Human Factors in Cybersecurity, Vol. 168, 2025, 57–69
https://doi.org/10.54941/ahfe1006143

Analysis of Large Language and Instance-Based Learning Models in Mimicking Human Cyber-Attack Strategies in HackIT Simulator

Shubham Sharma[1], Shubham Thakur[1], Megha Sharma[1], Ranik Goyal[1], Shashank Uttrani[2], Harsh Katakwar[3], Kuldeep Singh[4], Palvi Aggarwal[4], and Varun Dutt[1]

[1]Applied Cognitive Science Lab, Indian Institute of Technology Mandi, HP 175005, India
[2]Syneos Health India, Gurugram, Haryana - 122002, India
[3]Samsung Research Institute, Sector 135, Noida, UP 201304, India
[4]Department of Computer Science, College of Engineering, University of Texas at El Paso, El Paso, TX 79968, United States

ABSTRACT

As cyber threats grow more advanced, there is a urgent need for models that can simulate and anticipate adversarial behavior. While honeypots have been widely used in deception-based cybersecurity, less is known about how cognitive and AI-based models replicate human decision-making in realistic attack scenarios. This study uses the Team HackIT simulator to evaluate how well Instance-Based Learning (IBL) and GPT-4o, a Large Language Model (LLM), mimic human cyber-attack strategies across varied network topologies and sizes. The decay (d) and noise (σ) parameters included in the IBL model were calibrated from the ACT-R defaults (d $=$ 0.5, $\sigma = 0.25$), and ranged from 0.1 to 3. Calibrated IBL parameters (decay and noise) improved predictive accuracy, especially in smaller networks (MSE $=$ 0.060 for honeypots, 0.002 for real systems). With temperature (0.5, 1, 1.5) and top-k sampling (2, 3, 4) GPT-4o also aligned well with human behavior in 40-node networks (MSE $\leq$ 1.000) but performed less accurately in 500-node configurations (MSE up to 25.000). These findings provide insights into adversarial modeling and suggest that combining cognitive and AI-based approaches can enhance deception-aware cyber-defense strategies.

Keywords: Behavioral cybersecurity, Instance-based learning (IBL), Large language models (LLMs), Hackit simulator, Human decision-making, Behavior modeling

INTRODUCTION

Artificial intelligence (AI) and machine learning have advanced quickly, creating new opportunities to comprehend and mimic human decision-making in a variety of fields (Bhatt et al., 2023). Understanding adversarial tactics in cybersecurity contexts requires cognitive behavior, which is defined as the capacity to receive, analyze, remember, and use information in decision-making (Anderson, 1990). The field still lacks thorough research on how cognitive models like Instance-Based Learning (IBL) and large language models (LLMs) like GPT-4o can mimic human cyber-attack strategies, despite the notable advancements made in machine learning applications for anomaly detection and attack prevention (Wazid et al., 2022). In the field of cybersecurity, cognitive elements including action variety, frequency, and recency often impact adversarial decision-making (Gonzalez et al., 2003). Although cognitive models—especially those derived from IBL theory—have been used for dynamic decision-making tasks (Gonzalez & Dutt, 2011; Dutt & Gonzalez, 2012), little is known about how they may be used in cybersecurity scenarios including deceit. Nevertheless, there hasn't been a comprehensive investigation of the predicted accuracy of IBL models in detecting honeypots and normal systems, as well as how this accuracy varies with network size and topology. To close this gap and get a better understanding of adversarial behavior across different network topologies and sizes, this research calibrates the IBL model.

Concurrent with cognitive modeling, LLMs such as GPT-4o have emerged with strong capacities for processing large datasets, and imitating human behavior (Brown et al., 2020). For cybersecurity activities like automated penetration testing and vulnerability identification, LLMs have shown potential (Zhang et al., 2023). The potential of LLMs for autonomous red teaming was recently emphasized by Itonin et al. (2024), but their performance was not examined across various network topologies or parameter changes. In order to close these gaps, this study compares the performance of GPT-4o, IBL, and humans while assaulting networks with different topologies and sizes using the Team HackIT simulation program. In the dynamic environment provided by Team HackIT, players must navigate misleading honeypots intended to deceive attackers while identifying and exploiting vulnerabilities (Aggarwal & Dutt, 2020). We want to determine the unique advantages and disadvantages of these models in simulating the actions of human attackers by adjusting the network topology (Bus vs. Hybrid) and size (40 vs. 500 nodes), as well as by adjusting model parameters like decay and noise for IBL and temperature and top-k sampling for GPT-4o.

This study investigates the interplay between human and model judgments under various settings, in contrast to earlier research that concentrated on either algorithmic performance or human decisions alone. The study compares two approaches Instance-Based Learning (IBL) and GPT-4o using the Team HackIT simulation platform, across varying network topologies (Bus vs. Hybrid) and sizes (40 vs. 500). Both human and model behaviors were assessed across three dependent variables: total systems exploited, total honeypots exploited, and total real systems exploited. The primary objectives were to compare the fidelity of IBL and GPT-4o in reproducing human decision patterns, and examine the influence of network complexity on model accuracy. We speculated that the IBL model rooted in human cognitive principles would more accurately mirror human decision patterns, especially in simpler environments, while GPT-4o, with its exploratory tendencies, might exhibit greater variability in performance. To test this, we compared the predictive accuracy of each model against human behavior. Our analysis revealed some compelling differences between the two approaches, which offer valuable insights into their respective strengths and limitations in modeling adversarial behavior.

Background

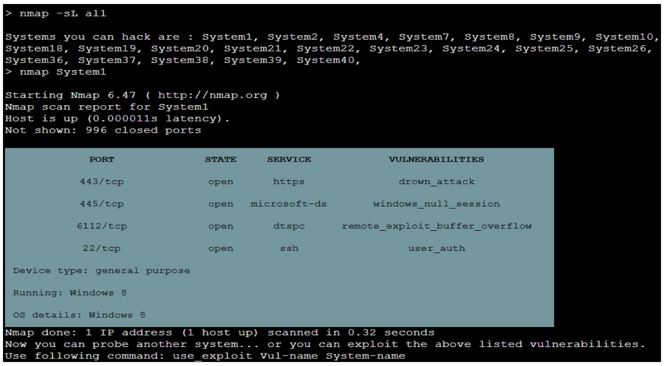
According to Aggarwal and Dutt (2020), researchers have created a program called HackIT that uses behavioral game theory ideas to correctly simulate a cyberattack scenario. Network size, honeypot dimensions, and deception strategies are some of the factors that greatly influence adversarial decisions. For example, in a deception-based game using honeypots, Katakwar et al. (2020) investigated the effects of varying network sizes on adversarial decision-making. In addition, Katakwar et al. (2022) created a computer cognitive model based on Instance-Based Learning Theory (IBLT) to mimic human behavior in deception scenarios. When calibrated with human data, the model showed that participants relied too much on frequency and recency.

Previous studies have shown the importance of deception in cyber protection. Rowe and Custy (2007), for instance, examined a number of misleading strategies to deceive attackers, highlighting the cognitive effects of deception. In a similar vein, Almeshekah and Spafford (2016) examined how misleading methods affect adversarial decision-making psychologically and demonstrated how strategically placed honeypots may deceive attackers and stall harmful activity. The predictive capabilities of IBL models in situations with different network topologies and sizes are still not well understood, despite these discoveries. The purpose of this research is to close this gap by using IBL models in various network topologies and adjusting their parameters to match observed human behavior. Greater decay and lower noise values in IBL models for smaller networks indicate that participants make fewer variable choices and depend more on recent experiences. Participants target ordinary web servers more often in medium-sized networks, perhaps because they are more used to previous successful operations. On the other hand, with bigger networks, participants investigate a wider range of systems due to increased noise and decreased decay (Katakwar et al., 2022). The current study investigates the relationship between these cognitive inclinations and network size and structure.

At the same time, LLMs such as GPT-4o show promise as cybersecurity tools. Because they were trained on large datasets, these models perform very well in real-time attack simulations, contextual reasoning, and vulnerability assessment (Brown et al., 2020; Radford et al., 2019). LLMs' effectiveness in offensive cyberattack simulations is still mostly unknown, despite the fact that they have been well researched in anomaly detection (Sommer & Paxson, 2010; Zhang et al., 2023). LLMs can be used for autonomous red teaming, according to recent research by Itonin et al. (2024); however, further research is needed to understand how LLM choices interact with human tactics in dynamic, deception-based situations. In this work, we fill this research gap by comparing the IBL and GPT-4o decision-making processes in a variety of cybersecurity situations using the Team HackIT tool. We examine how well the GPT-4o and IBL models identify honeypots and target actual systems by varying the temperature and Top-K parameters, as well as decay and noise.

Team HackIT

A web-based testbed called Team HackIT was created to support multiplayer cyber deception tests in a variety of cybersecurity situations, such as different network topologies and sizes. It offers a command-line interface for easy interaction and an interactive, game-like interface that mimics real-time cyberattacks (see Figure 1). Real systems and honeypots are the two main system kinds that the platform differentiates between. Honeypots are malicious systems designed to seem like authentic ones in order to trick attackers. Their main goals are to watch how attackers behave, evaluate their strategies, and protect legitimate systems by rerouting harmful activity. In the Team HackIT setting, participants work together in pairs. One person creates a virtual room to start the session, and the other person joins using a special room ID. After entering, the participant's first assignment is to list all of the systems that are accessible in order to evaluate the network. They then probe these systems to find important data, like open ports, active services, and vulnerabilities that may be exploited.



```
> nmap -sL all

Systems you can hack are : System1, System2, System4, System7, System8, System9, System10,
System18, System19, System20, System21, System22, System23, System24, System25, System26,
System36, System37, System38, System39, System40,
> nmap System1

Starting Nmap 6.47 ( http://nmap.org )
Nmap scan report for System1
Host is up (0.000011s latency).
Not shown: 996 closed ports

          PORT              STATE     SERVICE              VULNERABILITIES

         443/tcp             open      https                 drown_attack

         445/tcp             open   microsoft-ds         windows_null_session

        6112/tcp             open      dtspc        remote_exploit_buffer_overflow

          22/tcp             open       ssh                    user_auth

 Device type: general purpose

 Running: Windows 8

 OS details: Windows 8
Nmap done: 1 IP address (1 host up) scanned in 0.32 seconds
Now you can probe another system... or you can exploit the above listed vulnerabilities.
Use following command: use_exploit Vul-name System-name
```

Figure 1: Demonstration of "Nmap" command being used in the probe phase.

A data table detailing the network's current state and possible access sites is shown to participants during the probing phase (see Figure 1). Participants use this information to choose whether to carry out an attack, focusing on systems with known vulnerabilities. A successful exploitation simulates the exfiltration of sensitive data by giving participants access to the target system and enabling them to obtain a specific file (pin.txt). Team HackIT records every action taken by participants throughout the experiment, including system probes, attack plans, and successful exploits.

METHODS

Participants

A total of 84 participants willingly took part in this cybersecurity experiment; all students from the Indian Institute of Technology Mandi, participated in this study. Before starting the study, consent was obtained from all the participants. The total duration of the experiment was 10 minutes. The participant's demographics included both boys and girls, with ages ranging from 18 to 27 years (mean = 20.88, standard deviation = 1.67 years). Eighty-five percent of the participants were male, and fifteen percent were female. Out of those who took part, about 72.5% were bachelor's students, and 27.5% were master's students. The study was approved by the Indian Institute of Technology Mandi's Ethical Committee.

Experimental Design

Participants were randomly assigned one out of the four between-subjects conditions. These conditions were: Bus topology with small (40 systems) network size, Hybrid topology with small (40 systems) network size, Bus topology with large (500 systems) network size, Hybrid topology with large (500 systems) network size. The deception via honeypot was present in all the conditions, consisting of 50% honeypots and 50% real systems. The purpose of this design was to assess how participants' attack plans and decision-making were affected by the presence of deception. There were a total of 10 minutes given to the participants in order to attack as many systems as they could to maximize their reward. Participants were aware of the deception present but they were not aware of where actually the deception is present. Before entering the experimental tasks, participants were provided instructions regarding the objectives of the attack and the nature of the systems involved.

Procedure

Two bus configuration testbeds and two hybrid testbeds were set up for two network sizes (40 and 500). The participants received essential game instructions which included necessary commands for each stage. The questionnaire checked whether participants had fully understood the instructions before moving forward. The objective is to steal confidential file "pin.txt" within the duration of 10 minutes by breaking into as many real systems as possible. The procedure was divided into two stages: the Attack Phase and the Probe Phase. Participants in the Probe Phase scanned available machines using the "Nmap" command to find out about open ports, services that were operating, and potential vulnerabilities. After selecting the appropriate vulnerability to exploit, participants had to use the command "use_exploit" to break into the system in the following step. After completing an exploitation successfully users could locate "pin.txt" through the "ls" command in the system directory. The participants used "scp" to move the file from the target system to their own system. The participants received their final score after completing the session.

IBL Model

Instance-Based Learning (IBL) theory (Gonzalez et al., 2003; Gonzalez and Dutt, 2011; Dutt & Gonzalez, 2012; Lejarraga et al., 2012; Dutt et al., 2013), a cognitive framework that mimics human decision-making by depending on prior experiences (or instances), is the foundation for the machine learning model utilized in this study. The IBL theory postulates that our decision-making is influenced by past experiences which are there in memory. When something new happens, the model finds previous examples that are similar to it and compares them using predetermined similarity measures. The decisions made by human participants serve as the dataset for the model to learn from. In the context of HackIT, each scenario, such as the system number they are in, the running services or network protocols (FTP, TCP, HTTP, DNS, etc.) and the vulnerabilities (sql_injection, drown_attack, remote_auth, etc.) represent an instance. In order to create a judgment that resembles human decision-making patterns as nearly as possible, the IBL model retrieves and compares cases that are there in the memory.

IBL Model Calibration: Decay and Noise

Two critical factors during IBL model calibration included Decay and Noise. The model decreases its reliance on past instances through time or through changes in situation relevance. The model prioritizes recent relevant instances through its decay implementation, which determines decision-making priorities. Memory reliability based on recent experiences grows stronger when the decay (d) parameter value rises while the rate of memory degradation accelerates. The unpredictable elements of human decision-making can be attributed to noise. Multiple players in this game cause their choices to differ from typical patterns. Our model seeks to replicate human unpredictability through controlled noise implementation. The model receives different decay and noise values for calibration before it gathers decisions from each condition. There were a total of six values of both decay and noise in between 0 and 3, which makes a total of 36 combinations. Also, the model was calibrated on ACT-R default values of decay (d) and noise (σ) (d $= 0.5$, $\sigma = 0.25$).

GPT-4o Model

The Transformer architecture forms the base structure for advanced large language model (LLM) GPT-4o that depends on self-attention to make accurate token predictions from historical context (Vaswani et al., 2017). GPT-4o is well-suited for complicated reasoning in multi-step adversarial simulations because it can handle long text sequences and multi-modal inputs (text and images) within a context window of 25,000 tokens (Brown et al., 2020; Radford et al., 2019). GPT-4o learns to predict tokens from extensive datasets, such as academic literature, chats, and instructional texts, during pre-training (Raffel et al., 2020). Human assessors improve the model's decision-making to better suit human goals when reinforcement learning (RL) is used with human input (Ouyang et al., 2022). Because of its versatility, it can assess the plans of attackers and help create defenses that are more robust (Stiennon et al., 2020).

Top-K and Temperature Tuning

The way that GPT-4o makes judgments is greatly impacted by the temperature and Top-K factors. Top-K determines the number of likely tokens the model considers while generating answers. Lower Top-K values restrict the model to high-probability alternatives, resulting in more predictable outcomes, whereas higher Top-K values promote a more thorough examination of potential responses (Holtzman et al., 2020). Fan et al. (2018) claim that temperature controls randomness; higher values favor riskier, experimental decision-making, whereas lower values increase accuracy and predictability. This study systematically varied both variables to investigate the effects of Top-K and temperature on GPT-4o's ability to distinguish between real systems and honeypots. Lower temperature and higher Top-K improved precision and decision focus, enabling GPT-4o to mimic the decision-making processes of human attackers, but higher temperature and lower Top-K promoted exploration

RESULTS

IBL Results

Table 1 shows the results of the IBL model with ACT-R parameters and with calibrated parameters across different conditions. Here we can compare the mean squared error (MSE) obtained from both the parameters. We can observe that the MSE obtained is the least with calibrated parameters compared to ACT-R parameters. In the case of the Real systems exploited we can see that the MSE is the same for both the parameters across all the conditions, but if we look for honeypots exploited and total attacks, the MSE is lesser with calibrated parameters. In the case of Bus 500, the MSE is similar for both the parameters in honeypots exploited and real systems exploited.

Table 1: Different model parameters and MSE for the IBL model across different conditions.

Configuration	Model	d	σ	MSE Honeypot	MSE Real	MSE Total Attacks
Hybrid 40	With ACT-R Parameters	0.50	0.25	0.130	0.002	1.000
	With Calibrated Parameters	3.00	0.10	0.102	0.002	0.846
Bus 40	With ACT-R Parameters	0.50	0.25	0.058	0.002	0.410
	With Calibrated Parameters	3.00	0.68	0.058	0.002	0.314
Hybrid 500	With ACT-R Parameters	0.50	0.25	0.194	0.014	1.440
	With Calibrated Parameters	2.42	0.10	0.160	0.014	1.346
Bus 500	With ACT-R Parameters	0.50	0.25	0.102	0.006	1.440
	With Calibrated Parameters	2.42	1.84	0.102	0.006	1.440

Hybrid 40

In the "Hybrid 40" condition, the IBL model with ACT-R parameters ($d = 0.5, \sigma = 0.25$) resulted in an MSE of 0.1296 for the honeypot systems, 0.0016 for real systems, and 1.0 for total exploit. The MSE values marginally improved when calibrated parameters ($d = 3, \sigma = 0.1$) were used. They were 0.1024 for honeypot systems, 0.0016 for genuine systems, and 0.8464 for all systems exploited. Figure 2(a) displays the performance of the honeypot exploit for this condition, whereas Figures 1(b) and (c) display the real system exploits and the total number of exploits.

Bus 40

The MSE with ACT-R parameters for the "Bus 40" condition was low for real systems (0.0016), honeypot systems (0.0576), and total exploits (0.4096). The MSE for the total exploits decreased from 0.4096 to 0.3136 with the calibrated parameters (d = 3, σ = 0.68), however the MSEs for the honeypot and real systems stayed the same. Figure 2(a) compares the honeypot exploited accuracy, while Figure 2(b) and Figure 2(c) present the real system exploited and total systems exploited, respectively, for the Bus 40 condition.

Hybrid 500

In the "Hybrid 500" condition, with the original ACT-R parameters, the MSE was higher for honeypot (0.1936), real (0.0144), and exploit-based systems (1.44). The MSE values for honeypots exploited (0.16) and total systems exploited (1.3456) decreased with calibrated parameters (d = 2.42, σ = 0.1). However, for the real system exploited, the MSE stayed constant at 0.0144. The results for the honeypot, real system, and total systems exploited in Hybrid 500 are shown in Figure 2(a), Figure 2(b), and Figure 2(c), respectively.

Bus 500

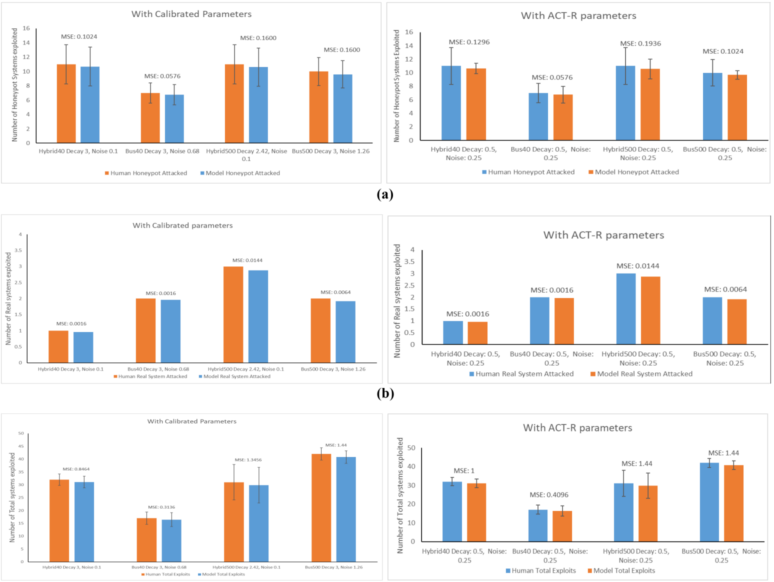
The model with ACT-R parameters got an MSE of 0.1024 for honeypot systems exploited, 0.0064 for real systems exploited, and 1.44 for total systems exploited in the "Bus 500" condition. The MSE values were not significantly altered by the calibrated parameters (d = 2.42, σ = 1.84), as the MSEs for the honeypots, real system, and total systems exploited stayed the same. The honeypot exploited results for this condition appear in Figure 2(a), and real system exploited and total systems exploited are shown in Figures 2(b) and 2(c).

GPT 4o Results

Comparative Accuracy of GPT-4o and Human Participants on Total Systems

Both GPT-4o and human participants exploited 45 systems in the "Bus 40 (Temperature = 1.5, Top-K = 4)" configuration, producing an MSE of 0.000, therefore proving perfect alignment. Additionally, in the "Hybrid 40 (Temperature = 1, Top-K = 4)" configuration, both groups exploited 25 systems once more, producing an MSE of 0.000, hence demonstrating GPT-4o's ability to replicate human decision-making in smaller, ordered network environments. GPT-4o exploited 32 systems in the "Bus 500 (Temperature = 1.5, Top-K = 3)" configuration, while human participants exploited 29 systems, producing an MSE of 9.000, therefore demonstrating GPT-4o's enhanced performance under greater exploration settings for larger network size. GPT-4o exploited 25 systems in the "Hybrid 500

(Temperature = 1.5, Top-K = 4)” configuration, surpassing the 19 systems exploited by human participants, thereby producing an MSE of 36.000, implying a larger inclination for over-exploitation in complicated hybrid topologies (see Figure 3(a)).

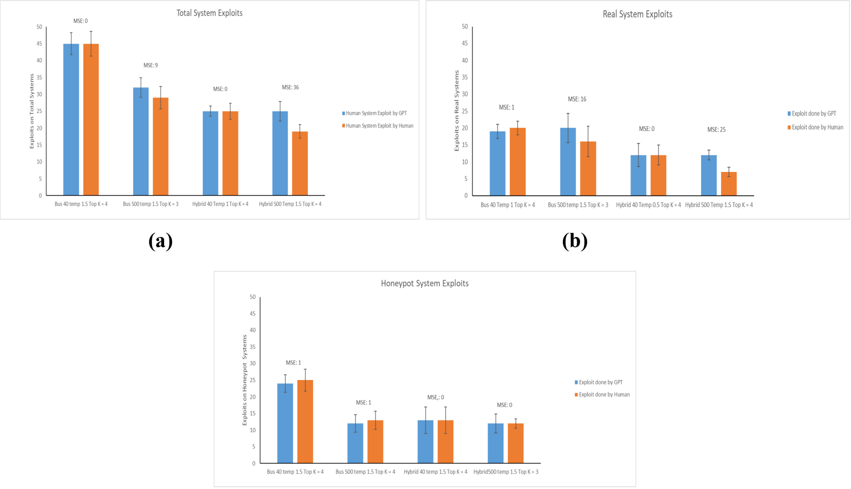


(a)

(b)

(c)

Figure 2: (a) Number of honeypots exploited by human and model. (b) Number of real systems exploited by human and model. (c) Number of total systems exploited by human and model.



(a)

(b)

(c)

Figure 3: (a) Comparison of total system exploited by Gpt-4o and human participants. (b) Comparison of real system exploited by Gpt-4o and human participants. (c) Comparison of honeypot system exploited by Gpt-4o and human participants.

Comparative Accuracy of GPT-4o and Human Participants on Real Systems

GPT-4o exploited 19 real systems in the “Bus 40 (Temperature = 1, Top-K = 4)” configuration, roughly paralleling the 20 systems exploited by humans, resulting in a mean squared error of 1.000. In the “Hybrid 40 (Temperature = 0.5, Top-K = 4)” configuration, both GPT-4o and human participants exploited 12 systems, resulting in an MSE of 0.000, thus demonstrating GPT-4o’s ability to replicate human strategies under stable configurations. GPT-4o surpassed human participants in the “Bus 500 (Temperature = 1.5, Top-K = 3)” configuration by exploiting 20 systems instead of 16, therefore producing a mean squared error (MSE) of 16.000. In the “Hybrid 500 (Temperature = 1.5, Top-K = 4)” configuration, GPT-4o exploited 12 systems, whereas humans exploited 7, giving a mean squared error (MSE) of 25.000, which suggests that GPT-4o adopted a more robust exploitation strategy in complicated hybrid environments (see Figure 3(b)).

Comparative Accuracy of GPT-4o and Human Participants on Honeypot Systems

In the “Bus 40 (Temperature = 1.5, Top-K = 4)” configuration, GPT-4o exploited 24 honeypot systems, nearly equivalent to the 25 systems exploited by humans, with a mean squared error of 1.000. Similarly, in the “Bus 500 (Temperature = 1.5, Top-K = 4)” configuration, GPT-4o exploited 12 honeypot systems, whereas humans exploited 13, therefore producing an MSE of 1.000, showing GPT-4o’s ability to replicate human decision-making in honeypot configuration. Both GPT-4o and human participants exploited 13 systems in the “Hybrid 40 (Temperature = 1.5, Top-K = 4)” configuration, producing an MSE of 0.000 and indicating great alignment. Similarly, in the “Hybrid 500 (Temperature = 1.5, Top-K = 3)” configuration, both groups exploited 12 honeypot systems once more, producing an MSE of 0.000, therefore attesting to GPT-4o’s flexibility in challenging attack scenarios (see Figure 3(c)).

Statistical Results

The mean squared error (MSE) was analyzed to compare the performance of the IBL and GPT models. The IBL group demonstrated a lower average MSE ($M = 0.73$, $SD = 0.58$) compared to the GPT group ($M = 2.37$, $SD = 3.25$) likely reason for this could be the exploratory nature of GPT models, indicating a potential advantage of the IBL model in predictive accuracy. One-way ANOVA test revealed a statistically significant difference in MSE between the IBL and GPT groups with $F(1, 35) = 4.43$, $p = .043$, partial $\eta^2 = .112$, which indicates a moderate group-level difference in prediction error.

DISCUSSION AND CONCLUSION

The significance of deception strategies like honeypots in defending vital systems from malevolent actors has long been highlighted by cybersecurity studies (Rowe & Custy, 2007). While more recent research has added computational cognitive models to simulate adversarial decision-making processes, earlier works examined how misleading strategies may misdirect attackers and postpone intrusions (Katakwar et al., 2020). In order to replicate human reactions in networks with different honeypot proportions, Katakwar, Aggarwal, and Dutt (2023) created a model based on Instance-Based Learning Theory (IBLT). By applying IBL and GPT-4o models to cybersecurity situations with different network sizes and topologies, this research builds on previous efforts by examining how well they function in bus and hybrid settings.

The outcomes of the Team HackIT simulations are consistent with the ideas of IBLT, which holds that recollection of past experiences specifically, frequency and recency effects influences decision-making. Due to the simpler, linear layout of smaller bus networks, which made it easier to remember previous accomplishments, participants showed a higher preference for attacking genuine systems. The hierarchical complexity of hybrid network setups, on the other hand, increased unpredictability and encouraged participants to engage in more exploratory activities. When adjusted for decay (d) and noise (σ), the IBL model more closely matched human performance under all circumstances, especially when it came to distinguishing between honeypots and real systems.

Interesting similarities between GPT-4o's performance and human decision-making processes were found. GPT-4o demonstrated a low mean squared error (MSE) in smaller, more organized networks, closely resembling the attack patterns of human players. The exploratory inclinations shown in human participants when confronted with uncertainty are mirrored in the model's capacity to modify its tactics via the use of Top-K sampling and temperature modifications. In line with Pu and Faltings (2011), lower temperatures produced more focused, exploitative tactics, while higher temperatures encouraged more exploratory activity. But in hybrid network settings, GPT-4o showed limits, which is in line with Zhang et al. (2023), who pointed out that LLMs have difficulties in contexts with more structural complexity.

These results have important ramifications for the field of human factors. Designing more user-friendly and flexible cybersecurity training resources may be aided by an understanding of how cognitive biases such as frequency and recency affect the choices made by cybercriminals. The potential for LLMs to function as dynamic elements in cyber-defense systems, automating red teaming procedures and mimicking a wide variety of hostile strategies, is shown by the comparative performance of GPT-4o. This kind of integration of cognitive and AI-based models promotes a more thorough, human-centered approach to the design of cybersecurity systems.

This research only examined bus and hybrid topologies, with a fixed honeypot fraction of 50%, despite these encouraging findings. Future research might examine how different honeypot distributions affect human decision-making and look at other network topologies like star or ring topologies. Furthermore, longitudinal research that looks at attacker behavior over lengthy stretches of time may provide further light on the tactical and cognitive adjustments attackers make in response to deception-based countermeasures. By providing an alternative viewpoint on adversarial decision-making, these models open the door to cybersecurity solutions that are more flexible, effective, and focused on people.

REFERENCES

Aggarwal, P., & Dutt, V. (2020). The role of information about opponent's actions and intrusion-detection alerts on cyber decisions in cyber security games. *A Peer-Reviewed Journal, 3*(4).

Almeshekah, M. H., & Spafford, E. H. (2016). Cyber deception: Using decoys to engage attackers and misinform. *Computers & Security, 67, 266–284*

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Lawrence Erlbaum Associates, Hillsdale, NJ.

Berrueta, L., Rios, R., Laorden, C., & Bringas, P. G. (2019). Anomaly detection in network traffic using unsupervised learning techniques. *IEEE Transactions on Network and Service Management, 16(1), 314–327.*

Bhatt, S., Sethi, T., Tasgaonkar, R., et al. (2023). Machine learning for cognitive behavioral analysis: Datasets, methods, paradigms, and research directions. *Brain Informatics, 10*(18).

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems, 33*.

Dutt, V., & Gonzalez, C. (2012). Making instance-based learning theory usable and understandable: The instance-based learning tool. *Computers in Human Behavior, 28*(4), 1227–1240.

Dutt, V., Gonzalez, C., & Lebiere, C. (2013). Cyber situation awareness: Modeling detection of cyber attacks with instance-based learning theory. *Human Factors, 55*(3), 605–618.

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science, 27*(4).

Gonzalez, C., Lerch, J. F., & Lebiere, C. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review, 118*(4), 523–551.

Holtzman, A., Buys, J., Du, L., et al. (2020). The curious case of neural text degeneration. *Proceedings of the International Conference on Learning Representations (ICLR), 2020.*

Itonin, A., Caldwell, S., & Richardson, B. (2024). Leveraging large language models for autonomous red teaming in simulating advanced ransomware attacks.

Katakwar, H., Aggarwal, P., Maqbool, Z., & Dutt, V. (2020). Influence of network size on adversarial decisions in a deception game involving honeypots. Frontiers in Psychology, 11, 527826.

Katakwar, H., Aggarwal, P., et al. (2022). Influence of probing action costs on adversarial decision-making in a deception game. *Springer, 2022.*

Katakwar, H., Singh, R., Mehta, S., et al. (2023). Modeling the effects of different honeypot proportions in a deception-based security game. *Human Factors in Cybersecurity, 91*(91).

Lejarraga, T., Dutt, V., & Gonzalez, C. (2012). Instance-based learning: A general model of repeated binary choice. *Journal of Behavioral Decision Making, 25*(2), 143–153.

Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*.

Raffel, C., Shazeer, N., Roberts, A., et al. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rowe, N. C., & Custy, E. J. (2007). Deception in cyber attacks. In *Cyber Warfare and Cyber Terrorism*.

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *In Proceedings of the 2010 IEEE Symposium on Security and Privacy (pp. 305–316).*

Stiennon, N., Ouyang, L., Wu, J., et al. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems, 33*, 3008–3021.

Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Wazid, M., Das, A., Chamola, V., & Park, J. H. (2022). Uniting cyber security and machine learning: Advantages, challenges, and future research. *ICT Express, 8*(1).

Zhang, Y., Li, H., Wang, J., et al. (2023). Exploring AI-driven cyber attack simulations with large language models. *Proceedings of the International Conference on Cybersecurity (ICC), 2023*.