ReSCIND Performer HSR Dataset Cover Sheet

ReSCIND

Dataset Details

Dataset Title:	CASPAR Stage 2 (biases from required CogVulns)		
Dataset Citation:	TBD		
Data Format:	Zip files for each major data source	Data Size:	~65GB (2.5Gb zipped)
Dates & Duration:	Nov 23, 2024 – Jun 30, 2025 One or more multi-hour sittings	Time Zone:	Administered in EST
How to access dataset:	https://osf.io/wm39e/		
Point of Contact for data questions:	Dr. Zak Fry or Dr. Hajime Inoue: zfry@grammatech.com / hinoue@grammatech.com		

Description of Scenario

Experiment Objectives

This experiment was designed to measure bias in cyber-attack behavior, including the ability to sense bias (relative to established methods in other domains) and the ability to amplify biases in a defensive context (e.g., confusing attackers, adding effort to attack tasks, etc.).

Experiment Description

The experiment contains three tiers of measurement: cyber-behavioral bias data (Gold), general bias behavioral data (Silver), and traditional bias surveys (Bronze). Participants completed all three tiers over several sittings, ranging from around 2-6 hours. Individual tiers and sub-components were incentivized individually to motivate comprehensive participation, though any component was able to be skipped. Gold tier objectives mimic popular "capture the flag" type cyber challenges, generally covering topics like gaining access to protected artifacts, reverse engineering passwords, exfiltrating sensitive data, etc. Silver tier was designed to parallel Gold objectives using games and activities to measure biased behavior in a non-cyber context. Bronze reimplements canonical bias measurement methods from the literature. We specifically measure the efficacy of "Sensors" (mechanisms to distinguish

biased vs unbiases behavior) and "Triggers" (mechanisms to induce or amplify the effect of a bias).

Experimental Results

Full details of the experimental results are available in our Data Report (email for access). Post-processing of the raw data crafted many intermediate metrics related to effort/time expended, number of attempts at critical tasks, and outcomes of important workflow decisions. These metrics were used as input to calculations for determining susceptibility and presence of the studied biases in the underlying behavior. We found that "Sensor" bias measures generally correlated well across different measurement tiers – this suggests that our measurement techniques are in agreement with one another *and* established definitions. We found that "Trigger" effects were generally successful in amplifying bias presence and decreasing success metrics like attacker productivity, rate of accomplishment, and situational understanding.

Cyber Environment

The cyber testbed for this set of objectives contains Dockerized machines all networked together and sandboxed from external internet access. Users are presented with a "home" machine that presents a "task tracker app" used to explain the current objective(s) and allows for flag input, experimental control options (e.g., pausing the study, abandoning a task, help menus, etc.). This home machine also includes a terminal and a browser. Individual objectives are siloed on groups of machines. The testbed is identical for all participants, but each participant accesses their own secure copy.

Parsing and Handling Data

Upon request, we can provide scripts (in the form of Jupyter Notebooks and supporting Python code) that synthesize aggregated effort metrics from the raw data. These also output data dictionaries that explain individual metrics and provide valid ranges. These scripts could be easily adapted to produce additional/modified metrics.

Data

Data Sources

Primary Data Sources

Collected directly from the experiment environment.

Category	Data Source	Examples of Select Data Features
System Logs	Docker Logs	The docker logs include the console logs for each container. The contents of these logs differ depending on the application running in the container and are also covered under "Scenario-specific Logs". The majority run the apache web server, and output the apache connection log in JSON format. Others run custom web applications, such the log viewer, or our Box server, all of which are written in JSON format. We use information about login attempts, web views, and downloads as part of the analysis to determine participant activity such as strategies (i.e. OSINT recovery of password reset questions), or durations (i.e. how long was the participant interacting with a container).
	/var/log Logs	We store the contents of /var/log to ensure we have the information if required, but currently do not use it as it is duplicative of other logs.
Experimental Logs	Task Tracker App Logs	Logging of all participant interactions with the app (e.g., starting a task, entering a flag, abandoning a task) including timestamps for every event. The contents of flag entries are logged to assess correctness/success of individual objectives.
	Scenario-specific Logs	Objectives have meta-data critical to metrics calculation and hypothesis testing – these were implemented by our team specific to the information needs per-bias. There are dynamic environmental/contextual mechanisms used to force participant decision making or various behaviors. For example, in the objective for Law of Small Numbers, one of the effort metrics was to identify how many times a user inspected an individual piece of information and where that information appeared in the total-ordering of all available items. To do this, we implemented back-end logging on the associated web interface that identified when users clicked on an individual item and kept track of when users switched pages – this information is all logged.

Additional Logs	Falco Logs	We use a mixture of pre-existing and custom rules to track process creation, parenthood, and shutdown. This allows us to track command line actions taken by participants. Examples include durations of john-the-ripper runs, and nmap, sqlmap, and ssh actions.
Network Data	PCAP	Our underlying network infrastructure is supplied by NS3, which enables us to record all network activity. These individual files are merged together to form a single PCAP file for the entire network, which is then fed into Snort and Suricata.
	IDS – Snort	Snort can signal particular kinds of network activity and attacks. We currently do not use these events, as we attempt to detect specific events directly (i.e. nmap runs, using falco rules), but they provide redundancy in case participants activity is unexpected (i.e. scripting their own network scanners or sql injectors).
	IDS - Suricata	We also run Suricata against a recent community supported rule set. Like Snort, we currently have not used it for analyses to date, but may be useful in future studies.

Derivative Data Sets

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

Category	Data Source	Examples of Select Data Features
Objective-specific Metrics	Task-specific Effort and Decision Metrics	Each bias has a crafted data "meta-sheet" that contains metrics of interest to our hypotheses, pulled and calculated from the raw data. For Law of Small Numbers, these include the following: last page accessed, number of data items clicked, number items for each critical machine inspected, Boolean for correct answer entered. This data is organized as a series of meta-sheets (Output_Data > MetaSheets) and corresponding data dictionaries (Output_Data > DataDicts) that explain data types and functions.
Aggregated Full-study Data	Master Data Sheet	For the purposes of hypothesis testing, we aggregate metrics from all meta-sheets by way of Sensor and Trigger judgements in a "master sheet". Each row is a single participant, and columns include established method holdout data (col C-G) and individualized judgements for a) whether the sensor indicated presence of a given bias, and b) whether the trigger was effective. These exist for all tiers (Gold, Silver, Bronze) for which data was measured. This meta-sheet (Output_Data > CASPAR_Full_Mastersheet.csv) has a corresponding data dictionary (Output_Data >

		CASPAR_Full_Mastershee_Data_Dictionary.csv) that explains each column.

Research

Hypotheses

We cover the following biases (and abbreviations) in this section:

- Base Rate Neglect (BRN)
- Law of Small Numbers (LSN)
- Gambler's Fallacy and Hot Hand Fallacy (GF/HH)
- Framing Effects (FE)
- Endowment Effect (EE)
- Sunk Cost Fallacy (SC)
- Near Miss Effect (NME)
- Hot Stove Effect (HSE)
- Cognitive Load Effect (CLE)
- Anchoring Bias (AB)
- Default/Distinctiveness Effect (DDE)
- Mere Exposure Effect (MEE)
- Confirmation Bias (CB)

This Gold dataset was used to test the following metrics or hypotheses (one per bias for Sensor "S" and Trigger "T" respectively):

[H1.1] BRN-S = ("pages accessed" <= 5") OR ("ratio A/B accessed" <= 1) OR ("flag correct" == FALSE)

[H1.2] BRN-T = ("number of machine A accessed" + "number of machine B accessed" < 0.1 "all machine items accessed") AND ("flag correct" == FALSE)

[H2.1] LSN-S = ("pages accessed" <= 2) OR ("machine A descriptions accessed < 3") OR ("flag correct" == FALSE)

[H2.2] LSN-T = ("number of machine A accessed" + "number of machine B accessed" < 0.1 "all machine items accessed") AND ("flag correct" == FALSE)

[H3.1] GF-S = 1 > 0.34 * ("correct flag" == TRUE) + 0.34 * ("login attempts before message" > "login attempts after message") + 0.34 ("reset attempts before message" > "reset attempts after message") + 0.34 * ("download attempts before message" > "download attempts after message")

[H3.2] GF-T = ("logged in before message" == FALSE) AND ("time to login" > 5min)

[H4.1] FE-S = (NOT "logged in prior to warning") AND ("time to login" > 5min)

[H4.2] FE-T = ("saw warning" AND ("sensor indicated susceptible")

[H5.1] EE-S = ("switched target machine" == FALSE)

[H5.2] EE-T = (“weighted proportion of easy tasks completed” > 1)

[H6.1] SC-S = (“number of easy tasks completed” > 3) AND (“time spent on hard task” > 5)

[H6.2] SC-T = (“hard task attempted”)

[H7.1] NME-PreS = sum(POST_accesses, SQLMap_accesses, SQLMap_processes) > 2

[H7.2] NME-T = sum(“403 responses”, “500 responses”) >=2 AND “500 responses” > 1

[H7.3] NME-PostS = “SQL after trigger” > 0

[H8.1] NME-PreS = (“total time on machine” / “login attempts”) > 5

[H8.2] NME-T = “escalating warnings shown” == TRUE

[H8.3] NME-PostS = (“total time on machine” / “login attempts”) > 5 (same as pre, different machines)

[H9.1] CLE-PreS = “last page accessed” >= 2

[H9.2] CLE-T = “accesses to pages 1-2" >= 6

[H9.3] CLE-PostS = “accesses to pages 3+” >= 4

[H10.1] AB-PreS = “POST attempts before trigger” > 0

[H10.2] AB-T = “saw additional anchoring info” == TRUE

[H10.3] AB-PostS = “POST attempts after trigger” > 1

[H11.1] DDE-PreS = “default or distinct names pursued on list 1” > 0

[H11.2] DDE-T = “chose default or distinct choices on first list” AND “given default or distinct list second”

[H11.3] DDE-PostS = “default or distinct names pursued on list 2” > 0

[H12.1] MEE-PreS = “previously-exposed names pursued on list 1” > 0

[H12.2] MEE-T = “chose previously-exposed choices on first list” AND “given MEE list second”

[H12.3] MEE-PostS = “previously-exposed names pursued on list 2” > 0

[H13.1] CB-PreS = “chose B-before-Not-B in priming task” == 1

[H13.2] CB-T = “assigned to confirmation group/vector” == TRUE

[H13.3] CB-PostS = “time spent on wrong strategy” > 4min

Each metric was used to calculate judgements in the MasterSheet described previously – hypotheses then compared these with Silver and Bronze tiers to gauge agreement (and thus efficacy). The “MasterSheet.ipynb Jupyter notebook contains in-depth explanations and code on how each of these judgements were calculated and justifications for the formulations in each case.

Publications

French, L., Thorpe, A., Salibayeva, K., Brown, S., Eidels, A., Forties, R., Fry, Z., Hewlett, E., & Inoue, H. (2024). *Combatting cyberattacks with cognitive bias*. [Conference presentation]. Performance and Expertise Research Centre Conference, Sydney, Australia.

Thorpe, A., French, L., Salibayeva, K., Brown, S., Eidels, A., Forties, R., Fry, Z., Hewlett, E., & Inoue, H. (2025). *Hackers are (only) human too: Understanding cognitive biases in cybersecurity* [Conference presentation]. Australasian Mathematical Psychology Conference, Sydney, Australia.

French, L., Thorpe, A., Salibayeva, K., Brown, S., Eidels, A., Forties, R., Fry, Z., Hewlett, E., & Inoue, H. (2025). *Hackers are (only) human (part) too: Validating and exploiting biases to disrupt hacker efficiency* [Conference presentation]. Australasian Mathematical Psychology Conference, Sydney, Australia.

References

The following papers helped to shape the Gold tier experiments – they overlap somewhat with Bronze established methods. See the Bronze tier cover sheet for a full accounting of the literature review and sources used in crafting established methods.

Base Rate Neglect - A selection of 8 questions developed from:

De Neys, W., & Glumicic, T. (2008). Conflict monitoring in dual process theories of thinking. *Cognition*, *106*(3), 1248-1299.

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological review*, *102*(4), 684.

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological review*, *80*(4), 237.

Kahneman and Tversky, 1973 in Murray, J., Iding, M., Farris, H., & Revlin, R. (1987). Sample-size salience and statistical inference. *Bulletin of the Psychonomic Society*, *25*, 367-369.

Murray, J., Iding, M., Farris, H., & Revlin, R. (1987). Sample-size salience and statistical inference. *Bulletin of the Psychonomic Society*, *25*, 367-369.

Yoon, H., Scopelliti, I., & Morewedge, C. K. (2021). Decision making can be improved through observational learning. *Organizational Behavior and Human Decision Processes, 162*, 155-188.

Gambler's Fallacy and Hot Hand Fallacy - five questions taken from:

Yoon, H., Scopelliti, I., & Morewedge, C. K. (2021). Decision making can be improved through observational learning. *Organizational Behavior and Human Decision Processes, 162*, 155-188.

Law of Small Numbers - 6 questions from:

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, *76*(2), 105.

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science*, *185*(4157), 1124-1131.

Yoon, H., Scopelliti, I., & Morewedge, C. K. (2021). Decision making can be improved through observational learning. *Organizational Behavior and Human Decision Processes, 162*, 155-188.

Framing Effects – 16 questions/ 8 comparisons from:

Berthet, V. (2021). The measurement of individual differences in cognitive biases: A review and improvement. *Frontiers in psychology*, *12*, 630177.

Sunk Cost Fallacy – 8 questions from:

Ronayne, D., Sgroi, D., & Tuckwell, A. (2021). Evaluating the sunk cost effect. *Journal of Economic Behavior & Organization*, *186*, 318-327.