Attackers have some prior beliefs: Understanding cognitive factors of confirmation bias on adversarial decisions

Harsh Katakwar[1][0000-0002-7476-8878], Cleotilde Gonzalez[2][0000-0002-6244-2918], and Varun Dutt[1][0000-0002-2151-8314]

[1] Applied Cognitive Science Laboratory, Indian Institute of Technology Mandi, Mandi, India
[2] Dynamic Decision-making Laboratory, Carnegie Mellon University, Pittsburgh, USA
katakwarharsh@gmail.com, coty@cmu.edu, varun@iitmandi.ac.in

Abstract. Cyberattacks are hazardous, and honeypot deception has been shown to be successful in combating them. Due to involvement of multiple factors in cyber situation, the adversary is likely to suffer from various cognitive biases. One of the many cognitive biases that affect adversarial decisions in cyberspace is confirmation bias. However, little is known about the cognitive mechanisms that drive confirmation bias in adversarial decision-making. To test for confirmation bias, one hundred and twenty participants were recruited via a crowdsourcing website and were randomly assigned to one of two between-subjects conditions in a deception-based cybersecurity simulation. Results revealed the presence of confirmation bias in adversarial decisions. Thereafter, a cognitive Instance-based Learning model was built involving recency, frequency, and cognitive noise to understand the reasons behind the reliance on confirmation bias. Results revealed that participants showed reliance on recent events and high cognitive noise in their decisions. We highlight the implications of our findings for cyber decisions in the presence of deception in the real world.

Keywords: Deception, Attack, Honeypot, Probe, Confirmation bias, Instance-based Learning Theory (IBLT), Cognitive models.

1 Introduction

Cyberattacks are the adversary's deliberate attempts to disable computer systems. In 2021, there was a dramatic increase of 105% in ransomware attacks among the different cyberattacks [1]. This rise in attack activity encourages the research community to develop adaptive solutions to enable in creation of a secure cyberspace. There are certain security solutions available to assist in countering cyberattacks, such as intrusion detection systems (IDSs), filtering strategies, and firewalls [2-5]. IDSs generate alarms on the detection of any suspicious activity [2,3]. IDSs are reliable; however, they may also generate false alerts resulting in financial losses [5]. Filtering solutions help remove unwanted content while still ensuring secure access. This strategy may result in bounded nonrational network agents reaching consensus[5]. In general, such an agreement could help detect cyberattacks before they become a hazard to cyberinfrastructure [5].

Firewalls also monitor network traffic and restrict incoming and outgoing packets as per established security policies [6]. However, these firewalls are not intelligent enough to distinguish between genuine authorized access and a malicious attempt[6]. Overall, these solutions might not be able to help in combating newer cyberattacks.

Cyber deception has been proven to be an effective way of combating cyberattacks [4]. The aim of cyber deception is to incorporate human factors into consideration in cyber circumstances and improve security tools to minimize cyber-attacks [4]. Cyber deception has been used with the help of honeypots, which appear as real systems [7]. This technique has been found to be effective in the detection and response to cyberattacks [8-10]. Recent research in the behavioral cybersecurity domain has focused on technological factors that drive adversarial decisions in cybersecurity. Some of them include network topology, timing and amount of deception, network size, and honeypot proportions [10, 11]. Aggarwal et al. [10] investigated the impact of timing and amount of deception on adversarial decisions, revealing that late deception increased the proportion of honeypot attacks when compared to early deception. Similarly, Katakwar et al. [10] investigated the impact of various network sizes on adversarial decisions in cyberspace. Besides these elements, adversarial decision-making is likely to be influenced by a set of predetermined adversarial strategies. These strategies might be convincing and efficient in a simplified view of the real world, but when applied in a real situation, the adversary may become prone to certain cognitive biases [12]. As a result, it is essential to investigate the role of cognitive biases in the adversary's decision-making.

Researchers have discovered that adversarial decisions in cybersecurity scenarios may suffer from cognitive biases such as confirmation bias, anchoring bias, sunk-cost fallacy, irrational escalation, loss aversion, and others [12,13]. Among these biases, confirmation bias, the tendency to select options that support one's own belief from a pool of information, has affected a majority of adversarial decisions [14]. However, research is yet to investigate how cognitive elements such as memory decay and cognitive noise aids in suffering from cognitive bias in adversarial decisions. One way of understanding the cognitive elements in dynamic environments is by building cognitive models based upon Instance-based Learning Theory (IBLT) [15-17]. Previously, IBLT-based cognitive models were able to explain the reason for adversarial decisions in various cyber situations [11,18].

The purpose of this research is to understand the cognitive factors that drive adversaries towards confirmation bias in cyber situations by computational cognitive modelling. First, we create a deception-based security game to study the presence of confirmation bias and build a cognitive model based upon IBLT that could account for adversaries' decisions. Hence, we developed a deception-based security game as a simulation environment and replicated real-world cybersecurity circumstances. We used a sequence of deception trials followed by non-deception trials and vice versa (non-deception trials followed by deception trials) in the game to check if the adversaries revealed confirmation bias. In the non-deception trial, the response from the network remains true, whereas in the deception trial the response from network is opposite to that of actual. If an adversary faces a non-deception trial before experiencing deception trial, the adversary develops a belief that the network's response is true. Thus, when the adversary transitions from a non-deception trial to a deception trial, the adversary will rely on confirmation bias and most likely target those webservers whose response is “regular webserver”, but in reality, are “honeypot webserver”. Similarly, if the transition happens from a deception trial to a non-deception trial, then the adversary builds a belief of a deceptive network, which

communicates a “honeypot webserver” as a regular webserver”. Thus, when the adversary transitions to a non-deception trial with her deceptive belief, the adversary would be driven by confirmation bias and attack honeypot webservers. Hence, we expect that participants are likely to suffer from confirmation bias whenever they transition from a deception to a non-deception experience or vice-versa. First, we provide a brief overview of a deception-based game. Thereafter, we discuss an experiment, where we investigate whether adversarial decisions suffer from confirmation bias. Following that, we present the results of the experiment. Furthermore, we describe the results of cognitive models based upon IBLT, which attempt to account for human decisions in the experiment. Finally, we talk about the real-world implications of the developed cognitive models.

2 DECEPTION GAME

Deception Game (DG) is a web-based game in which an adversary and the network play against each other [19]. It is expressed as DG (n, k, γ), where *n* denotes the number of webservers in the network, *k* represents the number of honeypots, *γ* depicts the number of probes the adversary makes before attacking the network. The DG had two kinds of webservers, regular and honeypot. The regular webserver is a real system in the network, whereas the honeypot is a fake system in the network. A trial in DG has two phases, a probe phase followed by an attack phase. In the probe phase, the adversary can probe some of these webservers or may not probe any of them. Probing in DG means clicking the button which denotes the webserver in the game's interface. In the probe phase, the adversary receives the information from the network based upon deception and non-deception trial. In deception trial, the network response is opposite to that of the actual. However, in the non-deception trial, the network’s response is same as that of actual. In the attack phase, the adversary may attack one of the webservers present in the network. Attacking a webserver in the DG means clicking on the button denoting a webserver in the network. After the adversary has made his decisions in the probe and attack stages, the adversary moves on to the feedback screen, where the rewards for the probe and attack stages of the DG are revealed. Table 1 below shows the payoffs for players for the probe and attack actions in the DG.

Table 1. Hacker's payoff during probe and attack stages.

Stage	Hacker's action	Payoff
Probe	Regular webserver	+5
	Honeypot webserver	-5
	No probe	0
Attack	Regular webserver	+10
	Honeypot webserver	-10
	No attack	0

3 METHODS

3.1 Experiment Design

A total of 120 participants were recruited from crowd-sourcing website called Amazon Mechanical Turk for the study, and they were randomly assigned to one of the two between-

subject conditions (each condition with 60 people). Each condition possessed a different sequence of deception and non-deception trials. Across both the conditions, the DG was configured as DG (20, 10, 10), where there were 20 webservers in the network, 10 of them were honeypot webservers, and the adversary got the option to probe 10 times before attacking the network. One of the two conditions had the first five trials as deception followed by non-deception (referred to as the D-N condition). In contrast, the other condition had the first five trials as non-deception followed by deception (referred to as the N-D condition). Table 2 shows the pattern of deception and non-deception trials across the D-N and N-D conditions.

Table 2. The sequence of deception and non-deception across the trials in D-N and N-D conditions.

Trial	D-N	N-D
1-5	Deception	Non-deception
6-10	Non-deception	Deception
11-15	Deception	Non-deception
16-20	Non-deception	Deception
21-25	Deception	Non-deception
26-29	Non-deception	Deception

In the experiment, the participant could take one of the three attack decisions (regular attack, honeypot attack, and not attack) in the probe and attack phases. In both the conditions and across all the transition trials (i.e., change in the trial from deception trial to non-deception trial or non-deception trial to deception trial), whenever the participant attacked the honeypot webserver, it showed the presence of confirmation bias in the adversarial decisions. In the D-N condition, the deception trial makes a belief in the adversary about the incorrect information from the network. Hence, when the transition happens from deception trial to non-deception trial, the adversary with the deceptive belief in the non-deception trial gets trapped by attacking a honeypot webserver. Since, in the non-deception trial, the response from the network is true. Similarly, in the N-D condition, the adversary forms a belief of a true response from the network. Once the transition happens from the non-deception trial to the deception trial, the adversary with the true belief gets trapped by attacking those honeypot webservers, which acts as regular webserver in the probe phase. Across the transition trials in both the conditions, participants who attacked honeypot webservers were found to have confirmation bias in their decisions. Hence, we labeled participants attacking honeypot webservers as "1" (confirmation bias) and those attacking regular webservers or not attacking the network as "0". The average of 1s and 0s gave the proportion of participants suffering from confirmation bias.

3.2 Stimuli

Figure 1 presents the probe stage's interface in the DG. It shows how participants were briefed about the task and the availability of various types of webservers in the network. Thereafter, the participant performing as an adversary may probe some of the webservers in the network or may not probe any. Thereafter, the participant proceeds to the attack phase. During the attack phase, the participant chooses to attack one of the network's webservers (see Figure 2). After that, the participant is awarded for the decisions made during the probe and attack stages (see Figure 3).

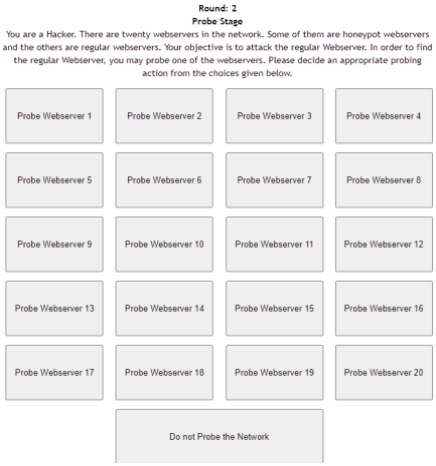

Fig. 1. Initial screen in the probe stage of DG.

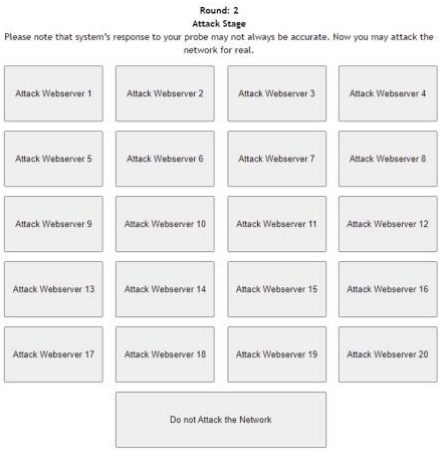

Fig. 2. Attack stage in the DG.

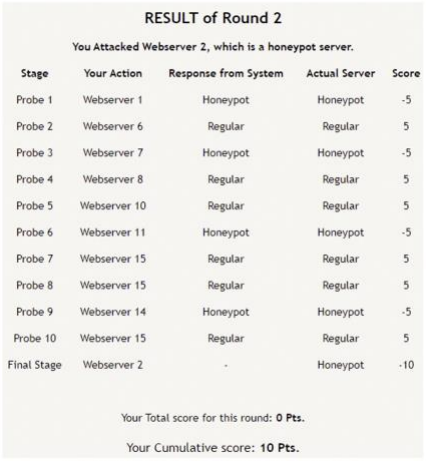

Fig. 3. Result of the complete round in the DG.

3.3 Participants

The study was conducted after the approval from the Ethics Committee of the Indian Institute of Technology Mandi (IITM/DST-ICPS/VD/251). Participants were recruited via a crowdsourcing website called Amazon Mechanical Turk [20]. 73% were males, while the remaining 27% were females. The age of the participants varied from 19 to 54 years (Mean = 31 years; Standard deviation = 6 years). 94% of the participants possessed a college degree, while the remaining participants did not have a college degree. On successful completion of the study, participants were renumerated with INR 50 (USD 0.67). After the completion of study, one among the top-three scorers of the study was rewarded with gift voucher of INR 500 (USD 6.69).

3.4 Procedure

Participants performing as adversaries were instructed about their goal in DG and were informed about the remuneration for their participation in the study. In addition, they were also instructed about the presence of deception trials in the DG. Also, they were asked to increase their score over the trials in the DG in probe and attack stages. Once the study was over, the participants were renumerated and thanked for their participation.

3.5 Results

We performed a single sample t-test to compare the proportion of participants with confirmation bias across the different transition trials of D-N and N-D conditions. In D-N condition, the proportion of participants suffering from confirmation bias was significantly higher compared to the unbiased proportion (33%) for the following trials (see Figure 4): trial 6 (t (59) = 2.612, p < .05), trial 11 (t (59) = 2.612, p < .05), trial 16 (t (59) = 3.947, p < .001), trial 21 (t (59) = 2.357, p < .05), and trial 26 (t (59) = 2.869, p < .01). Similarly, in the N-D condition, the proportion of participants suffering from confirmation bias was significantly higher compared to the unbiased proportion (33%) for these trials (see Figure 5): trial 6 (t (59) = 3.397, p < .001), trial 11 (t (59) = 3.131, p < .01), trial 16 (t (59) = 3.947, p < .001), and trial 26 (t (59) = 3.131, p < .01). However, for the trial 21 in the N-D condition, there was no significant difference between proportion of participants with confirmation bias and the unbiased proportion (0.33) (t (59) = 1.853, p = .069).

4 IBL MODEL

IBLT is a theory of decisions from experience in complex scenarios [15-18, 21]. As per prior research, cognitive models based upon IBLT have accounted for human decisions in different dynamic situations. An instance is an IBL model (based upon IBLT) consists of a situation, decision, and utility triplet. The situation in the instance corresponds to the current situation. The decision in the instance denotes the action taken in the current situation, and the utility refers to the outcome received in the current situation. When a decision is to be made, instances for the option are recalled from memory. These instances are then blended. The blended value for the option j in trial t is denoted as:

$$V_{j,t} = \sum_{i=1}^{n} p_{i,j,t} x_{i,j,t}$$

where, $p_{i,j,t}$ is the probability of retrieval of the instance i for the option j in trial t, which is proportional to the instance's activation; $x_{i,j,t}$ denotes the utility value, for instance, i for the option j in the t_{th} trial of the experiment. The above equation calculates the blended value for each option which is the weighted product of observed outcomes and the probability of recalling the instances containing those outcomes. In each trial of the experiment, the model chooses the option with the maximum blended value. The activation of an instance is computed using the following equation:

$$A_i = \ln\left(\sum_{t_{p,i}\in\{1,\ldots,t-1\}} \left(t - t_{p,i}\right)^{-d}\right) + \sigma * \ln\left(\frac{1-\gamma_{i,t}}{\gamma_{i,t}}\right)$$

In the above equation, d and σ are the free parameters known as memory decay and cognitive noise. respectively; t is the current trial; $t_{p,i}$ denotes the previous trial in which the outcome with instance i occurred, and $\gamma_{i,t}$ depicts the random number chosen from the uniform distribution between 0 and 1.

The decay parameter d is used to account for reliance on recent information. The higher the value of the d parameter, the greater the dependence on recent information and the faster the memory degradation. The parameter accounts for the variance in instance activation from one trial to the next. The model's instance structure was made up of the webserver decision, ground truth, and utility value associated with it. In the IBL model for an adversary, the webserver decision in the instance denoted the webserver number probed or attacked by the adversary. The ground truth represented the type of webserver the opponent investigated or attacked, i.e., regular and honeypot. The third attribute of the instance structure in the model was utility value referred to the reward linked with the adversary's decision and the ground truth for the decision (see Table 1).

4.1 Calibration of model parameters

We obtained the calibrated value of d and σ using human data for different experimental conditions. In the IBL model, we aimed to minimize the average Mean Square Deviations (MSD) of decisions between humans and models across the transition trials. The average MSD is the aggregate of MSDs for different attack decisions (regular webserver attack, honeypot webserver attack, or not attack) in transition trials. The MSD for attack decision is defined as,

$$MSD = \frac{1}{5}\sum_{t\in\{6,11,16,21,26\}} (model_t - human_t)^2$$

where t denotes the transition trials (where switch happens from deception to non-deception or non-deception to deception). The $model_t$ and $human_t$ depict model and human decisions in the transition trial t for different attack decisions, respectively. If the value of the average MSD was small, then better is the model's fit to human data. The values of d and σ parameters for both model participants were optimized using the Genetic Algorithm, an optimization algorithm. This optimization algorithm utilizes bio-inspired operators such as mutation, crossover, and selection to create better solutions for optimization problems. The crossover and mutation rates in the genetic algorithm (GA) were set at 80% and 1%, respectively. The GA stopped when there was no change in MSD for a successive of 20 generations.

4.2 Model Results

Table 3 shows the average MSD and calibrated values of the free parameters (i.e., d and σ) in both conditions. In both conditions, the cognitive noise and decay were quite high, showing variability and recency in decisions. In addition, the model was able to predict human decisions across both conditions very accurately.

Table 3. Model parameters and average MSD value across the two conditions.

Condition	Memory decay (d)	Cognitive noise (σ)	Average MSD
D-N	2.71	7.49	0.002
N-D	3.10	7.61	0.002

Figure 4 denotes the proportion of participants with confirmation bias across the different transition trials in the D-N condition in DG.

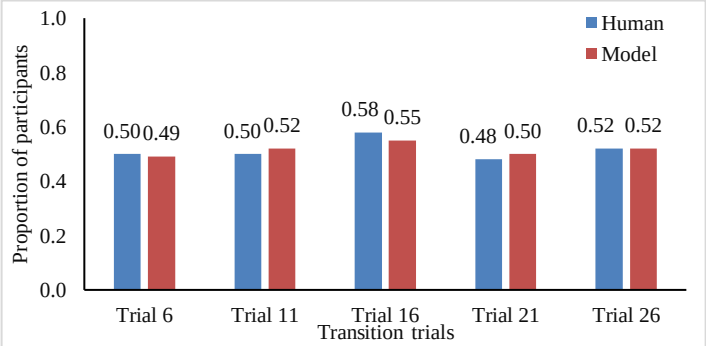

Fig. 4. Proportion of participants with confirmation bias confirmation bias across different transition trials in the D-N condition.

Figure 5 denotes the proportion of participants with confirmation bias across different transition trials of the N-D condition in the DG.

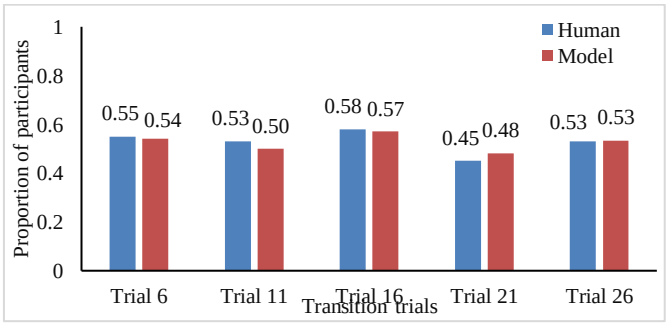

Fig. 5. Proportion of participants with confirmation bias across transition trials in N-D condition of DG.

5 DISCUSSION AND CONCLUSION

Cyberattacks are threatening and they require intelligent security solutions for countering them. Deception via honeypots is likely a viable solution to counter cyberattacks. Cyber deception has helped in improving cyber defense, where the adversary has been found to be vulnerable to different cognitive biases.

First, we discovered that when participants transitioned from a deception to a non-deception trial, they exhibited confirmation bias by attacking more honeypot webservers in the network.

This is because, while switching from a non-deception trial to a deception trial, the adversary in earlier trials acquired true information while probing a webserver in the network. This gave the adversary the notion of no deception in the network. As a result, in a deception trial, the adversary made decisions based on the same belief and attacked the honeypot webserver. However, during the transition from a deception trial to a non-deception trial, the adversary on probing webservers received false information, leading to a deceptive belief. Hence, the adversary decided to make inverted decisions, and with this belief, the adversary exhibits confirmation bias and attacked more webservers which resemble to be regular webservers but were actually honeypot webservers. Furthermore, participants showed high cognitive noise and memory decay in both conditions. The high cognitive noise indicates that participants' decisions varied from trial to trial, while the high memory decay indicates that they did rely on recent events. One possible explanation for these results is that the adversary received inconsistent replies from the network as a result of the transitions of deception and non-deception trials in a short span of time. This baffled the adversary, prompting him to make arbitrary decisions based on pre-existing assumptions. Furthermore, adversaries relied upon the most recent experiences to make decisions exhibiting confirmation bias.

Our findings are limited because they are based on a laboratory experiment. Thus, some of the findings should be interpreted in that light. Conditions in real-world cybersecurity scenarios may differ from those in a lab-based experiment. However, participants had no prior knowledge of what trials possessed deception in DG. They were also unfamiliar with the mapping of honeypots and regular webservers in DG, which was done randomly in each trial. Some of these characteristics may resonate with real-world cybersecurity situations. One application of the developed cognitive model is that it might be useful for penetration testing to find exploitable vulnerabilities. In addition, the developed model can assist in generating predictions on the proportion of confirmation bias in other unexplored sequences of deception and non-deception trials. In future studies, we intend to study how the length of deception and non-deception trials affects adversary decisions in cybersecurity circumstances. In addition, we would like to investigate the presence of other cognitive biases in adversarial decisions in complicated and dynamic cybersecurity circumstances. Following that, we intend to create cognitive models to aid in our understanding of the various cognitive elements that affect adversarial choices in cybersecurity situations.

6 REFERENCES

1. A. Taylor, “Ransomware cyberattacks surged in 2021 according to a new report | Fortune,” *Fortune*, 2022. https://fortune.com/2022/02/17/ransomware-attacks-surge-2021-report/ (accessed Feb. 20, 2022).
2. P. Aggarwal and V. Dutt, “The role of information about opponent’s actions and intrusion-detection alerts on cyber decisions in cyber security games,” *Cyber Secur. A Peer-Reviewed J.*, vol. 3, no. 4, pp. 363–378, Jun. 2020, Accessed: Jan. 18, 2022. [Online]. Available: https://hstalks.com/article/5815/the-role-of-information-about-opponents-actions-an/.
3. K. Scarfone and P. Mell, “Guide to intrusion detection and prevention systems (idps),” *NIST Spec. Publ.*, vol. 800, no. 2007, p. 94, 2007.
4. N. C. Rowe and E. J. Custy, “Deception in cyber attacks,” in *Cyber Warfare and Cyber Terrorism*, IGI Global, 2007, pp. 91–96.

5. Y. Shang, “False Positive and False Negative Effects on Network Attacks,” *J. Stat. Phys.*, vol. 170, no. 1, pp. 141–164, Jan. 2018, doi: 10.1007/S10955-017-1923-7.
6. “ICT in Education - Victoria L. Tinio - Google Books.” https://books.google.co.in/books/about/ICT_in_Education.html?id=eZyZOQAACAAJ&redir_esc=y (accessed Oct. 11, 2021).
7. M. H. Almeshekah and E. H. Spafford, “Cyber security deception,” in *Cyber Deception: Building the Scientific Foundation*, Springer International Publishing, 2016, pp. 23–50.
8. P. Aggarwal, C. Gonzalez, and V. Dutt, “Cyber-security: Role of deception in cyber-attack detection,” in *Advances in Intelligent Systems and Computing*, 2016, vol. 501, pp. 85–96, doi: 10.1007/978-3-319-41932-9_8.
9. H. Katakwar, P. Aggarwal, Z. Maqbool, and V. Dutt, “Influence of Network Size on Adversarial Decisions in a Deception Game Involving Honeypots,” *Front. Psychol.*, vol. 11, p. 2385, Sep. 2020, doi: 10.3389/FPSYG.2020.535803/BIBTEX.
10. P. Aggarwal, C. Gonzalez, and V. Dutt, “Looking from the hacker’s perspective: Role of deceptive strategies in cyber security,” *2016 Int. Conf. Cyber Situational Awareness, Data Anal. Assessment, CyberSA 2016*, Jul. 2016, doi: 10.1109/CYBERSA.2016.7503288.
11. P. Aggarwal, C. Gonzalez, and V. Dutt, “Modeling the effects of amount and timing of deception in simulated network scenarios,” in *2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA)*, Jun. 2017, pp. 1–7, doi: 10.1109/CyberSA.2017.8073405.
12. R. Gutzwiller, K. Ferguson-Walter, S. Fugate, and A. Rogers, “‘Oh, Look, A Butterfly!’ A Framework For Distracting Attackers To Improve Cyber Defense:,” *https://doi.org/10.1177/1541931218621063*, vol. 1, pp. 272–276, Sep. 2018, doi: 10.1177/1541931218621063.
13. K. J. Ferguson-Walter, D. S. Lafon, and T. B. Shade, “Friend or Faux: Deception for Cyber Defense,” *J. Inf. Warf.*, vol. 16, no. 2, pp. 28–42, 2017, doi: 10.2307/26502755.
14. C. K. Johnson, R. S. Gutzwiller, J. Gervais, and K. J. Ferguson-Walter, “Decision-Making Biases and Cyber Attackers,” in *2021 36th IEEE/ACM International Conference on Automated Software Engineering Workshops (ASEW)*, 2021, pp. 140–144, doi: 10.1109/ASEW52652.2021.00038.
15. C. Gonzalez, J. F. Lerch, and C. Lebiere, “Instance-based learning in dynamic decision making,” *Cogn. Sci.*, vol. 27, no. 4, pp. 591–635, 2003, doi: https://doi.org/10.1016/S0364-0213(03)00031-4.
16. C. Gonzalez and V. Dutt, “Refuting data aggregation arguments and how the instance-based learning model stands criticism: A reply to Hills and Hertwig (2012),” *Psychol. Rev.*, vol. 119, no. 4, pp. 893–898, 2012, doi: 10.1037/A0029445.
17. C. Gonzalez and V. Dutt, “Instance-Based Learning: Integrating Sampling and Repeated Decisions From Experience,” *Psychol. Rev.*, vol. 118, no. 4, pp. 523–551, Oct. 2011, doi: 10.1037/A0024558.
18. V. Dutt, Y. S. Ahn, and C. Gonzalez, “Cyber situation awareness: Modeling detection of cyber attacks with instance-based learning theory,” *Hum. Factors*, vol. 55, no. 3, pp. 605–618, Jun. 2013, doi: 10.1177/0018720812464045.
19. N. Garg and D. Grosu, “Deception in honeynets: A game-theoretic analysis,” *Proc. 2007 IEEE Work. Inf. Assur. IAW*, pp. 107–113, 2007, doi: 10.1109/IAW.2007.381921.
20. W. Mason and S. Suri, “Conducting behavioral research on Amazon’s Mechanical Turk,” *Behav. Res. Methods*, vol. 44, no. 1, pp. 1–23, Mar. 2012, doi: 10.3758/s13428-011-0124-6.
21. V. Dutt and C. Gonzalez, “Making Instance-based Learning Theory usable and understandable: The Instance-based Learning Tool,” *Comput. Human Behav.*, vol. 28, no. 4, pp. 1227–1240, Jul. 2012, doi: 10.1016/J.CHB.2012.02.006.
22. M. Mohsin, “Security policy management for a cooperative firewall,” *Doctoral Dissertation, Aalto University*, 2018.