# ReSCIND Performer HSR Dataset Cover Sheet

ReSCIND

## Dataset Details

| | | |
|---|---|---|
| Dataset Title: | CIRCE Study 3: Confirmation Bias (S2:CB) | |
| 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: **Confirmation Bias Data Set**. Charles River Analytics, Cambridge, Massachusetts, USA. | |
| Data Format: | Zip file (.zip, .7z) archives of .xlsx and .json files | Data Size: ~4.1 GB |
| Dates & Duration: | • One 1.5-hour survey session, collected as early as mid-September 2024 for some participants<br>• Dec/8/2024 – Feb/25/2025: Two 1-hour cyber sessions per participant | Time Zone: US Eastern Standard Time |
| How to access dataset: | https://osf.io/q96nd/ | |
| Point of Contact for data questions: | Spencer Lynn<br>slynn@cra.com<br>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 confirmation bias 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 CB study attempted to exploit the confirmation bias CogVuln by providing the attacker with given a general description of the data to be extracted (financial files in one version, human resource files in another). The goal was to use the attacker's assumptions about what such data looks like to provide distractor files that seem reasonable, while hiding the real files with alternate file names and extensions.

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

Providing attackers with observations congruent with their expectations resulted in attackers concentrating on alternate targets that fit their expectations, rather than searching for evidence that they were being misled.

Bias sensors (specificity vs breadth of search focus, time-pressure, directory access) predicted susceptibility, showing ecological validity and were predictive of bias trigger impacts, indicating sensor effectiveness.

Bias trigger effectiveness results showed that attackers tended to exfiltrate fewer target files (rate of attack success) in the presence of the trigger.

### *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.

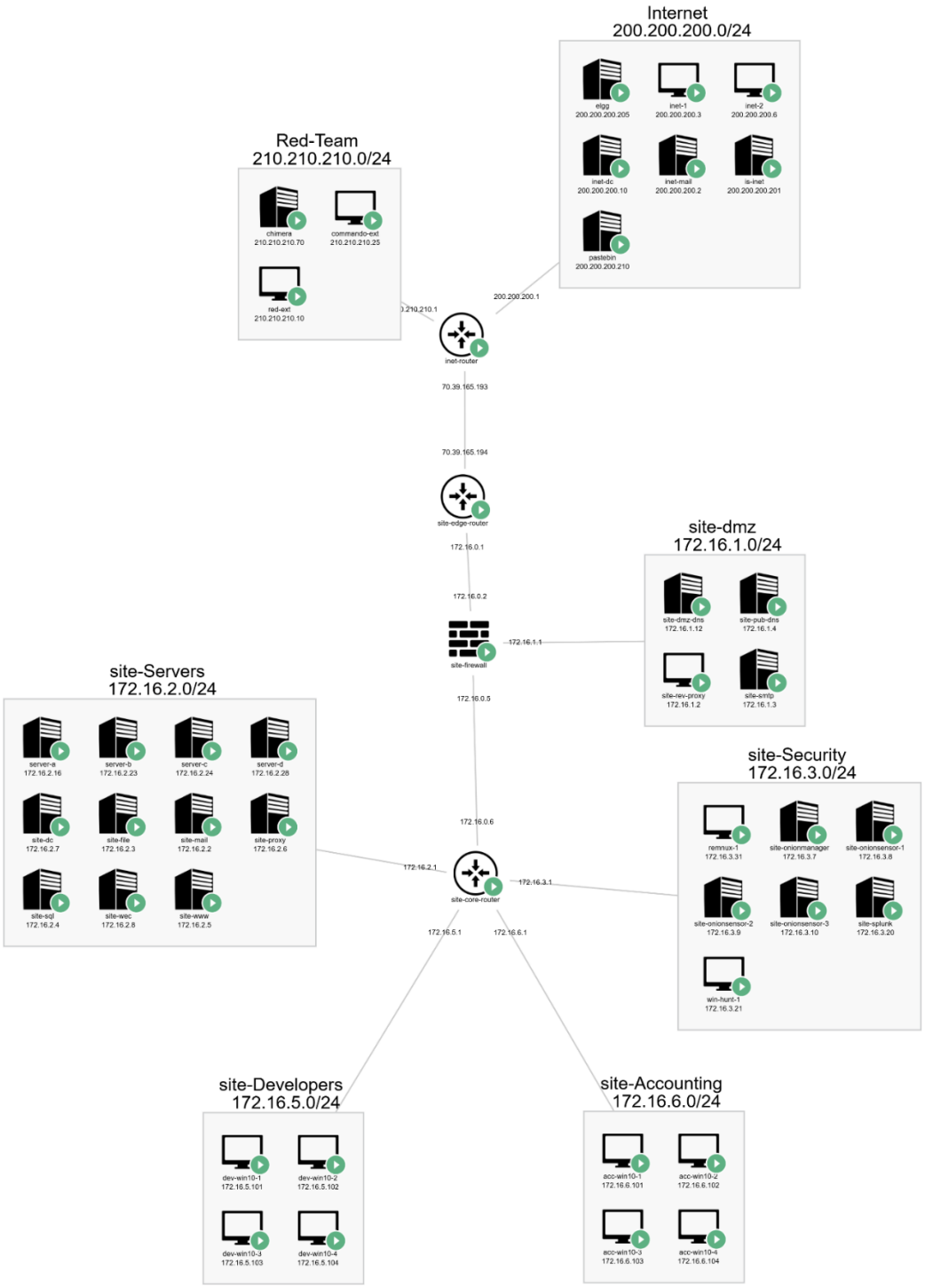


*Figure 1: CIRCE Confirmation Bias network topology.*

# 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 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 |

| | | |
|---|---|---|
| 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 CB dataset was used to address the following hypotheses:

| Category | Detailed Hypotheses |
|---|---|
| A: Sensor ecological validity | **Hypothesis CB1-SEV-1:** Normalized value of *search focus* sensor will be within 1.5 standard deviations of the normalized value of the confirmation bias established measure<br><br>**Hypothesis CB1-SEV-2:** Inverse of the normalized value of *search broadness* sensor will be within 1.5 standard deviations of the normalized value of the confirmation bias established measure<br><br>**Hypothesis CB1-SEV-3**: Normalized value of the *command verbosity* sensor will be within 1.5 standard deviations of the normalized value of the confirmation bias established measure<br><br>**Hypothesis CB1-SEV-AGG:** Correlation-weighted aggregate of the sensors will be within 1.5 standard deviations of the normalized value of the confirmation bias established measure |

| Category | Detailed Hypotheses |
|---|---|
| B: Trigger effectiveness | **Hypothesis CB1-TE-1:** *Target file exfiltration* performance metric (**rate of attack success**) will show a decrease from control to experimental condition, with Cohen's d ≥ 0.5<br><br>**Hypothesis CB1-TE-2:** *Distractor file exfiltration* performance metric (**cognitive effort**) will show an increase from control to experimental condition, with Cohen's d ≥ 0.5<br><br>**Hypothesis CB1-TE-3:** *Noise file exfiltration* performance metric (**detectability**) will show an increase from control to experimental condition, with Cohen's d ≥ 0.5<br><br>**Hypothesis CB1-TE-4:** *Number of file search commands* performance metric **(cognitive effort)** will show an increase from control to experimental condition when Confirmation Bias susceptibility is high, with Cohen's d ≥ 0.5 |
| C: Sensor effectiveness | **Hypothesis CB1-SE-1:** *Search Focus* sensor value will correlate with performance reductions from control to experimental conditions<br><br>**Hypothesis CB1-SE-2:** *Search Broadness* sensor value will inversely correlate with performance reductions from control to experimental conditions<br><br>**Hypothesis CB1-SE-3**: *Command Verbosity* sensor value will correlate with performance reductions from control to experimental conditions<br><br>**Hypothesis CB1-SE-AGG:** Correlation-weighted aggregate of the sensor values will correlate with performance reductions from control to experimental conditions |
| D: Trigger ecological validity | **Hypothesis CB1-TEV-1:** Confirmation bias established measure will correlate with decrease in *target file exfiltration* from control to experimental conditions<br><br>**Hypothesis CB1-TEV-2:** Confirmation bias established measure will correlate with increase in *distractor file exfiltration* from control to experimental condition<br><br>**Hypothesis CB1-TEV-3:** Confirmation bias established measure will correlate with increase in *noise file exfiltration* from control to experimental condition<br><br>**Hypothesis CB1-TEV-4:** Confirmation bias established measure will correlate with increase in *number of file search commands* from control to experimental condition |

*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.