ReSCIND Performer HSR Dataset Cover Sheet

ReSCIND

Dataset Details

Dataset Title:	Guarding Against Malicious Biased Threats (GAMBiT) Experiment 3		
Dataset Citation:	GAMBiT Experiment 3 Dataset [2508.20963] Guarding Against Malicious Biased Threats (GAMBiT) Experiments: Revealing Cognitive Bias in Human-Subjects Red-Team Cyber Range Operations Beltz, B., Doty, J., Fonken, Y., Gurney, N., Israelsen, B., Lau, N., Marsella, S., Thomas, R., Trent, S., Wu, P. and Yang, Y.T., 2025. Guarding Against Malicious Biased Threats (GAMBiT) Experiments: Revealing Cognitive Bias in Human-Subjects Red-Team Cyber Range Operations. *arXiv preprint arXiv:2508.20963*.		
Data Format:	Available on S3 bucket, zip files	Data Size:	2.8 TB
Dates & Duration:	February 1, 2025 – March 26, 2025 Two 8-hour days per participant	Time Zone:	EST/EDT
How to access dataset:	IEEE Dataport link: Guarding Against Malicious Biased Threats (GAMBiT) Experiment 3 | IEEE DataPort		
Point of Contact for data questions:	Quanyan Zhu, NYU quanyan.zhu@nyu.edu		

Description of Scenario

Experiment Objectives

This experiment observed how skilled attackers behaved during a cyber attack scenario. The goal was to collect detailed behavioral data to help researchers develop new ways to classify attacker actions and decision-making patterns.

Experiment Description

Over two days, 22 red team participants were given access to a simulated enterprise network (a "cyber range") and instructed to conduct self-paced cyberattacks. In the beginning, participants received operational instructions and credentials for initial network access. From there, they pursued realistic objectives—such as identifying sensitive systems and exfiltrating valuable data—at their own pace.

The network contained embedded "triggers"—carefully designed cues like misleading credentials or deceptive file names—used to study how attackers respond to uncertainty. Participants also completed periodic surveys and maintained written notes to document their reasoning and tactical choices throughout the exercise.

Cognitive Vulnerabilities targeted

Bias	Descriptive Phrase	Indicators
Loss Aversion	Emotional weighting of outcomes.	Biased behaviors include prioritizing preserving what one already has instead of aiming for greater gains (endowment), such as focusing on found credentials
Base Rate Neglect	Statistical misjudgments	Biased behaviors include ignoring general statistics or probabilities such as access of valid admin credentials. Rational behaviors include testing account privileges.
Availability	Bias in memory and recall	Biased behaviors include making decisions based on what is easiest to remember or most recent in one's mind and overestimating the importance of rare but memorable events, such as recalling and attempting publicized Apache 2.4.50 vulnerability.
Confirmation Bias	Selective, belief driven data processing	Biased behaviors include seeking information that supports one's existing beliefs and dismissing evidence that does not align with one's initial assumptions, such as viewing failed attempts of found malformed SSH keys as "almost working" rather than as actual failures.
Sunk Cost	Effort-related persistence	Biased behaviors include continuing with a failing plan because one has already invested effort or resources, such as persisting in using commands despite repeated session termination, focusing on recovering their perceived process.

Experimental Results

Analysis of cyber data, skills tests, self-reports, and operational notes found that higher-skilled individuals made more progress in cyber attacks.

Cyber Environment

Experiment 3 used the SimSpace Cyber Force Platform to design and implement the GAMBiT cyber range, which simulated an enterprise business information system. This cyber environment comprised approximately 40 virtual devices organized into subnetworks, incorporating routers, switches, and user traffic commonly found in operational networks. Image 1 illustrates the most recent network topology of the GAMBiT cyber range. Each participant operated within a designated cyber range and initiated all challenges using a virtual machine running Kali Linux.

Each participant had identical IP addresses within their assigned range, ensuring consistency across individual environments. The network included restricted subnets designated for managing the environment during the engagement, which were classified as no-strike targets. The experiment environment consisted of one range with 25 triggers.

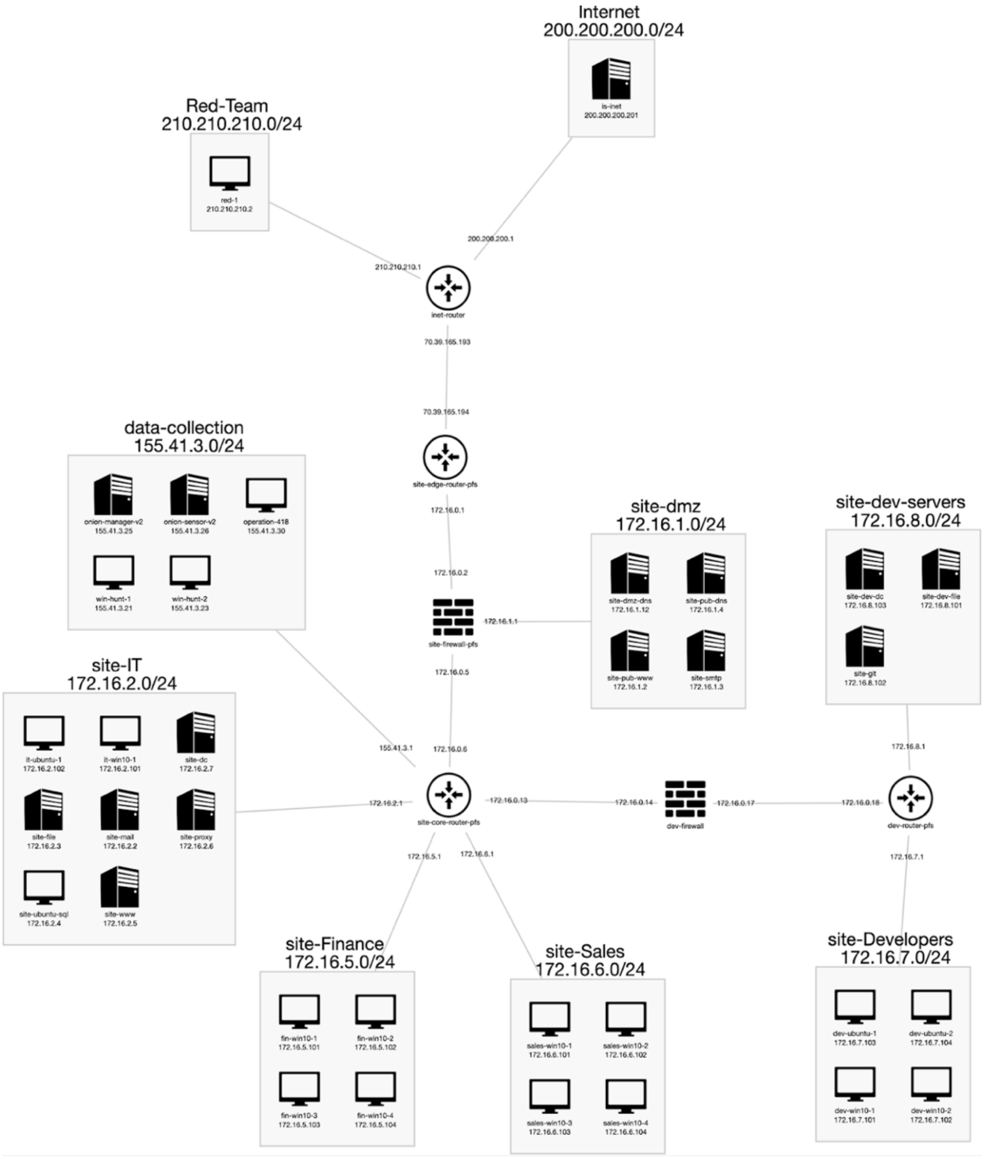

Image 1: Network topology for GAMBiT cyber range for Experiment 3, for reference only.

Data

Data Sources

Primary Data Sources

These data were collected directly from the cyber range experiment environment.

Category	Data Source	Examples of Select Data Features
Self Reports/Background Data *	Screening Questionnaire	Years of experience, type of cyber experience, team size, preferred OS, length of campaign.
	Demographics Questionnaire	Age, gender, native language, education level
Self Reports/Psychometric Data *	Cognitive Reflection Test (CRT)	3 multiple-choice items to assess the ability to override an intuitive but incorrect response and engage in more deliberate, analytical thinking. (Frederick, S. 2005)
	Big Five Inventory extra-short form (BFI-2)	15 items to indicate personality (Soto & John, 2017)
	General Risk Propensity Scale (GRiPS)	8 items on tendency to risky behaviors (Zhang et al., 2019)
	Adult Decision-Making Competence Scale (A-DMC)	Resistance to Framing Positive (7 items) & Negative (7 items) and Resistance to Sunk Cost (10 items) (Bruine de Bruin et al., 2007)
Self Reports/ Questionnaires	Applied Techniques	Hourly Stage reports (X.1-X.3): intended/applied MITRE ATT&CK techniques
	Reasoning and Affect Changes	Hourly Stage reports (X.1-X.3): 5-point Likert scale items on reasoning (6 items) and mood (5 items) changes
	OPNOTES	CherryTree file with Operation Notes from participant.
Network Data	PCAP	Timestamps, source & destination packets & protocols, payloads
Network Data Kali Host Data	NIDS - Suricata	Monitors network traffic for suspicious activities based on predefined rules
	Keylog	Keylogger where each line records an individual keystroke. Particularly useful as it collects text that participants copy to their clipboard.

	Terminal histories	Bash and zsh histories, timestamps, order of commands

** Provide a citation for each psychometric assessment in the References section below.*

Derivative Data Sets

These datasets were created from aggregating, analyzing, curating, and labeling the source data.

Data Source	Examples of Select Data Features
Clean Log	Result of running a post-processing script on Admin VM to remove certain keystrokes for better readability.
GAMBiT Exp 1+2 Screening, Demographics, Psychometrics, Command Logs.xlsx **README GAMBiT Exp 1+2 Screening, Demographics, Psychometrics, Command Logs.pdf**	This excel file includes preprocessed data organized by participant, including screening questions and skills test scores, demographics, psychometrics, and processed data from command line logging. This file contains information for participants in all three experiments. More information in the DATA readme file.

Research

Hypotheses

The GAMBiT Experiment 3 dataset was used to answer the following hypotheses:

[H1] Participant behaviors will be influenced by the presence of triggers associated with biases listed in table below.

[H2] Expert participants will perform better than open-division cyberattackers.

Highlights of trigger-specific and global trigger impact findings:

Trigger-specific Behavioral Impacts:

All the triggers met IARPA program-metrics with at least one behavioral impact in the expected direction, though not all were statistically significant. Due to the open world design, some triggers were difficult to isolate from network observables, thus multiple regression analysis was used to identify the contribution of a trigger to behavioral impacts. Note that multiple regression analysis is a foundational technique used in Structural Equation Modeling (SEM), which was originally proposed as our analysis method in the proposal, Statement of Work, and the first version of the CogVuln playbook.

The Loss Aversion trigger may be considered the most successful trigger among those encountered. In this trigger, the attacker faces a choice of using found credentials that can carry an increased risk of detection and resulting loss of access, or using a known exploit that takes more time and effort. This trigger drew statistically significant portions of MITRE Techniques, commands, and hacker time to areas outside of the attack path. Since being on the attack path is the only way for the attacker to make progress, this trigger reduced attacker progress and wasted attacker resources. The mean proportion of attack techniques targeting a VM associated with the Loss Aversion trigger is 4 times greater for the trigger group vs. the control group.

The Confirmation bias trigger includes an aliased command that terminates their session unexpectedly forcing a rational attacker to stop and reassess, or irrationally persist repeatedly beyond initial failures, which could also increase discoverability. This trigger found a statistically significant impact on attackers in their frequency of samba commands in the expected direction. Multiple regression analysis also show significant effects on increasing the time to task completion and attacker resources wasted by the trigger group.

The Availability trigger included valuable sounding filenames to prompt attackers to download files that can increase detectability. Indeed, there was a marginally significant effect in increased count of download and other file interactions for the trigger group. At the same time, the trigger group spent less time with the files, suggesting less checking and more risky behaviors.

The Base Rate Neglect trigger included additional false admin accounts in the cyberrange that attempted to entice the hacker to interact with them. Although not statistically significant, there was a positive effect size for participants who encountered the trigger, indicating longer and more commands used in the trigger group compared to the control group.

The Sunk Cost trigger involves choosing between attempts to identify contents of protected files vs. repeated attempts to unlock password protected files solely based on filename salience and assumed value. This trigger did not directly show statistically significant differences between trigger and control group, but multiple regression analysis showed marginally significant effects in increased time until task completion and increased attack resources wasted. Qualitative inspection suggests attackers used trigger specific information in unexpected ways such as using prior passwords from other contexts in other locations.

Overall, three of the five triggers had at least marginally significant differences between the trigger vs. non-trigger groups. The data also suggests all triggers impacted participants in the direction consistent with expectations, i.e. worse performance for the trigger group. Collecting additional data targeted to the triggers in Phase II may increase statistical power and confidence in the effect of the triggers. It is important to note that the triggers are realistic network artifacts, and sensors do not rely on deception or even the existence of triggers. For example, repeated attempts of admin logins can occur with or without fictional credentials.

Global Behavioral Impacts:

The bias triggers collectively impacted hacking behaviors negatively across multiple metrics. Compared to the control group, Exp 2 Trigger group generally (1) made less progress on the assigned hacking objectives, and (2) wasted proportionally more attack resources in terms of commands, MITRE Attack Techniques, and time targeting VMs irrelevant to the hacking objectives (i.e., not on the attack path). Further, the behavioral impacts of triggers tend to be less for expert than novice hackers. Taken together, this study provided strong confirmatory evidence that the designed triggers derived from cognitive heuristics/biases are effective distractors that attackers are naturally gravitating towards, and can be further refined to be cyber defense mechanisms.

Analysis on different subsets of the experimental data provided other important findings. First, sensitivity of the metrics for behavioral impacts vary by trigger design and placement on the cyber range (with respect to the hacking scenario). Observability of behavioral impacts for triggers embedded on relevant VMs tend to be more difficult than those embedded on irrelevant VMs; however, rate of encounter for triggers embedded in VMs not on the attack path can be rather low (as in survey response rates). Trigger design can dictate which behavioral metrics to be sensitive or relevant. Sunk cost triggers can lengthen time to complete hacking tasks. One incidental observation is that encountering triggers in a VM might actively keep participants focused or engaged on that VM rather than moving to subsequent targets. This had the opposite effect of accomplishing more hacking objectives. Although this was a single incident, such an effective distractor trigger may inspire more elaborate defenses that can potentially corral and keep attackers occupied in predictable environments.

Publications

Shuo Huang, Frederick Jones, Nikolos Gurney, David Pynadath, Kunal Srivastava, Stoney Trent, Peggy Wu, and Quanyan Zhu (2024). PsybORG+: Modeling and Simulation for Detecting Cognitive Biases in Advanced Persistent Threats. In proceedings of 2024 IEEE Military Communications Conference (MILCOM). Washington, DC. Oct 28 to Nov 1, 2024.

References

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. Journal of personality and social psychology, 92(5), 938.

Frederick, S. (2005). Cognitive reflection and decision making. Journal of Economic perspectives, 19(4), 25-42.

Soto, C. J., & John, O. P. (2017). Short and extra-short forms of the Big Five Inventory–2: The BFI-2-S and BFI-2-XS. Journal of Research in Personality, 68, 69-81.

Zhang, D. C., Highhouse, S., & Nye, C. D. (2019). Development and validation of the general risk propensity scale (GRiPS). Journal of Behavioral Decision Making, 32(2), 152-167.