

Exploiting Base Rate Neglect to Disrupt and Distract Cyber Attackers

K Raghav Bhat
Arizona State University
kbhat4@asu.edu

Robert S. Gutzwiller
Arizona State University
rgutzwil@asu.edu

Sean Guarino
Charles River Analytics
sguarino@cra.com

Spencer Lynn
Charles River Analytics
slynn@cra.com

Benjamin Clegg
Montana State University
benjamin.clegg@montana.edu

Joel Hypolite
Charles River Analytics
jhypolite@cra.com

Michael Seiffert
Assured Information Security
sieffertm@ainfosec.com

Michael Locasto
Narf Industries
michael.locasto@narfindustries.com

David Kelle
Charles River Analytics
dkelle@cra.com

Max Slocum
Assured Information Security
slocumm@ainfosec.com

Curt Wu
Charles River Analytics
cwu@cra.com

Scott Harrison
Charles River Analytics
sharrison@cra.com

Matthew Revelle
Montana State University
matthew.revelle@montana.edu

Susan Latiff
Charles River Analytics
slatiff@cra.com

Abstract

Oppositional human factors (OHF) seeks to exploit tendencies in human thinking to disrupt cyber attackers. One tendency is base rate neglect (BRN), where individuals overlook the likelihood of an event during reasoning, and instead base judgements on salient surface details. An expert sample of cyber red teamers completed cognitive bias survey measures, followed by missions in a cyber range. In the range, features on a server consistent with a vulnerability but out of context (extremely low base rate) were used to test whether these experts would ignore such base rates. BRN was found, including meaningful, significant performance reductions, suggesting a real, valid path for OHF techniques. Further, this approach can be employed even where bias susceptibility predictions for an attacker are unavailable.

Keywords: cybersecurity, human factors, decision-making, cognitive bias, oppositional human factors.

1. Introduction

Cyber defenders face an increasingly asymmetric challenge. While attackers evolve rapidly, employing novel techniques and leveraging cutting-edge tools like generative AI, and often backed by significant resources available to advanced persistent threats (APTs), defenders are frequently hamstrung by

reactive strategies and rigid protocols. For example, this imbalance is evident in the 2025 *CrowdStrike Global Threat Report*, where the average breakout time (how quickly an attacker moves laterally across a network) fell to just 48 minutes, with the fastest observed breakout occurring in 51 seconds (CrowdStrike, 2025). The challenge defenders face suggests the need for novel defense mechanisms.

Use of oppositional human factors (OHF) attempts to rebalance this asymmetry by targeting the human adversary rather than technical artifacts (Gutzwiller et al., 2018). OHF makes the case that any guideline intended to make a human's work easier can be usefully reversed when considering the experience of the attacker within the system. In cyber defense, the goal is not only to disrupt but also to potentially control the attacker, making tasks more difficult to accomplish, behavior more identifiable to defenders, and/or actions more time-consuming. Together these will increase the likelihood of detection and intervention. Incorporating cognitive aspects of attackers is not an entirely novel concept; however, it has mostly been explored through the lens of cyber deception. Traditional deception strategies have relied on mechanisms such as honeypots, decoy systems, and false file shares to mislead and delay adversaries (Ferguson-Walter et al., 2021; McKneely et al., 2023). Human-subject experiments have empirically demonstrated that these methods can increase attacker

confusion, prolong the duration of attacks, and, in some cases, prevent goal completion among red-team participants (Ferguson-Walter et al., 2018; Shade et al., 2020).

However, deception represents just one technique within a much broader set of strategies aimed at targeting attacker cognition. Exploiting specific cognitive vulnerabilities in human decision-making by shaping attacker perception or manipulating their choice environment are possible (Johnson, 2020; Johnson et al., 2021). The existence of *cognitive vulnerabilities* in cyber operations is substantiated by emerging empirical findings (Aggarwal et al., 2024; Cranford et al. 2021; Gutzwiller et al., 2024; in press; Hitaj et al., 2025). Gutzwiller et al. (2019) identified evidence of cognitive biases and heuristic-driven decision-making among professional red teamers in a realistic network experiment, including anchoring, sunk cost fallacy, and the “take-the-best” heuristic. Later analysis of 139 professional red teamers over a two-day experiment found extensive evidence of framing and confirmation bias (Gutzwiller et al., 2024). These findings demonstrated that cognitive biases can be systematically triggered and measured in cybersecurity settings. The current work builds on existing strategies by leveraging cognitive biases to make decoys more compelling. In this conceptualization, cognitive vulnerability exploitation complements deception methods.

This paper explores how base rate neglect (BRN) – a cognitive bias - can be operationalized to exploit attacker decision-making in realistic scenarios. BRN emerged as a viable candidate after review by subject matter experts, due to its robust empirical support in published literature, measurable effects in cyber testbed environments, and the availability of validated instruments to assess individual susceptibility to the bias. Then, based on cyber expert analysis, we focused on a specific heuristic used by attackers associated with BRN: the tendency for attackers to try to exploit outdated software, despite the broader degree to which the network is updated. When attackers pursue the old vulnerability despite its inconsistency with the rest of the system, they commit BRN i.e., misjudging probability by focusing on salient details while underweighting the likelihood that such a service would exist on an updated network.

1.1. Representativeness Bias & BRN

Base rate neglect is conventionally nested under the more general ‘representativeness’ *heuristic*. Gigerenzer and Gaissmaier (2011) define a heuristic as “a strategy that ignores part of the information, with the goal of making decisions more quickly and/or accurately than more complex methods”. The classical

explanation as to why this strategy is employed relies on dual-process theory, which posits that human cognition operates through two different pathways: a fast, effortless, intuitive and pattern-based system (i.e. System 1) and a slower, much more deliberate and rule-based system (i.e. System 2) (Evans and Stanovich, 2013; Kahneman, 2011; Stanovich, 1999). Heuristics rely on System 1 thinking, where efficiency is prioritized by limiting information search, and computation time, trading effort for speed.

For instance, when evaluating whether an email is legitimate, users often rely on surface-level cues such as professional formatting, company logos, and familiar sender names - rather than systematically verifying sender addresses, embedded links, or metadata (more effortful). As most genuine emails share the surface-level characteristics, individuals assume an email exhibiting them is also safe. This approach speeds decision-making, but increases susceptibility to phishing attacks, where attackers may intentionally mimic the appearance of legitimate communications.

Representativeness heuristic. The use of heuristics can lead to systematic biases and errors (Tversky & Kahneman, 1974) leading individuals to misestimate risk, misclassify events, or perceive false correlations. The representativeness heuristic is a mental shortcut that helps “ease” decision making by comparing information to our pre-existing mental stereotypes. Specifically, a person following this heuristic evaluates the probability of an uncertain event on the basis of: i) the similarity of its properties to its reference population and ii) how closely it conforms to patterns or mechanisms underlying its occurrence (Kahneman and Tversky, 1972).

Considering the previous example of phishing emails: rather than verifying security markers, users judge legitimacy based on how closely an email resembles their mental model of a genuine corporate message. In this case, similarity to the expectation overrides a systematic evaluation, leading to a mistake. This heuristic has tangible effects in defensive security settings; Hitaj et al. (2025) showed that manipulating sample size insensitivity in CTF scenarios altered attacker behavior.

Base Rate Neglect (BRN). Our work focuses on BRN, i.e. when individuals ignore or underweight prior base rates in their judgements, leading to erroneous assessments. One account (Bar-Hillel, 1980, 1990) suggests BRN occurs when people regard the base rate itself as being of low relevance to the decision compared to the other features present that are of high relevance. BRN is among the most extensively documented biases.

To illustrate, consider the example from Gigerenzer and Hoffrage (1995), with the descriptor information in brackets normally not shown to participants: *"The probability of breast cancer is 1% for a woman at age forty who participates in routine screening [base-rate]. If a woman has breast cancer, the probability is 80% that she will get a positive mammography [hit-rate]. If a woman does not have breast cancer, the probability is 9.6% that she will also get a positive mammography [false-alarm rate]. A woman in this age group had a positive mammography in a routine screening. What is the probability that she actually has breast cancer? __%."*

Though the test seems reliable, the actual chance is only about 8%. Because the disease is rare (1% prevalence), false positives (about 10%) far outnumber true positives. People ignore the low base rate and focus on the positive result, overestimating risk. At its core, this reflects BRN: overweighting specific evidence (e.g., the positive result) while underweighting prior information, so judgments are driven more by salient details than by base rates.

BRN's role in adversarial decision-making remains underexplored. Our study represents the first documented attempt to examine BRN within an OHF framework, with skilled cyber attackers, acting in an actual cyber-attack environment. Attackers often form assumptions that guide categorization, prediction, and preference, for example, regarding the role or vulnerability of a device on a network, while disregarding prior probabilities that contradict those assumptions. Thus, BRN can be strategically exploited to shape adversary expectations and subtly steer their decision-making in controlled directions.

1.2. Induction and Attacker Monitoring for Susceptibility

Application of BRN to the cyber adversarial environment is a key contribution in the current work. Creating *realistic* techniques for inducing, monitoring, validating and testing how the technique works, were all advanced significantly.

For realism, a core component of using BRN and other OHF techniques is the ability to monitor attacker behavior for susceptibility to a given OHF method or bias. These methods are evolving, moving from more traditional qualitative assessments to quantitative and discrete scoring methodologies (Gutzwiller et al., 2024; Vang & Revelle, 2024; IARPA, 2025).

We leverage two factors to further evolve these techniques. First, we identify and operationalize *bias indicators* based on activities that the attacker can take in the environment that we thought would be reflective of vulnerability to the bias. (We refer to '*bias indicators*' simply as '*indicators*' throughout the

remainder of this paper). We further attempted to validate the accuracy of these indicators by comparing them to a survey-administered established measure (EM) of the cognitive bias (Berthet, 2021, base rate neglect items from Study 2). In this way, we get a quick sense of whether the indicators have validity.

Second, we use the bias indicator-outcome relationship to see whether a participant's vulnerability to the bias was related to performance shifts expected between control (no exploitation of the bias) versus experimental conditions (explicit exploitation of the bias). If performance degradation is related to the bias as reflected by indicator activity or scores on the survey measure, then we are more confident the OHF technique worked as expected.

Additionally, operationalizing bias components was crucial to maintain cyber realism. Our approach to scenario generation was highly interactive with cyber and psychological SMEs as part of a broader design team. This provided deep understanding of the cognitive vulnerability we were seeking to exploit, and cyber analyst assessment and input to ensure we were exploiting it using realistic and valid manipulations in the cyber environment. In our scenario, we expose out-of-date service vulnerabilities that are simple and reliable to exploit, on hosts that have otherwise up-to-date patches, to lure the attacker down an apparent attack path that the defender controls.

Finally, this experiment was followed by a unique questionnaire to assess whether participants' estimations of the frequency of encountering various host services matched the assumptions upon which we based our design. Incorporating convergent validation ensures we are rigorously advancing OHF.

1.3. Experiment Hypotheses

The cyber attacker scenario introduces an outdated vulnerability (the neglected 'base' rate) into an otherwise up-to-date network environment, with the objective of presenting an apparently easy target that the attacker should recognize as out-of-place and ignore as an obvious lure if correctly paying attention to base-rate information about the network.

We tested the validity of the developed BRN indicators (i.e., that they are measuring relative susceptibility to BRN; and whether this susceptibility or the susceptibility measured by our EM relates to performance changes). To assess validity, we examined correlations between each indicator value and the EM (Berthet, 2021), with the further expectation that the EM will align with performance changes. Therefore we tested three hypotheses: ***H1***: the values of each BRN bias indicator will correlate with the score from the BRN EM. ***H2***: higher indicator values (more bias and bias susceptibility

detected) will correlate with greater changes in performance metrics. ***H3***: higher BRN in the established survey correlates with greater changes in performance.

The second area assessed effectiveness in degrading performance, specifically, whether exploiting BRN reduced performance relative to the control condition, where BRN manifests as well but was not targeted. ***H4***: attacker performance will be reduced under the experimental condition, compared to the control condition, resulting in increased time wasted, increased cognitive effort, and detectability, and decreased rate of attack success.

2. Methods and Materials

2.1. Participants

This study was part of a multi-experiment effort aimed at testing and validating the use of cognitive vulnerabilities (i.e., heuristics and biases) for cyber defense. Participants were recruited through professional networks within cyber challenge communities, using snowball sampling to expand the reach. The sample included professional hackers, red team members, penetration testers, and individuals from broader hacking communities. Participants were compensated at a rate of \$150 per hour. The significant compensation was chosen due to the niche expertise of the target population, who are typically difficult to recruit and retain—particularly in a longitudinal, multi-experiment research effort.

A larger participant pool was recruited for the overall effort, from which only a subset took part in the current study. Thirty-three (33) participants (all males) completed this study, with an average age of 32.3 years (SD = 7.89), with ages ranging from 19 to 46. In terms of education, 17 reported having a bachelor's degree, 12 had a master's, 1 a doctorate, and 1 a high school diploma or equivalent; 2 participants reported having completed some college but no degree. Of the 33 participants, 26 were native English speakers and 7 were non-native. Most reported high levels of English fluency: 24 = "mastery," 7 = "advanced," and only 2 = "upper-intermediate" levels. Regarding employment status, the majority were currently working for pay (28), while 4 identified as students, 1 = unemployed, and 0 opting to withhold a response.

A self-reported skill inventory was used during recruitment to assess cyber knowledge and relevant abilities. The inventory consisted of five offense-oriented skill categories adapted from the NIST inventory of attacker skill sets (Cybersecurity and Infrastructure Security Agency, 2025). Each item was rated on a 1–4 scale, and a cut-off was established at

an average rating of 2.0 or higher. In conjunction scores on a knowledge test and independent SME evaluation of cyber performance data was leveraged for participant inclusion in our study's data. On average, participants scored 2.99 across categories (*Exploitation (5 items)*, M= 2.79, SD= .61; *Host security (5)*, M= 2.87, SD= .75; *Network skill (5)*, M= 3.26, SD= .53; *System admin skill (2)*, M= 3.17, SD= .58, *Tool ratings (9)*, M= 2.88, SD= .53). Scores across six knowledge-based cyber test questions developed by MITRE were also good, and on average, participants missed fewer than one question (M= 0.85 incorrect, SD= 0.76), indicating proficiency.

2.2. Cyber Range, Scenario and Bias Indicator Design and Data Collection

Cyber Range. The study was conducted online, using the SimSpace Cyber Force platform, a high fidelity commercial, military-grade cyber range test bed routinely used across the US Department of Defense. Splunk, a commercial logging tool, enabled collection of a wide variety of command interactions. Security Onion, a platform for threat hunting, enabled capture of specific exploit uses. These data collection nodes were explicitly off-limits as attack targets.

Scenario Design. A focused cyber scenario was constructed within SimSpace, comprising a network of virtual machines configured to support and reflect the objectives of the study. The attacker started on an attacker host (with an available set of familiar cyberattack software) and proceeded with a mission to infiltrate a target network to either exfiltrate target data or establish persistence on the network. Four different test beds were developed to enable experimentation. First, two versions of each scenario were developed, to enable within-subject study participants so we could minimize the learning impacts from the first session. Second, for each version of the scenario, a control and experimental condition were developed. Order effects were mitigated by randomly assigning participants to different versions of the control and experimental, and different orders between them.

Bias Manipulation Design. To exploit the cognitive bias, a scenario was devised in which the attacker has gained access to a host on a company's corporate network and has an objective to access internal servers and exfiltrate key intellectual property (including source code and data). The attacker lands on Host A, and can see Server B and Server C. Each server has a path traversal vulnerability on them.

External inspection of the servers, through service/ version information, indicates that the servers are recently updated. One of the servers has a vulnerable service running that is recent, aligning well with the update status of the server (Apache 2.4.49 in

version 1; Bazaar in version 2). The other runs an outdated service that does not align with a freshly patched Windows 10 environment (BusyBox in version 1; Spring in version 2). BusyBox is a lightweight legacy utility not typically deployed on Windows servers, and the Spring version used (CVE-2014-3625) is over a decade old, both highly unlikely to be found in modern day, maintained systems. From a base rate perspective, the attacker should look at the update status of the servers, and conclude that the old software is suspicious, as it does not fit into the broader server status. However, it is anticipated that attackers who are susceptible to BRN will ignore that fact and focus on the old server as much (if not more) than the new server, as it will be viewed as a higher potential security gap.

Both servers were present in both the control and experimental conditions. In the *control*, the old server had only noise files that provided source data to explore but did not resemble the target the attacker was tasked to exfiltrate. In the *experimental* condition, the same noise files (~100 per server) were present along with ~20 distractors (e.g., data that looked like the target of interest, but were not actual targets). Distractors were placed in similar directory regions to the noise files, though were not perfectly collocated.

It was anticipated that attackers who are susceptible to BRN will spend more time exploring the server with the out-of-place vulnerability, particularly when distractor data is available on that server (experimental condition). Importantly, in this scenario, we are not aiming to create BRN in the

experimental condition; rather, *BRN comes into play regardless of the condition* but our belief is that participants who are susceptible to BRN would be more likely to continue to pursue well-designed distractor data that exploit that bias when such data are present on those servers.

Data Collection and Assessment. Bias indicators to detect susceptibility to BRN were implemented as a series of tools for extracting and interpreting data. These tools included (1) data collectors designed to collect a stream of potentially relevant data (e.g., all command shell activity, all PowerShell activity, etc.), (2) state abstractors that were designed to pull some pertinent state information out of a data collection stream (e.g., the count of stealthy commands within the command shell activity), and (3) bias indicators, which were intended to interpret the state information in context with adversary behaviors and infer bias susceptibility, opportunity, or context. Data collectors ran on the range, while the state abstractors and bias indicator ran on the collected data post-session. Further assessment of the outcomes to create indicators and output variables as described below.

Table 1 provides the list of bias indicators that were collected in the cyber range. Most indicators focused on observed behavior on the out-of-place server, as a direct indicator of susceptibility to BRN. We also collected an indicator of how time pressured the participant felt, as time pressure is a moderator that can impact susceptibility to BRN (as well as other heuristic-based biases). Table 2 provides the list of performance measures collected for the scenario.

Table 1: Bias indicators for BRN

Indicator	Measurement	Description and Use
Exploration Depth	Count of file access commands on the out-of-place server, and count the number of unique directory paths explored.	We anticipated more susceptible individuals will search a greater range of directories within the out-of-place server.
Exploit Density	Count of number of times the out-of-date exploit is called (total number of file access commands)	We anticipated more susceptible individuals will use the exploit more often to scan and exploit the out-of-place server.
Focus on Exploiting Out of Place Server	Count use of exploit using cURL for BusyBox, or Suricata for the other exploits, and compute ratio of commands exploiting the out-of-place server to those spent exploiting the updated server.	We anticipated more susceptible individuals will spend more activity focused on the out-of-place server.
Time-Pressured Behavior: Command Verbosity	Rate (per minute) of commands used by the attacker	Moderator literature shows heuristic-based biases (including BRN) are more prevalent when subjects are under time pressure.

Table 2: Performance measures for assessing BRN impacts

Name	Measurement	Description and Use
Time Wasted: Time Exploiting Out of Place Server	Σ time windows of out-of-place exploits to recent exploits.	We treat the measured time after using an exploit, to the time using a different exploit, as a time block where participants were focused on a given exploit. Participants more susceptible to BRN will spend more time accessing the out-of-place vulnerability than the recent vulnerability.
Rate of Attack Success: Target File Exfil	% of target files exfiltrated	Participants more susceptible to BRN will be less likely to get data from the recent vulnerability server, reducing rate of attack success.
Cognitive Effort: Distractor File Exfil	# of distractor files exfiltrated	Participants more susceptible to BRN will be more likely to download and try to interpret data from the out-of-place vulnerability server, creating more cognitive load.
Detectability: Noise File Exfil	# of noise files downloaded	Participants who are more susceptible to BRN will be more likely to download noise files, creating more overall traffic that can increase detectability.
Detectability: Interactions with Out of Place Server	# of interactions with the out-of-place server	Participants who are more susceptible to BRN will have more interactions with the out-of-place server. Given that this is a known lure that valid users should never interact with, each interaction increases detectability.

Demographic, Skill, Bias and Moderator Questionnaires. After screening and agreeing to informed consent, participants completed a survey including demographic questions, including age, gender, education, employment status, native language, and attacker skill/experience. The survey then administered a series of psychometric questionnaires which included validated instruments for assessing biases. Participants' baseline susceptibility to BRN was assessed with a standardized questionnaire and served as a validated benchmark against which indicator-based behavioral observations in the cyber environment could be compared. The four item assessment from Berthet (2021) was selected due to high reliability (Cronbach $\alpha = 0.70$). These questions featured classic BRN scenarios involving probabilistic reasoning under uncertainty, for example: *"In Greece, 15% of women marry before the age of 25. Katerina is a 23-year-old Greek woman who has been in a serious relationship for three years. What is the probability that Katerina will marry before she is 25?"*. The survey also included measures of other cognitive biases (confirmation bias, anchoring effects, sunk cost fallacy, loss aversion, and framing effects), and potential moderators such as the Cognitive Reflection Test (CRT; Toplak et al., 2014), Adult Decision-Making Competence (Bruine de Bruin et al., 2007), the short form PANAS (MacKinnon et al. 1999), the Big Five Inventory short form (Soto & John, 2017), as well as the General Risk Propensity Scale (Zhang et al. 2018). The presentation order of these metrics and assessments was pseudo-randomized to control for potential scale administration order effects.

Post-Session Survey. A post-session survey was administered immediately following each SimSpace session and included 5-point Likert-scale questions assessing task-related confusion and time pressure, both overall and at peak moments, and two open-ended questions. We omit the results for space.

Base Rate Survey. A base rate survey was administered (post-session) after a separate experiment at a later date, to assess participants perceived base rates regarding services. A single question was repeated across ten items, phrased as follows: "Of 1,000 Windows 10 servers with up-to-date patches you might currently encounter in typical use, how many would you estimate to have the following [item] : [Provide a number between 0 and 1,000 for each item]" The ten items included relevant, expected low base rate elements (e.g., outdated software and services used related to the BRN manipulation) and unrelated distractor items expected to be frequently encountered as a sanity and attention check. Items were presented in a randomized order.

2.3. Procedure

The study employed a within-subjects design. Following an initial two-hour initial session for consent, surveys, and familiarization with the experimental environment, participants then returned at later dates to complete each of two separate one-hour cyber operation sessions in SimSpace. The SimSpace session started with participants receiving an "operational brief", that provided task instructions based on the cyberattack task and goal of interest in the session within SimSpace. Participants received an introductory description of the task at hand that conveyed intelligence-based context of the tasks and

the mission goals. Salient features of a brief included: (1) the stage of the current attack at the start of the task, (2) important details about the computer network as could plausibly have been derived from prior attacker-gained intelligence about the network, and (3) a statement of the attacker's objective(s). The briefs did not constitute manipulation of any independent variables.

Participants then logged into the test bed to play the role of a cyberattacker and navigated through a network, performing reconnaissance, infiltrating, and attacking network nodes, with ultimate goals focused on achieving specific attack objectives (largely focused on data and intellectual property theft). Participants could take brief breaks at any time but were not allowed to pause the session. Participants were not allowed to bring/install their own hacker toolsets, previously created scripts, etc., for deployment on the test bed.

Once logged into the attacker virtual machine (their staging ground) they could use the cyberattack software tools we provided against the target network, as well as on the test bed. Participants worked independently from one another; each attacker participated alone within their own test bed instance. Participants were provided with multiple decision points throughout the network path, which allowed us to determine the presence (or absence) of cognitive bias and the impact of the biasing technique.

As described earlier, two versions of the scenario (V1 and V2) were developed along with the control (C) and experimental (E) conditions (manipulated within-subjects). This created four counterbalanced groups as the variation among the orders of four scenarios, with each participant completing one control and one experimental session, varying the version and order (two sessions per participant). After both SimSpace sessions were completed, a post-session survey was administered (we omitted results for space), and later the base-rate post survey. The timing ensured participants did not receive any cues or hints regarding the bias exploitation during their tasks, helping to preserve the ecological validity of attacker behavior.

3. Results

Where bias indicator or survey responses fell outside 5 SD from the mean, these values were removed (resulting in two removed data points from the indicator data overall, $n=33$ otherwise). When correlations are reported, given the expert population and low n , we generally report correlations $> .25$ even when they may not have reached the standard .05 criterion for significance. Marginal significance is

reported when p is less than or equal to .1 but greater than .05.

Base Rate Neglect EM. The results were scored for accuracy on a 0-100 scale. Participants ($n=33$) exhibited relatively low bias ($M = 78.03$, $SD = 31.1$) when compared to the results of Berthet (2021) but showed substantial variability.

Base Rate Neglect Indicator Validity and Effectiveness. Indicator validity was first assessed through comparison with the survey measure. To facilitate this comparison, indicator scores and the BRN survey measure were rescaled (i.e., normalized as a proportion of their respective maximum values). Two outliers (>5 SD from the mean) were removed before this analysis.

To test H1, we examined the correlation between the bias indicators and the established BRN survey measure. No correlation was significant across the relationships between the survey measure and the aggregated bias indicators (no $\rho > -.21$, or $p < .24$). Therefore, **H1 was not supported**. This leaves open several possibilities: on the one hand, the survey-based measures may not result in the bias due to lack of participant engagement in them or other factors; and on the other, it could be that the indicators really do not align, or that we simply lack sufficient power to detect meaningful correlations.

Spearman's rank-order correlations examined the relationship between aggregated indicator values and the differences between control and experimental outcome measures to determine whether indicator outputs meaningfully predicted differential behavioral responses to experimental conditions (H2). Here, we found several promising results. *Focus on exploiting out-of-place vulnerability* did significantly correlate with performance reductions from control to experimental conditions, specifically with the aggregate number of interactions with the old exploit ($\rho = .376$, $p = .031$). *Command verbosity* was marginally correlated with the change in noise file exfiltration from control to experiment, ($\rho = .311$, $p = .078$). *Exploration depth* marginally correlated with the number of distractor files exfiltrated in aggregate ($\rho = .294$, $p = .097$) and with the number of interactions with the old exploit ($\rho = .293$, $p = .098$). However, *Exploit density* did not correlate with performance reductions from control to experimental conditions (no $p < .197$).

Combined, the evidence **partially supports H2**; we do indeed see differential effects on meaningful outcomes based on our detection of BRN susceptibility through bias indicators. We caveat that these results must still be synthesized with the lack of clear relationship between the indicators and survey measure of BRN.

Finally, to evaluate whether baseline susceptibility to bias (as measured by BRN survey) predicted the magnitude of behavioral change in response to experimental manipulation (H3), Spearman correlations were computed between BRN survey scores and difference scores (experiment – control) for each outcome variable. None of the relationships were significant (all p values $< .14$, p values $> .28$), and therefore **H3 was not supported**.

3.1. Effectiveness in Degrading Performance

OHF technique effectiveness was measured by comparing control and experimental conditions using Wilcoxon signed-rank tests, as the data were non-normal and comparisons were within-subject. Because of the non-normality, the use of r as a measure of effect size was used, calculated as = test statistic (Z) / square root (# of cases). ‘ R ’ can be interpreted using the same approximate ranges as Cohen’s d .

We first assessed for any ordering effects of scenario variant, as well as condition ordering using Mann-Whitney U tests. Each was counterbalanced. Tests between the scenario variant orders on changes in performance revealed the order of variants exposure may have influenced the results of only one of the performance differences: the change in number of distractor files exfiltrated ($Z = -2.40$, $p = .016$) significantly differed between variants, showing higher values when participants experienced a ‘ $v_2 \rightarrow v_1$ ’ order versus the ‘ $v_1 \rightarrow v_2$ ’ order). For potential condition ordering, Mann-Whitney tests were conducted and also only revealed a single relationship; the number of noise files exfiltrated ($Z = -2.36$, $p = .018$), was affected by the condition order with significantly greater number for $C \rightarrow E$ groups versus $E \rightarrow C$. This could indicate a lingering effect of the experimental conditions.

Bearing in mind the above influence of variant and order was only on the number of distractor or noise files exfiltrated, **we observed several significant, and marginally significant effects on performance**. Specifically: *Distractor File Exfiltration* was significantly less in control (median = 0) versus experimental condition (median = 1; $Z = -3.67$, $p < .001$, $r = .45$). *Number of Interactions with Out-of-Place Vulnerability* were significantly fewer in control (median = 40) to experimental condition (median = 104; $Z = 2.147$, $p = .032$, $r = .27$). *Noise File Exfiltration* was significantly less in control ($M = .45$) to experimental condition ($M = 1.12$; $Z = -2.15$, $p = .032$, $r = .26$). The median rate was 0 in both conditions, however. *Time Exploiting* was increased from control (Med = 8.86) to experimental (Med = 13.59) conditions but only a marginally significant finding ($Z = -1.48$, $p = .14$, $r = .18$). However, *Target File Exfiltration* did not show a

decrease from control to experimental condition. ($Z = -.598$, $p = .550$, $r = .07$).

Base Rate Neglect Survey. A subset of participants ($n = 23$) completed a base rate survey, in which Apache, Bazaar, BusyBox, and Spring were of particular interest as rare. Apache ($M = 214$ out of 1000, Med = 150) indicated participants perceived it as having a low base rate, though with some variability. Bazaar was perceived as rarer ($M = 45$, Med = 10), and BusyBox and Spring followed similar patterns ($M = 120$ and 136; medians = 20 and 50, respectively), indicating the majority recognized their rare deployment on Windows servers.

As a manipulation check for data quality, participants also estimated base rates for services known to be *native* to Windows environments: Remote Desktop, Windows File Sharing, Microsoft Internet Information Server (IIS), and Microsoft Exchange. These were consistently rated much higher (e.g., medians = 800 out of 1000 for Remote Desktop and File Sharing), aligning with expectations and widespread use. Together this data raises confidence that our base rate manipulation used a genuinely ‘rare’ set of items according to the mental model of our participants.

4. Discussion

The study aimed to investigate the feasibility of exploiting Base Rate Neglect (BRN) as a defensive mechanism through the OHF framework. In particular, we sought to determine whether exploiting BRN could impair attackers’ performance by manipulating perceived risk and value through lures and distractor data. Given the novelty of examining BRN within cyber adversarial decision-making, our findings provide meaningful insights into the application of exploiting cognitive biases as a defensive strategy.

Bias Indicator Validity and Effectiveness. To assess indicator validity, we examined the correlation between bias indicator values and a validated psychometric instrument (Berthet, 2021). Contrary to expectation, H1 was not supported, as no significant correlation emerged between bias indicators and the survey-based BRN measure. This suggests a possible disconnect between behavioral manifestations of the bias and performance on the established survey measure, or it may reflect limitations in the sensitivity of the survey or the statistical power of the sample. To assess functional validity, we examined whether indicator values predicted changes in performance. Results offer partial support for our hypothesis (H2) that higher indicator values will predict greater performance declines under the bias manipulation. The exploitation of out-of-place vulnerability indicator further demonstrated the strongest evidence

of functional validity, and both exploration depth and command verbosity indicator still showed association with performance change (though weaker). **In general, our approach appears to result in degraded attacker performance.**

A puzzling finding was that our established BRN survey measure did not predict performance impacts. It is possible that the chosen EM, relying on accuracy across only four items, provided insufficient range given the high overall performance within our sample. The lack of correlation may reflect limited sensitivity of the measure or inadequate variability among participants. Yet, the practical performance change was encouraging and suggested OHF techniques are useful even if susceptibility is unavailable.

Effectiveness in Degrading Performance. The experimental manipulation demonstrated practical effectiveness. Participants significantly increased interactions with the out-of-place vulnerability, exfiltrated more distractor and noise files, and showed a marginal increase in time wasted under experimental conditions. These findings support H4, demonstrating that **exploiting BRN can degrade attacker performance and increase their detectability**, providing defenders opportunities for detection and disruption. A possible alternative explanation for the performance differences is the (slight) difference in the number of files between the two conditions. However, noise file availability was high and consistent across conditions, and yet noise exfiltration and time wasted both increased only when distractors were present. Moreover, bias indicators (ex. exploration depth, focus on out-of-place server) predicted these condition shifts, supporting the interpretation that BRN-like tendencies were exploited rather than a simple effect of file volume. Even so, we acknowledge that file content and distribution may have contributed to the effect.

Overall, using a highly skilled sample of cyber red team members and operating in a high fidelity realistic range, cyber attacker performance can be influenced by creating an OHF scenario exploiting the BRN cognitive bias. This is a fundamental advancement of the general OHF theory and strong evidence in ecologically valid and realistic settings for its efficacy. Importantly, the results demonstrate that OHF techniques exploiting cognitive vulnerabilities, such as the one presented here, can serve not only to delay attackers but also to actively shape and monitor their behavior by using these vulnerabilities as affordances that draw them into observable, controlled pathways.

5. Limitations

We had no female representation in our participant population. There is limited evidence that

gender influences representativeness biases (Ohlert & Weißenberger, 2015, found males showed *less* BRN than females; and so did Lucena et al., 2021). Our unique sample therefore suggested less susceptibility, but nevertheless there were still BRN impacts: our BRN EM did show reasonable variance, though no correlation with performance. Additionally, none of the bias indicators showed significant correlation with the EM of BRN, indicating that they did not (alone) reliably track individual differences in susceptibility.

This is not surprising; most bias studies aim to demonstrate the existence of a cognitive bias, rather than produce a range of potential susceptibility even if one is produced. This mismatch between group-level convergence and individual-level variation highlights a limitation about the bias indicators about whether they lack precision to differentiate between individuals (or attackers). Further research should explore whether aggregated indicator information allows individual tailoring for OHF disruptions. Lastly, only the experimental condition contained distractor files, while noise file volume was consistent across conditions. This asymmetry may limit direct comparison of distractor exfiltration and may partly explain increased noise file grabs; future work should separate file volume effects from bias-driven behavior.

6. Acknowledgements

We thank Brian Gzemski for scenario design and Christina Lewis, Tommy Tran, Grace Oswald, and Jasmine Vang for data collection and participant management. This research is based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA) under ReSCIND program. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government.

7. References

- Aggarwal, P., Nowmi, S. R., Du, Y., & Gonzalez, C. (2024). Evidence of Cognitive Biases in Cyber Attackers from An Empirical Study. *HICSS*, 934–943.
- Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, 44(3), 211-233.
- Bar-Hillel, M. (1990). Back to base rates. In *Insights in decision making: A tribute to Hillel J. Einhorn* (pp. 200–216). University of Chicago Press.
- Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in adult decision-making competence. *Journal of Personality & Social Psychology*, 92(5), 938.

- Cranford, E. A., Gonzalez, C., Aggarwal, P., Tambe, M., Cooney, S., & Lebiere, C. (2021). Towards a Cognitive Theory of Cyber Deception. *Cognitive Science*, 45(7).
- CrowdStrike (2025). *2025 Global Threat Report*.
- Cybersecurity and Infrastructure Security Agency. (2025). NICE Framework. <https://niccs.cisa.gov/workforce-development/nice-framework>
- Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3), 223-241.
- Ferguson-Walter, K. J., Major, M. M., Johnson, C. K., & Muhleman, D. H. (2021). Examining the efficacy of decoy-based and psychological cyber deception. In *30th USENIX Security Symposium*, 1127-1144.
- Ferguson-Walter, K., Shade, T., Rogers, A., Trumbo, M. C. S., Nauer, K. S., Divis, K. M., ... & Abbott, R. G. (2018). The Tularosa Study: an experimental design and implementation to quantify the effectiveness of cyber deception. *Sandia National Lab TR (SAND2018-5870C)*
- Gigerenzer, G., & Gaissmaier, W. (2011). Heuristic decision making. *Annual review of psychology*, 62(2011), 451-482.
- Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: frequency formats. *Psychological Review*, 102(4), 684.
- Gutzwiller, R. S., Ferguson-Walter, K. J., & Fugate, S. J. (2019). Are cyber attackers thinking fast and slow? Exploratory analysis reveals evidence of decision-making biases in red teamers. *Proceedings of the Human Factors and Ergonomics Society*, 63, 427-431.
- Gutzwiller, R. S., Lewis, C., Pharmed, R., et al. (In press). Oppositional Human Factors. In Fausett, Keebler, Lazzara, & Schuster (Eds.), *Handbook of Human Factors in Cybersecurity Systems*.
- Gutzwiller, R. S., Rheem, H., Ferguson-Walter, K. J., Lewis, C. M., Johnson, C. K., & Major, M. (2024). Exploratory Analysis of Decision-Making Biases of Professional Red Teamers in a Cyber-Attack Dataset. *Journal of Cognitive Engineering and Decision Making*, 18, 37-51.
- Gutzwiller, R. S., Ferguson-Walter, K., Fugate, S., & Rogers, A. (2018). "Oh, look, a butterfly!" A framework for distracting attackers to improve cyber defense. *Proceedings of the Human Factors and Ergonomics Society*, 62, 272-76.
- Hitaj, B., Denker, G., Tinnel, L., McAnally, M., DeBruhl, B., Bunting, N., Fafard..., & Starink, D. (2025). A case study on the use of representativeness bias as a defense against adversarial cyber threats. *2025 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*, pp. 680-689.
- Intelligence Advanced Research Projects Activity. (2023). ReSCIND: Reimagining Security with Cyberpsychology-Informed Network Defenses. *ODNI*. <https://www.iarpa.gov/research-programs/rescind>
- Johnson, C. K., Gutzwiller, R. S., Ferguson-Walter, K. J., & Fugate, S. J. (2020). A cyber-relevant table of decision making biases and their definitions. *ResearchGate TR*. <https://www.researchgate.net/publication/344106644>
- Johnson, C. K., Gutzwiller, R. S., Gervais, J., & Ferguson-Walter, K. J. (2021). Decision-making biases and cyber attackers. In *36th IEEE/ACM ASEW*, 140-144.
- Johnson, C. K., Van Tassel, R. W., Shade, T., Rogers, A., & Ferguson-Walter, K. (2024). Adversarial cognitive engineering (ACE) and defensive cybersecurity: leveraging attacker decision-making heuristics in a cybersecurity task. *HICSS*, 974-983.
- Kahneman, D. (2011). *Thinking, fast and slow*. New York, NY: Farrar, Straus and Giroux.
- Kahneman, D., & Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3(3), 430-454.
- Kahneman, D., & Tversky, A. (1973). Psychology of prediction. *Psychological Review*, 80, 237-251.
- 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. *Personality and Individual Differences*, 27(3), 405-416.
- Mckneely, J., Sell, T., Straub, K., & Thomas, D. (2023). Defensive Cyber Maneuvers to Disrupt Cyber Attackers. *HICSS*, pp. 6736-6745.
- Ohlert, C. R., & Weissenberger, B. E. (2015). Beating the base-rate fallacy: an experimental approach on the effectiveness of different information presentation formats. *Journal of Management Control*, 26(1), 51-80.
- Shade, T., Rogers, A., Ferguson-Walter, K., Elsen, S. B., Fayette, D., & Heckman, K. E. (2020). The Moonraker Study: an experimental evaluation of host-based deception. *HICSS*. 10.24251/HICSS.2020.231.
- Soto, C. J., & John, O. (2017). The next Big Five Inventory. *Journal of Personality & Social Psych*, 113(1), 117.
- Stanovich, K. E. (1999). Who is rational? Studies of individual differences in reasoning. Erlbaum.
- 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, 1275-1289.
- Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-31.
- Tversky, A., & Kahneman, D. (1981). Judgments of and by Representativeness (No. TR-3, Stanford University).
- Vang, J., & Revelle, M. (2024). Formalizing cognitive biases for cybersecurity defenses. *ACM SIGSAC*, 4991-4993.
- 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), 152-67.
- Zhao, C. (2018). Representativeness and similarity. Available at SSRN 3979419.