ReSCIND Performer HSR Dataset Cover Sheet

ReSCIND

Dataset Details

Dataset Title:	CIRCE Study 3: Representativeness—Base Rate Neglect (S3:RBRN)		
Dataset Citation:	Guarino, S., K.R. Bhat, B.A. Clegg, R.S. Gutzwiller, S. Harrison, J. Hypolite, D. Kelle, S.S. Latiff, C.M. Lewis, M.E. Locasto, M. Revelle, M. Sieffert, M. Slocum, T.T. Tran, C. Wu, and S.K. Lynn. (2025). Context-driven Interventions through Reasoning about Cyberpsychology Exploitation: **Representativeness-Base Rate Neglect Data Set**. Charles River Analytics, Cambridge, Massachusetts, USA.		
Data Format:	Zip file (.zip, .7z) archive of .xlsx and .json files	Data Size:	~3.7 GB
Dates & Duration:	One 1.5-hour survey session, collected as early as mid-September 2024 for some participants Oct/21/2024 – Dec/20/2024: Two 1-hour cyber sessions per participant	Time Zone:	US Eastern Standard Time
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 were generated from an experiment that examined defenses based on the base rate neglect facet of the representativeness heuristic 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The RBRN study attempted to exploit the base rate effect CogVuln by introducing an out-of-place, outdated vulnerability (the neglect target) in an otherwise up-to-date network environment, with the objective of providing an apparently easy target that the attacker should, if paying attention to base-rate information on the network, ignore as an obvious lure.

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

Using simple, out-of-place targets can induce base rate neglect, which can then be defensively exploited by inserting additional honey-data in those out-of-place targets.

Bias sensors (ratio of commands on the out-of-place target to commands to commands on the valid target, exploration depth, command verbosity, and density of commands on the out-of-place target) predicted susceptibility, showing ecological validity and were predictive of bias trigger impacts, indicating sensor effectiveness. Sensors were also predictive of bias trigger impacts, indicating sensor effectiveness.

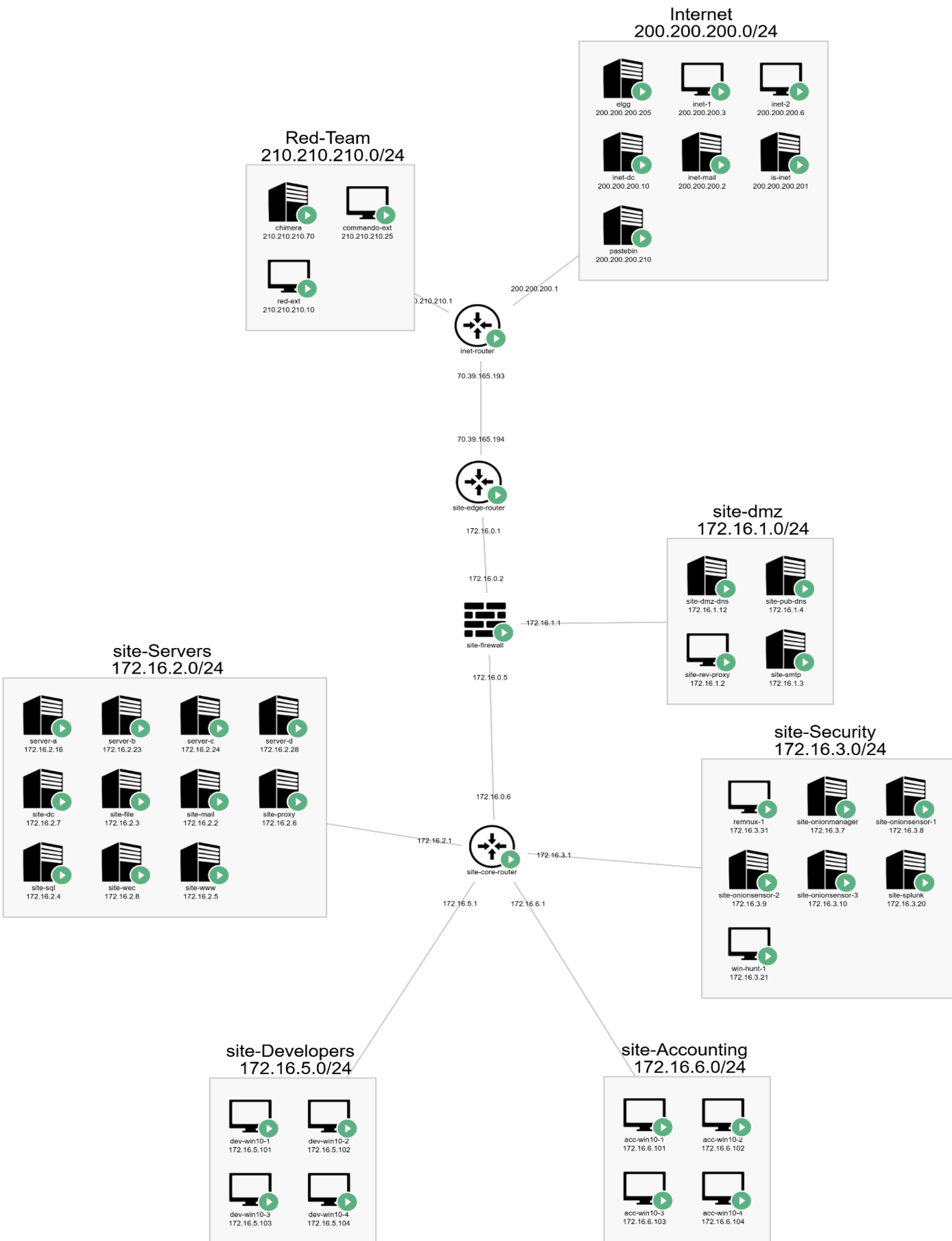
Bias trigger effectiveness results showed that attackers were willing to waste time on the out-of-place server, in itself showing base-rate neglect behavior. Exploitation of base rate neglect showed increased cognitive effort (measured by the number of distractor files exfiltrated), increased detectability (measured by number of interactions with the neglect vulnerability and by the number of noise files exfiltrated), and increase in time wasted (measured by time spent trying to exploit the neglect vulnerability).

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 Representativeness—Base Rate Neglect 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)
	Base Rate Neglect Scenario Check	Post-session survey to assess estimated base rates of examples in the scenario
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

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 and State Abstractor 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

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

Research

Hypotheses

The RBRN dataset was used to address the following hypotheses:

Category	Detailed Hypothesis
A: Sensor Ecological Validity	**Hypothesis**: A normalized count of directories explored on the BusyBox server will be within 1.5 standard deviations of normalization of the established base rate neglect measure **Hypothesis:** A normalized count of commands exploring exploitability and permissions will be within 1.5 standard deviations of normalization of the established base rate neglect measure **Hypothesis:** A normalized assessment of the time spent exploiting the BusyBox server will be within 1.5 standard deviations of normalization of the established base rate neglect measure **Hypothesis:** A normalized assessment of time pressure response will be within 1.5 standard deviations of normalization of the established base rate neglect measure **Hypothesis:** Each of the above hypothesis variables will be correlated with increases in the established measure for base rate neglect (correlation coefficient of 0.3 or higher).
B: Trigger Effectiveness	**Hypothesis:** Attackers will spend more time on the BusyBox server when the trigger is active **[increased wasted time]** **Hypothesis:** Attackers will have an increased rate of commands on the BusyBox server when the trigger is active **[increased attacker detectability]** **Hypothesis:** Attackers will successfully exfiltrate fewer of the target files when the trigger is active **[decreased rate of attack success]** **Hypothesis:** Attackers will download a larger percentage of decoy files when the trigger is active **[increased cognitive effort]**
C: 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).
D: Trigger Ecological Validity	**Hypothesis:** Increases in the established measure outcome will be correlated with larger trigger impacts (Hypothesis B dependent variables).

Publications

Publications and Conference Presentations

Clegg, B.A. (2025). CIRCE: Context-driven interventions through reasoning about cyberpsychology exploitation. In: Advancing Cyber+Human Research Session, 28th Annual CyberPsychology, CyberTherapy and Social Networking Conference, Sydney, Australia, 15-17 July.

Guarino, S., D. Kelle, C. Wu, M. Slocum, M. Sieffert, K.R. Bhat, R. Gutzwiller, and M. Neisser. (In press). Challenges and solutions in using virtual testbeds to study hacker cognitive constraints. Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC), Orlando, Florida, 1-4 December 2025.

Vang, J., & M. Revelle. (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

Data Sets

Guarino, S., K.R. Bhat, B.A. Clegg, R.S. Gutzwiller, S. Harrison, J. Hypolite, D. Kelle, S.S. Latiff, C.M. Lewis, M.E. Locasto, M. Revelle, M. Sieffert, M. Slocum, T.T. Tran, C. Wu, and S.K. Lynn. (2025). Context-driven Interventions through Reasoning about Cyberpsychology Exploitation: **Anchoring Bias Data Set**. Charles River Analytics, Cambridge, Massachusetts, USA. Available from https://osf.io/q96nd/.

Guarino, S., K.R. Bhat, B.A. Clegg, R.S. Gutzwiller, S. Harrison, J. Hypolite, D. Kelle, S.S. Latiff, M.E. Locasto, M. Revelle, M. Sieffert, M. Slocum, T.T. Tran, C. Wu, and S.K. Lynn. (2025). Context-driven Interventions through Reasoning about Cyberpsychology Exploitation: **Asymmetric Dominance Data Set**. Charles River Analytics, Cambridge, Massachusetts, USA. Available from https://osf.io/q96nd/.

Guarino, S., K.R. Bhat, B.A. Clegg, R.S. Gutzwiller, S. Harrison, J. Hypolite, D. Kelle, S.S. Latiff, C.M. Lewis, M.E. Locasto, M. Revelle, M. Sieffert, M. Slocum, T.T. Tran, C. Wu, and S.K. Lynn. (2025). Context-driven Interventions through Reasoning about Cyberpsychology Exploitation: **Confirmation Bias Data Set**. Charles River Analytics, Cambridge, Massachusetts, USA. Available from https://osf.io/q96nd/.

Guarino, S., K.R. Bhat, B.A. Clegg, R.S. Gutzwiller, S. Harrison, J. Hypolite, D. Kelle, S.S. Latiff, C.M. Lewis, M.E. Locasto, M. Revelle, M. Sieffert, M. Slocum, C. Wu, and S.K. Lynn. (2025). Context-driven Interventions through Reasoning about Cyberpsychology Exploitation: **Loss Aversion-Endowment Effect Data Set**. Charles River Analytics, Cambridge, Massachusetts, USA. Available from https://osf.io/q96nd/.

Guarino, S., K.R. Bhat, B.A. Clegg, R.S. Gutzwiller, S. Harrison, J. Hypolite, D. Kelle, S.S. Latiff, C.M. Lewis, M.E. Locasto, M. Revelle, M. Sieffert, M. Slocum, T.T. Tran, C. Wu, and S.K. Lynn. (2025). Context-driven Interventions through Reasoning about Cyberpsychology Exploitation: **Representativeness-Base Rate Neglect Data Set**. Charles River Analytics, Cambridge, Massachusetts, USA. Available from https://osf.io/q96nd/.

References

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

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

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.

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.

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.

Teovanović, P. (2019). Individual differences in anchoring effect: Evidence for the role of insufficient adjustment. *Europe's Journal of Psychology*, *15*(1), 8.

Teovanović, P., Knežević, G., & Stankov, L. (2015). Individual differences in cognitive biases: Evidence against one-factor theory of rationality. *Intelligence*, *50*, 75-86.

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.

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.