ReSCIND Performer HSR Dataset Cover Sheet CIRCE Project

ReSCIND

Dataset Details

Dataset Title:	CIRCE Study 1: Loss Aversion—Endowment Effect (S1: LAEE)	
Dataset Citation:	CIRCE Team (authors list TBD). (2025). Context-driven Interventions through Reasoning about Cyberpsychology Exploitation, Study 1: Loss Aversion—Endowment Effect [Data set]. Charles River Analytics, Cambridge, Massachusetts, USA. DOI TBD.	
Data Format:	Zip file (.zip, .7z) archive of .xlsx and .json files	Data Size: ~3.7 GB
How to access dataset:	Email: www.cra.com/projects/circe (Not online yet)	
Point of contact for data questions:	Spencer Lynn Email: slynn@cra.com www.cra.com/projects/circe	

Description of Scenario

Experiment Objectives

As part of the IARPA ReSCIND program, this experiment was designed to determine the efficacy of cognitive biases and heuristics ("CogVulns") as cyber-psychological network defenses. The data described here were generated from an experiment that examined defenses based on the endowment effect facet of loss aversion using a realistic cyber challenge and experienced red teamers as a proxy for hackers.

Experiment Description

CIRCE cyber experiments began with an on-line questionnaire session to survey hacker stills, established measures of CogVuln susceptibility, demographics, and psychological characteristics. Following the survey session, the experiment comprised two one-hour, within-subject sessions. Sample size was 34 participants, working alone (not together as a team). Participants attacked a network implemented in the SimSpace Cyber Force network simulation environment. Participants were given a specific mission and provisioned with required resources. The two sessions were pseudo-randomly assigned to be treatment (CogVuln trigger present) or control (no CogVuln trigger), differing in mission specifics to mitigate learning across sessions.

The LAEE study attempted to exploit the endowment effect CogVuln by providing an attacker with an endowment (i.e., easily gained access to a target node) and then threatening that endowment with the intent of making the attacker work harder to maintain it.

This experiment was designed to assess the efficacy of bias susceptibility sensors, trigger effectives, and associations with established measures of the CogVuln in the psychology literature and personality and demographic characteristics of the attackers.

Experimental Results

Loss aversion variation and susceptibility were successfully captured by bias sensors and exploitation of attacker susceptibility to endowment effect impacted attack effectiveness.

Bias susceptibility sensors (including, e.g., exploitation time, use of stealthy commands, monitoring for defensive activities, and files access verbosity) can be useful in combination to predict susceptibility. Scores from different sensors associated with different established measures for loss aversion, indicating ecological validity. Sensors were also predictive of bias trigger impacts, indicating sensor effectiveness.

Bias triggers were defensively effective in two areas: (1) participants had a reduced attack success rate, as indicated by a reduced number of successfully exfiltrated target data files. (2) Participants made limited progress toward their goal, as indicated by a reduced number of observations of commands that were related to completing the exfiltration kill chain.

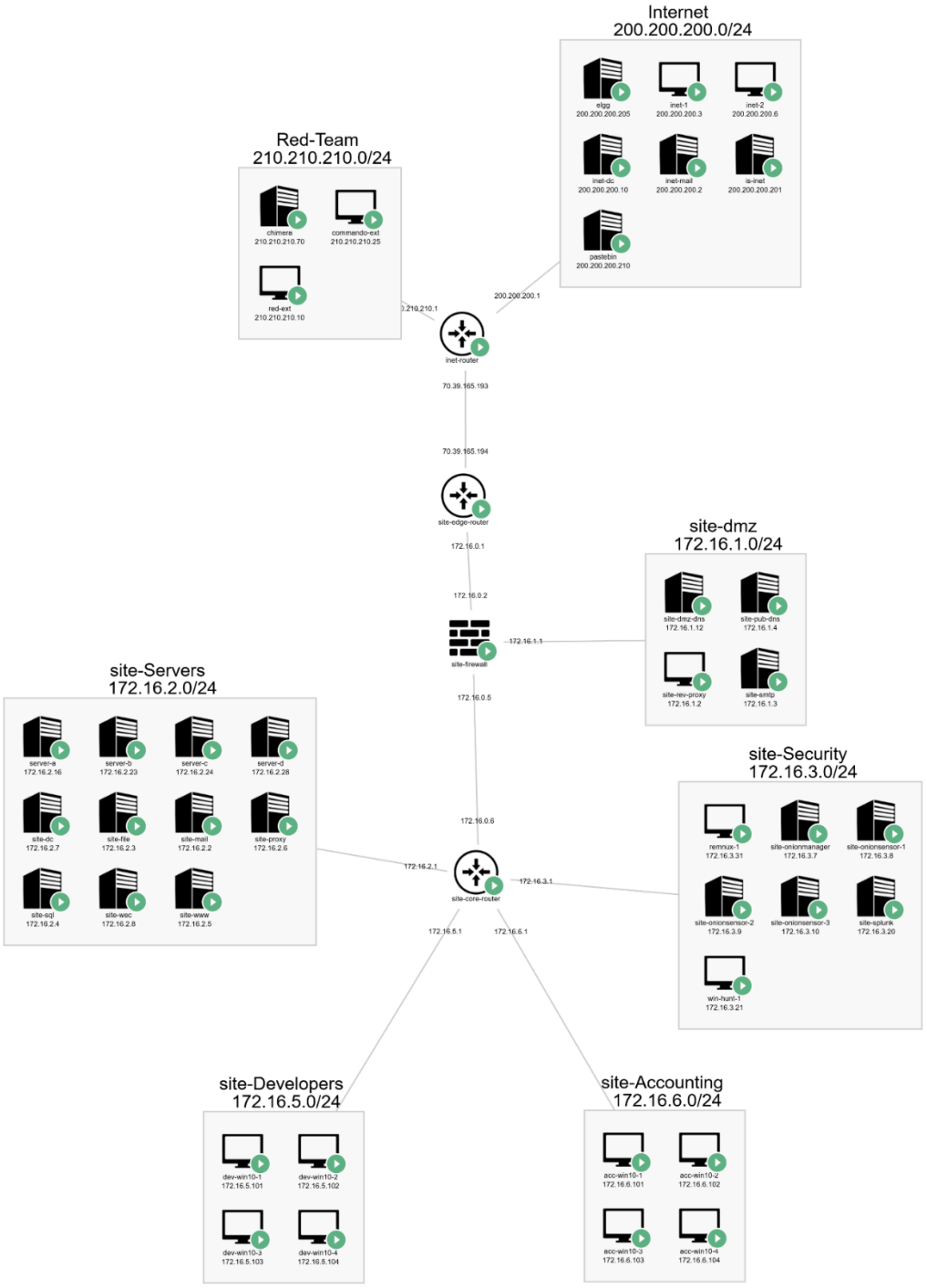
Cyber Environment

The cyber range comprised a network of virtual machines implemented in the SimSpace Cyber Force network simulation environment. The scenario has subnets of computers presenting a range of targets for the attacker.

Participants log into to their own instance of the test bed remotely (e.g., from home). To ensure control of experiments, participants were not able to deploy their own hacker toolsets, previously created scripts, etc., on the test bed. Once logged into the attacker virtual machine (their staging ground) on the test bed, they use cyberattack software tools provided to them against the target network.

The testbed network topology is illustrated in Figure 1.

Figure 1: CIRCE Loss Aversion—Endowment Effect network topology



DATA

DATA SOURCES

Primary Data Sources

Data collected directly from the experiment environment.

Category	Data Source	Examples of Select Data Features
Survey Data	Qualtrics survey metadata	Participant ID, timestamps, and other deidentified metadata from the survey host platform, Qualtrics
	Demographics	Gender, age, education level, English fluency, current employment
	Attacker skill/experience	Skill across five cyber domains from National Institute of Standards and Technology (NIST)/National Initiative for Cybersecurity Education (NICE). Also six MITRE-provided skill items.
	Psychometric questionnaires)	Short form positive and negative affect schedule (PANAS; MacKinnon et al. 1999), the 30-item Big Five Inventory (BFI-2S; Soto & John, 2017) emotional and personality scales, the General Risk Propensity Scale (GRiPS; Zhang et al. 2018)
	CogVuln established measures)	Base-rate neglect (Berthet, 2021), numeracy (Cokely et al., 2012), the Cognitive Reflection Test (Toplak et al., 2011), measures of loss aversion (Berthet, 2021), representativeness (Adult Decision-Making Competence; Bruine de Bruin et al., 2007), confirmation bias (Berthet, 2021), sunk cost fallacy (Teovanovic et al., 2015; ADMC, Bruine de Bruin et al., 2007), and anchoring bias (a modified version of Teovanovic, 2019)

Network Data	Splunk	Host monitoring from windows event logging, network monitoring via suricata, network monitoring via zeek, custom command line logging on kali

Derivative Data Sets

Datasets created from aggregating, analyzing, curating, and labeling the source data.

Category	Data Source	Examples of Select Data Features
Data Collector Output	Logfiles, Splunk database queries	A data collector queries logs and other raw-form cyber data for specific events. Outputs is a JSON file. Cyber activities of interest are observations that specific bias sensors and trigger evaluators process. Features include command line and PowerShell I/O, login events, file access, exfiltration events
State Abstractor Output	Data Collector output	A state abstractor receives data from data collectors and outputs a stream of data with measurements taken at specific intervals (e.g., 1 minute) as determined by adjustable parameters. Features include measures that bias sensors and trigger evaluators summarize, such as time to exploit a host and command stealthiness, and relevant data collectors, bias sensors, and trigger evaluators
Session Information	Data Collector output	Participant ID#, scenario version, experimental condition, CogVuln study ID#
Bias Sensor Data	Data Collector and State Abstractor output	A bias sensor receives data from state abstractors and outputs a stream of sensor data. Features include sensor measure name, time interval from start of session (set by state abstractor), score per time interval, contributing state abstractors, and relevant scenarios

Trigger Evaluator Data	Data Collector and State Abstractor output	A trigger evaluator receives data from data collectors and state abstractors. It outputs a single value that measures a specific trigger's effectiveness. Features include the evaluator name, associated triggers, contributing data collectors, applicable CogVulns, time interval from start of session (set by state abstractor), score per time interval

RESEARCH

Hypotheses

The LAEE dataset was used to address the following hypotheses:

Category	Detailed Hypotheses
A: Bias Sensor Ecological Validity	**Hypothesis**: A normalized assessment of the time taken to establish a foothold on a second host will produce a value within 1.5 standard deviations of the normalized established measure result for the endowment effect. **Hypothesis:** A normalized assessment of the inverse of noisiness of tools used to gain a foothold on the second host (looking at both PowerShell commands and network scan noisiness) will produce a value within 1.5 standard deviations of the normalized established measure result for the endowment effect. **Hypothesis:** A normalized assessment of the behavioral indicators of paranoia on the second host will produce a value within 1.5 standard deviations of the normalized established measure result for the endowment effect. **Hypothesis:** Each of the above hypothesis variables will be correlated with increases in the established measure for the endowment effect (correlation coefficient of 0.3 or higher).

B: Bias Trigger Effectiveness	**Hypothesis:** Attackers will take longer to establish persistence on a second host in the experimental (trigger) condition than in control (no trigger) condition. **[Increase in Time to Task Completion]** **Hypothesis:** Attackers will spend more time investigating defender threats in the experimental (trigger) condition than in the control (no trigger) condition. **[Increase in Time Wasted and Cognitive Effort Spent]**
C: Bias Sensor Effectiveness	**Hypothesis:** Increases in the susceptibility sensor values (Hypothesis A dependent variables) will be correlated with decreases in performance values (Hypothesis B dependent variables). **Hypothesis:** Increases in sensor values for loss aversion will be correlated with a reduced likelihood to shift targets to a different host (and/or a longer delay before shifting targets to a different host) amongst B, C, or D.
D: Bias Trigger Ecological Validity	**Hypothesis:** Increases in the established measure outcome are correlated with larger trigger impacts (Hypothesis B dependent variables). **Hypothesis:** Increases in established measure outcomes will be correlated with a reduced likelihood to shift targets to a different host (and/or a longer delay before shifting targets to a different host) amongst B, C, or D.

Publications

1. Vang, J., & Revelle, M. (2024). Formalizing Cognitive Biases for Cybersecurity Defenses. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security* (CCS '24, pp. 4991-4993). https://doi.org/10.1145/3658644.3691403

References

1. Berthet, V. (2021). The Measurement of Individual Differences in Cognitive Biases: A Review and Improvement. Frontiers in Psychology, 12(February), 1–12. https://doi.org/10.3389/fpsyg.2021.630177
2. 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), 938–956. https://doi.org/10.1037/0022-3514.92.5.938
3. Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. Judgment and Decision making, 7(1), 25-47.
4. Mackinnon, A., Jorm, A. F., Christensen, H., Korten, A. E., Jacomb, P. A., & Rodgers, B. (1999). A short form of the Positive and Negative Affect Schedule: Evaluation of factorial validity and invariance across demographic variables in a community sample. Personality and Individual differences, 27(3), 405-416.
5. 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. https://doi.org/10.1016/j.jrp.2017.02.004.
6. Teovanović, P. (2019). Individual differences in anchoring effect: Evidence for the role of insufficient adjustment. Europe's Journal of Psychology, 15(1), 8.
7. Teovanović, P., Knežević, G., & Stankov, L. (2015). Individual differences in cognitive biases: Evidence against one-factor theory of rationality. Intelligence, 50, 75-86.
8. Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory & cognition, 39(7), 1275-1289.
9. 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.