead to theory development addressing questions about how cyber security expertise can be identified, quantified, and trained. In addition to behavioral measures our plan is to conduct a functional magnetic resonance imaging (fMRI) study of neural processing patterns that can differentiate persons with different levels of cyber security expertise. Our second aim is to quantitatively assess the impact of attackers' thinking strategies – conceptualized by psychologists as heuristics and biases – on their susceptibility to defensive techniques (e.g., "decoys," "honeypots"). Honeypots are an established method to lure attackers into exploiting a dummy system containing misleading or false content, distracting their attention from genuinely sensitive information, and consuming their limited time and resources. We use the extensive research and experimentation that we have carried out to study the minds of successful chess players in order to study the minds of hackers with the ultimate goal of enhancing the security of current systems. This paper outlines our approach.**

Keywords*—Cognitive newro science, cyber security, honeypot, chess expert, fMRI, decoys, hackers

I. INTRODUCTION

We are facing an urgent and growing need to better understand expertise in computing domains relevant to cyber security. Cyber threats to software systems, information, and physical infrastructure continue to rise as society becomes increasingly reliant upon a digital platform for transactions, data archiving, and social interactions. Threats to cyber security can be launched from anywhere in the world and may be aimed at government or industry targets enabling widespread breaches of privacy and disruption of human activities. Perhaps most disturbing are recent reports that major physical infrastructure including the electrical grid, water treatment facilities, and food packaging industries are vulnerable to cyber-attacks [1] part of the effort to predict, interpret, and defend against such cyber security risks it will be necessary to better understand the human element that is active in cyber warfare, especially on the offense side [2]. In addition to understanding human behavior related to cyber-attack, it will be increasingly important to identify abilities within individuals that are predictive of cyber skills that would be useful on the defense side.

Critical, yet wide open questions regarding human skill and cyber expertise remain unaddressed. These include fundamental questions such as what types of individuals are effective at cyber skills and how can the relevant abilities be identified and measured? In the current proposal we focus on addressing these questions from a cognitive expertise perspective [3]. Experts are defined as having exceptional skills in a particular domain and their skills can be measured within laboratory tasks relative to novices or initiates [4]. There is a rich history of expertise-related research in behavioral psychology, but such research has been largely limited to the study of expertise in games [5] sports [6], or in a limited set of occupational domains such as nursing (7) and medicine [8]. We direct our initial specific aim at developing and validating measures of expert performance in cyber expertise through the investigation of competency on numerous challenge problems that require definable skills to solve. We plan to quantify and weight these skills based on difficulty to form a battery of expertise scores relevant to cyber problem solving, or hacking.

Experimentally measuring and quantifying expert skills relevant to cyber security is a key step in the identification of talent in this area. As societies continue to transfer more data and infrastructure control into the domain of cyberspace, the needs for skilled individuals at the human operations end will continue to increase. This may be particularly true in cyber security, where constant monitoring and mitigation of new and evolving threats will continue to escalate into the future. The ability to defend against cyber security attacks critically depends upon understanding the skills, motivations, and the abilities of the attacker. It is unlikely that individuals with cyber expertise are simply good general problem solvers. Rather, we must address what specific abilities in particular specialty areas are important in cyber expertise, as has been evident in most other areas of human expertise [3]. We do not yet have a handle on what these specialty areas are or how they may be related to one another. Thus, an important initial goal of the current proposal is to begin to break ground on this research front.

There is a growing need for training in cyber security. In addition to identifying individuals who show aptitudes toward cyber security, or hacking skills, it is also important to engage these individuals in training. There is currently a wide array of computer science coursework and lab experience that is useful toward training individuals on the hardware and software skills in cyber security. What is currently missing is training aimed at understanding the human element in this area. This includes insight into the motivations, tendencies, and competencies of attackers, how to detect them, and how to best defend against subtypes of attackers.

We have conducted some preliminary studies ted measuring expert processing in the domain of chess. While chess is not directly relevant to cyber expertise, the lessons we have learned about measuring expert behavior and brain function are relevant to how we will measure and interpret information gathered in the proposed studies on cyber experts. Hackers, like chess players, must recognize patterns in data, information, and sequences within problems or exploits. This process requires recognition of surface level data (e.g. programming code) and translating it into actions

978-1-4673-6213-9/13/$31.00 ©2013 IEEE

ISI 2013, June 4-7, 2013, Seattle, Washington, USA

that are stored at a deeper level in the expert. Similar operations occur within the game of chess and as such, we describe some of the important developments we have begun to understand through our work, which are relevant to the measure of expertise in the cyber domain.

II. STUDYING CYBER SECURITY THROUGH NEUROSCIENCE

We have two aims motivated by the issues discussed in Section I. They are:

Aim 1: *Characterize and measure the cognitive components of problem solving and decision making in cyber security expertise.*
Aim 2: *Measure the role of cognitive heuristics and biases in cyber security attack under conditions requiring avoidance of traps.*

Below we discuss the aspects surrounding these aims.

Understanding expertise through cognitive neuroscience: In addition to purely behavioral research, we need to employ neuroimaging techniques to further explore cyber expertise. Modern neuroimaging methods such as functional Magnetic Resonance Imaging (fMRI) have altered the field of cognitive psychology by providing greater detail about the biological systems that give rise to memory, reasoning, and decision making. In our prior work (see preliminary studies section) we have investigated questions about the nature of expertise by measuring task-based neural activity (as indexed by Blood Oxygen Level Dependent (BOLD) signal). A key method enabling new inferences about expertise using this technique is the careful design of tasks that reflect a particular type of expertise and contrasting conditions that are similar in many respects, but critically do not reflect an individuals' expertise.

The advantage of brain-based techniques is that they can provide sensitive information about the similarity or differences between experimental conditions complementing traditional behavioral measures. This is particularly useful when combined with accuracy and response times for solving problems or responding to stimuli. An example of the utility of combining behavioral and fMRI studies comes from our prior work on expertise in the game of chess. In these studies we compared visual perception of chess-related configurations to perception of pictures of human faces, an area in which nearly all people have great expertise. It is essential to our social lives that we rapidly and effortlessly identify and recognize faces. Our facility with faces enables us to rapidly recognize and read the intentions of new people we have just met as well. Similarly, an expert chess player can look at a chess game in progress and rapidly identify which side is winning, how close the game is, and where there are specific threats and opportunities within the game. These abilities superficially resemble the speed and accuracy of identification observed within human face recognition.

We experimentally tested for similarities between face and chess perception within chess experts in an interference paradigm (described further in the preliminary studies section). Results indicated that similar levels of interference, a marker of expertise, were observed for face stimuli and chess stimuli [9]. Initially, it would appear that face and chess expert perception share a high degree of similarity. We conducted a second study using fMRI to compare activation within visual cortex toward face and chess stimuli (see Preliminary study 2). We had hypothesized that there should be overlapping activity within the brain indicating that the two domains of expertise share overlapping mechanisms. Contrary to our expectations, there was very little overlap between face and chess perception within experts [10]. This pair of findings highlights the possibility that measurements of expertise at the behavioral level can only provide partial information about the way expertise develops and is organized. This lesson is possibly even more important in more complex domains.

We need to develop a set of fMRI studies designing and implementing experimental conditions that engage processes relevant to cyber expertise and control conditions that do not engage such expert processes. This will enable us to develop new markers of expertise relevant to subtypes of cyber security expertise. Such markers could then be related to scores from the tests we will develop to measure types of cyber expertise and can also be related to individuals' performance on bias-susceptibility (in specific aim 2).

Understanding offense will inform better defense: A critical aspect to understanding cyber expertise is to develop measurements of the skills that are relevant to carrying out attacks. Identifying the abilities of attack-side hackers would improve defensive capabilities through better informing defenders about where to look for attacks, how to detect them, and how to best mitigate against them. A key area of investigation (*Measure the role of cognitive heuristics and biases in cyber security attack under conditions requiring avoidance of traps)* addresses how attackers are vulnerable to detection and delay.

Research on judgment and decision making has identified key biases that operate within individuals. Some such biases tend to become more ingrained and difficult to avoid as individuals gain expertise [11]. Such biases operate across domains and have proved resistant to techniques aimed at removing them . If the operation of cognitive biases could be better identified it would be possible to leverage this knowledge to build better defense-side solutions. Among the robust cognitive biases shown to operate within experts are the following:

- "set effects" (irrationally performing a non-productive sequence of actions which have been successful in the past) [12]
- confirmation bias (the tendency to look for evidence supporting a favored hypothesis) (Wason, 1960; 1968)
- "sunk cost thinking" (the tendency to stick with a specific strategy because much prior effort has been invested in it) [13]
- "representativeness" (acting on a superficial resemblance between a current situation and structurally different situations encountered in the past) [14]
- "availability" (the ease of recalling instances in which a certain action has succeeded in the past) [15].

These biases are likely to operate across domains and present particular challenges when one deals with ambiguous information, a characteristic evident in offense-side cyber security situations in which a hacker must guess, infer, and attempt to gain information about a system before being able to compromise it.
Related to biases, are heuristics which are useful shortcuts or assumptions that typically apply well across a range of situations. Following heuristics leads to high levels of efficiency, as they enable rapid decision making that will prove to be accurate a majority of the time. Heuristics have a down side however, as they can be misapplied under certain exceptional situations. The misapplication of heuristics can lead to extremely poor inferences and logical inconsistencies. Humans tend to invoke heuristics, particularly in cases of high familiarity and facility with information, including cases of high expertise. The over-application or misapplication of heuristics may represent cases in which higher levels of expertise leave an attacker in the position to

make errors. Such errors could be exploited by defenders provided they understood the types of biases frequently used by different subtypes of attackers.

Attracting Attackers with Honeypots: Specific aim 2 directly targets the relationship of heuristics and biases to expertise. The ability to identify the operation of specific biases in different subtypes of attack specialty could enable defense-side cyber security specialists to design traps, decoys, and honeypots to lure and delay attackers. Honeypots/honeynets are methods commonly used to trap attackers using dummy content that looks superficially attractive to an attacker [16]. They attract an attacker by superficially mimicking a real system with genuine sensitive information. Typically honeypots do not contain genuine content and there is not a legitimate reason for an individual to access their services, thus any activity within the honeypot can be considered to be malicious. Once the hacker has been attracted to the honeypot it is in the best interest of the defender to keep them there. The longer an attacker spends interacting with the honeypot, the more time and resources they waste while they are distracted from genuine services, data, and operations. Additionally, the use of a honeypot enables defenders to observe the tactics and operations of an attacker to better understand their goals and methods [17].

Once an attacker has been attracted to a honeypot the goal of keeping him or her there can be facilitated by leveraging knowledge about human bias and heuristic usage. The honeypot may appear initially attractive, but if it is determined to be uninteresting, or worse yet, a trap, an attacker is likely to leave it alone. Thus, constructing compelling honeypots is essential to their effectiveness. Knowledge about human cognitive biases can be adapted to compel attackers to remain trapped in honeypots for greater amounts of time wasting more of their resources, while being observed by defenders. We aim to address how this can be achieved by studying the impact biases including set effects, confirmation bias, representativeness, and availability, which are likely to slow down attacks, thereby exacerbating the problems imposed upon attackers by honeypots.

III. PRELIMINARY STUDIES

Our approach is based on the preliminary studies we have carried out towards a behavioral study of chess experts. We will adapt these methods to study cyber security and how hackers think and behave individually as well as in groups. Below we list these studies.

Preliminary Study 1: Comparing expertise in human perception: A behavioral study of chess experts.

Rationale: Cognitive psychology methods are a powerful tool for the study of expertise. Face processing has several distinctive hallmarks that researchers have attributed either to face-specific mechanisms or to extensive experience distinguishing faces (expertise). Here, we examined the face-processing hallmark of selective attention failure—as indexed by the congruency effect in the composite paradigm—in a domain of extreme expertise: chess. Chess, like hacking, requires accessing elaborate stored sequences from surface level stimuli, then rapidly implementing these sequences on the fly.

Preliminary Study 2: fMRI Activation in Response to Expertise.

Rationale: In Preliminary study 2 we sought to further understand the basis for expertise in chess and face processing. The methods of fMRI enable investigators to ask research questions about the neural basis for cognitive processes. In the case of experts in chess, we investigated whether chess and face processing share similar mechanisms within the brain. If the neural responses to chess and faces are similar, they ought to evoke similar regional activation of perceptually relevant visual cortex. This hypothesis would appear to follow from the findings of Preliminary Study 1. If however, expertise is organized differentially within the brain and forms of expertise differ at a mechanistic level, we may not see substantial similarity in the neural activation associated with chess and face processing.

IV. OUR APPROACH

In support of specific Aim 1, *to characterize and measure the cognitive components of problem solving and decision making in cyber security expertise,* we are planning to develop behavioral tasks that capture key aspects of offensive cyber security. A supporting part of Aim 1 will be achieved through the development and implementation of fMRI tasks that will inform the underlying neural mechanisms of cyber expertise subcategories. Next we will design problem solving tasks that incorporate honeypots, which must be detected and avoided in support of Specific Aim 2: *Measure the role of cognitive heuristics and biases in cyber security attack under conditions requiring avoidance of traps*. The experimental honeypots will be designed around established cognitive biases and evaluating human performance on such problems will inform us about the type of problems that are most likely to distract attackers.

An important part of problem solving is evaluating surface-level information in order to determine its potential for deeper analysis (Gentner, 1983). We will use neuroimaging to better understand the process by which cyber experts evaluate stimuli for relevance that would then receive greater attention in actual cyber situations. The stimuli featured in the fMRI experiment will be drawn from ten of the top 25 software vulnerabilities derived from the 2011 CWE/SANS report (MITRE, 2011). The ten areas we will focus on include SQL injection, OS command injection, buffer overflow, missing authentication, missing authorization, incorrect authorization, Integer overflow, URL redirection, execution with unnecessary privileges, and use of dangerous functions. We predict that experts will be capable of identifying vulnerabilities within code, and screening out uninteresting code. The BOLD activation to these processes will be informative as to whether detecting exploitable vulnerabilities is similar to simply detecting errors in code and whether neural activation varies within detection-sensitive brain regions, or regions of interest (ROIs) along with expertise level (as indexed by the CSE).

An important part of fMRI research is to create control conditions that are highly similar to the conditions of interest. In this experiment, participants will be presented with lines of standard code sequentially, with each line being shown for 2s. Some sequences of code will contain an interesting vulnerability. These will be the trials of interest. Control stimuli will include code that contains a syntax error. These stimuli will look superficially the same as the experimental vulnerability stimuli and will require a similar response detection process, but critically this condition will not be relevant to cyber attack in the same manner. Additionally, we will include a third condition in which the code is uninteresting, or inoccuous (containing no vulnerability or syntax error) requiring only a button press toward the detection of a pre-determined phrase embedded within some screens.

V. SUMMARY AND DIRECTIONS

In this paper we have discussed the need to study the minds of hackers so that we can develop better systems. Specifically we will use s combination of behavioral and neuroscience techniques including fMRI methods to study the minds of the hackers. We have conducted preliminary studies on studying the minds of highly successful chess players and have obtained some promising results. We believe that our methods can be applied to cyber security. We also believe that to combat cyber terrorism, we need to be several steps ahead of the hackers. Therefore developing secure systems without understanding the kinds of hackers will not be the best approach. Our approach combines both: that is studies the minds of hackers, examine the developments in secure systems and applied the lessons learned from the behavioral and neuro-scientific studies to develop better systems.

The planned studies will provide new behavioral and brain measures for several major areas of cyber security expertise. Specifically, these studies will: 1) provide new measures which can be used to identify subtypes of cyber security experts, 2) provide new knowledge that may be useful in understanding the skills, motivations and abilities of offense-side hackers, 3) provide evaluations of the types of cognitive biases that are most active in subtypes of attackers, 4) provide evidence of what types of deception methods may be most promising for the design of honeypots, and 5) provide information on which cyber expertise skills share a common functional neural architecture, and which may further inform the categorization of subtypes of attackers. Additionally, all information collected in this proposal will have broad relevance to training of cyber experts and in identifying the tendencies and skills offense which can be applied toward better methods on defense.

REFERENCES

[1] Cyberware 2011 Westport, CT: Praeger.

[2] Pfleeger, SL & Caputo, D.D. : Leveraging behavioral science to mitigate cyber security risk.

[3] Ericsson K . A. ,; Expertise. Tyler J. Towne. Article first published online: 22 APR 2010 . DOI: 10.1002/wcs.47. Copyright © 2010 John Wiley & Sons, Ltd.

[4] Chi, M.T.H. (2006). Methods to assess the representations of experts' and novices' Knowledge . In K.A. Ericsson, N. Charness, P. Feltovich, & R. Hoffman (Eds.), Cambridge Handbook of Expertise and Expert Performance. (Pp. 167-184), Cambridge University Press.

[5] Gobet, F., & Charness, N. (2006). Chess and games. Cambridge handbook on expertise and expert performance (pp. 523-538). Cambridge, MA: Cambridge University Press.

[6] Duffy, L. J., Baluch, B., & Ericsson, K. A. (2004). Dart performance as a function of facets of practice amongst professional and amateur men and women players. International Journal of Sport Psychology, 35, 232-245.

[7] Ericsson, K. A. (2007). An expert-performance perspective on medical expertise: Study superior clinical performance rather than experienced clinicians! Medical Education, 41, 1124-30.

[8] Harley, E. M., Pope, W. B., Villablanca, P., et al. (2009). Engagement of fusiform cortex and disengagement of lateral occipital cortex in the acquisition of radiological expertise. Cerebral Cortex, 19: 2746–54.

[9] Boggan, A. L., Bartlett, J. C., & Krawczyk, D. C. (2012). Chess Masters show a hallmark of face processing for chess. Journal of Experimental Psychology: General, 141, 37-42.

[10] Krawczyk, D. C., McClelland, M. M., & Donovan C. (2011). A hierarchy for relational reasoning in the human prefrontal cortex. Cortex, 47, 588-597.

[11] Bilaliü M, McLeod P, Gobet F (2010) The mechanism of the Einstellung (Set) effect: a pervasive source of cognitive bias. Curr Direct Psychol Sci 19:111–115.

[12] Luchins, A. S. (1942). Mechanization in problem solving. Psychological Monographs, 54, No. 248.

[13] Arkes, Hal; Blumer, Catherine (1985). "The Psychology of Sunk Cost". Organizational Behavior and Human Decision Process 35: 124–140.

[14] Tversky, A. & Kahneman, D. (1974). Judgments and Uncertainty: Heuristics and Biases. Science, New Series, 185 (4157), 1124-1131.

[15] Tversky, A. & Kahneman, D. (1973). Availability: A Heuristic for Judging Frequency and Probability. Cognitive Psychology, 5 (2), 677-695.

[16] Cohen F. (2011) Use of Deception Techniques: Honeypots and Decoys University of New Haven 182. The Handbook of Information Security, Volume III Threats, Vulnerabilities, Prevention, Detection and Management.

[17] Cheswick, B. “An evening with Berferd in which a Cracker is Lured, Endured, and Studied”,

http://www.tracking-hackers.com/papers/berferd.pdf , 1991.