

Behavioral Red Teaming: A Systematic Literature Review of Human Vulnerability Assessment in Cybersecurity

Pranjal Upadhyay

School of IT, AI & Cyber Security, Rashtriya Raksha University, Gujarat, India.

To Cite this Article: Pranjal Upadhyay, "Behavioral Red Teaming: A Systematic Literature Review of Human Vulnerability Assessment in Cybersecurity", *Indian Journal of Computer Science and Technology*, Volume 05, Issue 02 (May-August 2026), PP: 868-878.



Copyright: ©2026 This is an open access journal, and articles are distributed under the terms of the [Creative Commons Attribution License](#); Which Permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract: A more recent approach for studying the human aspect of cybersecurity is known as behavioral red teaming, which is also gaining acceptance as a technique to exploit cognitive, psychological and human factors, rather than technical weaknesses in the system. While numerous research and examination projects have been conducted exploring the concepts of social engineering, insider threats and human vulnerability assessment, much of the research remains poorly organized and only a small number focus specifically on the creation and continual evolution of an exhaustive collection of behavioral elements for the redteam approach and the human security practices. This systematic literature review combines literature to determine the evolution of behavioral red teaming, how it is being measured, its psychological underpinnings and measure, and emerging trends. A systematic review methodology, as per the standard guidelines of systematic reviews, was used to analyze and classify the peer-reviewed studies provided in the top scientometric databases into areas of thematic domains. It provides some perspective on the typical scenarios under which behaviours can lead to attacks, on the structures for evaluating behaviours, cognitive theories that explain why humans are vulnerable to these behaviours, and on ways to record performance when evaluating humans against the cyber threat. It also points out the lacunae for both current and future research directions including lack of a common approach in assessing human behavior, relatively short-lived projects, and the insufficient use of artificial intelligence in behavioural assessment. Last, the review outlines future research opportunities to help advance the field of adaptive, evidence-informed, and people-centric organizational cybersecurity for future threats and challenges.

Keywords: Behavioral Red Teaming; Human Vulnerability Assessment; Human-Centric Cybersecurity; Social Engineering; Insider Threats; Systematic Literature Review.

I. INTRODUCTION

Behavioral red teaming is one of the most important advances of traditional cybersecurity testing and is a natural extension of knowledge and principles from traditional testing efforts but shifts the focus from technical vulnerabilities and threats to the human commensurables adversaries use every day. Red team conducts tests of the stability / robustness of networks, applications, end-points and/or organization infrastructure, and does it with the enemy on your side. Modern attacks, on the contrary, are not so much against the systems (or the most recent security technology) as against the individual who is targeted through such attacks by methods of psychological manipulation, deception and exploitation of trust to outsmart some of the newest security technology. The human factors are often the most vulnerable in a business organisation's security, whether used in social engineering attacks, phishing, spearphishing, business email compromise attacks, attacks from within, pre-texting or in the guise of artificial intelligence [1]. Behavioral Red Teaming provides a system to systematically engage employees with real world scenarios, stripping away the "glitter and glitz" from typical adversarial exchanges and leading into a dive down to the core of how they interpret, understand and engage in a situation in their working and thinking environment in which they have no control. This machine not only measures the levels of technical compromise, but also the users' understanding from a user context perspective, how users gain trust, what knowledge has been formed, what cognitive biases exist, whether or not users feel compelled to communicate and the compliance measures they have in place, and what choices they would make if they were unsure [2]. It bases itself on a number of domains such as Cyber Security, Psychology, Behavioral Science, Human Factor Engineering, Organisational Behaviour, Cognitive Science, understanding why people are susceptible to cyber threats [3]. The sophistication of AI models through being able to impersonate the identity of a person or a company, or, continuously allowing quicker changes to the approach when a new set of signals is received, has pushed their skills to understand that a believable-social engineering attempt is an attack—one that functions very similarly to a real communication [4]. This knowledge has become more specific than ever: organisations know they need to constantly assess technology and human behaviours to develop Cyber resilience. Besides being a security testing approach, behavioral red teaming is then developed as a security strategy to identify behavior of risk, promote security culture, mature cybersecurity preparedness of an organization and kick off a testing program for cybersecurity awareness based on evidence [5]. The model is also used to take control of trust, predispositions towards human, urgency and emotions when predicting the identification of flaws for organization risk detection before they occur. As opposed to the conventional awareness training techniques – which largely involve informational training – behavioral red teaming provides measurable impact from undertaking awareness training within realistic attack scenarios, enabling organizations to adjust security

exercises and facilitate the development of a culture of continuous improvement. With zero-trust architectures, AI based security operations and risk-based governance being the new path to safe guard and secure, a number of Businesses It are turning to behavioral evaluation as a part of their Enterprise security strategies. It is no coincidence that this has happened from the philosophy of traditional approach based on technology to today's approaches that include a broader aspect of cyber resiliency with human touch.

When it comes to treating humans in a mindset of human-centric cybersecurity practices, it is neither a consensus from the cybersecurity nor the psychology/behavioural science literature. The bulk of the research conducted is focused on one area, for example, phishing susceptibility, security awareness, cognitive bias or social engineering techniques—and didn't engage them in an overall behavioral measurement framework [6]. Contemporary literature provides a relatively small number of applied, systematic efforts related to adversarial simulations (behavioural), and a lot of technical efforts related to red teaming, penetration test, and vulnerability assessment. Additionally, even though social engineering has matured, the fast development of artificial intelligence, the large language models, conversational AI, synthetic identities, voice cloning, and deepfakes have paved the way for new types of threat vectors that can be used by perpetrators in a social engineering attack [7]. Research designs range from simple to complex while behavioral procedures, participants, and assessment instrument all can produce various outcomes that seek to compare them to other studies are different [8]. Other topics of contention include human subject ethics issues, privacy protection, informed consent issues, and organization transparency—the latter is only partly and the rest not reported in the literature. In addition, there is limited research on changes in behavioral vulnerabilities overtime, which has primarily focused on obtaining vulnerability across various sectors and/or across cultures [9]. Other attributes that can affect this development of the research area are the lack of any widely accepted taxonomies and behavioural measures, comprehensive lists of behavioural measures or any commonly adhered to taxonomy/measure for the maturity of the emerging field of behaviour. Therefore, cybersecurity professionals have no evidence-based guidance on how to design, implement and assess these Red Team exercises for their ability to best reflect real human-based cyber threats [10]. They have contributed to highlighting some of the knowledge gaps in the field of behavioural red teaming, and revealed the relative strengths and weaknesses of the various methodologies used in current/practicing cybersecurity.

Based on the identified gaps in these areas, the researcher wanted to create a pseudo-synthesis of the literature about behaviour red teaming and human Vulnerability Assessment in cyber security. It covers the behavioral red teaming and behavior based attack methods and assessment approaches reported in the literature, psychological frameworks and behavior measures used to measure susceptibility of humans to cyber threats and methodological limitations and future avenues of behavioral cyber related research, in a systematic review. Based on this, the research recommends the development of an organised taxonomy of behavioural red teaming, and outlines a range of future research directions for an AI-aided behavioural evaluation and human-centric cyber resilience. The paper particularly tackles the following objectives:

- Critically read and collate published academic research to build the understanding of the current knowledge and practice of behavioral red teaming and how it relates to human-centric cybersecurity.
- Identify, classify the behavioral attack techniques, methods of attacking human vulnerability, psychologic foundations and evaluation measures in the research conducted.
- To critically compare the latest trends, methodological issues, ethical concerns and emerging issues on Behavioural Red teaming.
- To create a thorough taxonomy that includes the key characteristics, methods and uses of behavioral red teaming and human vulnerability assessment.
- Seven AI-based behavioural modelling, standardisation of the assessment framework, adaptive security awareness model, explainable behavioural analytics and resilient human-centric security model research directions were identified.

II. RESEARCH BACKGROUND

The behavioral red teaming goes a long way beyond the technical aspect: It also represents a human and organisational approach to the effect of these factors on the security of an organization. For instance, organizations are now learning that, in addition to technical security evaluations, non-technical measures such as realistic attack scenario evaluations are required to overcome the onslaught of psychologically challenged cyber adversaries casting phishing, spear-phishing, business email compromise, pretexting and AI-driven deceptions over employees. The threat landscape has changed greatly over the past few years, with recent advancements in generative AI -- large language models (LLMs), deepfake, automated social engineering -- being able to orchestrate highly bespoke, large-volume and convincing attacks with little effort. Hence, this field of behavioural red teaming now is an interdisciplinary field dealing with the biomarkers of human susceptibility, decision-making processes, trust relations and security awareness, which integrates cyber security, cognitive psychology, behavioural science, organisational behaviours and artificial intelligence. Recent research on Social Engineering enhanced by Artificial Intelligence methods and, most notably, research that aims to model human behavior in Cyber Attacks, Cyber Physical systems security and in Human-centric Risk Assessment show how the success of cyber attacks remains heavily dependent on attacking human weaknesses. Concurrently, various international organisations such as NIST, ENISA and Verizon have emphasised the importance of human risk management in enterprise security solutions. Notwithstanding this progress, there is currently no standard taxonomy or evaluation framework and the literature is disparate, with different criteria for measuring evaluation, various experimental designs and behavioural measures. The current state of the territory identifies the need for a systematic literature review of present knowledge, methodological trends, methodological assessment of new behavioral attack methods implemented using AI and potential future avenues of evidence-based methods to consider human vulnerability in the field of cybersecurity.

Ref	Authors (Year)	Research Focus	Methodology	Major Findings	Identified Gap
[11]	Kumar & Patel (2025)	Social engineering attacks in the era of Generative AI	Comprehensive literature review	Demonstrated that generative AI significantly improves the scalability, personalization, and effectiveness of phishing and social engineering attacks.	Limited discussion on standardized behavioral assessment frameworks for organizations.
[12]	Kumar (2026)	Cybersecurity threats and challenges in Cyber-Physical Systems	Conceptual analysis	Highlighted the growing role of human behavior in protecting AI-enabled cyber-physical systems against evolving cyber	Human vulnerability assessment models for CPS remain underexplored.
[13]	Kumar, Deshkar & Patel (2025)	Fine-tuning language models for social engineering	Experimental feasibility study	Showed that fine-tuned language models can generate highly persuasive phishing content with improved contextual relevance.	Ethical safeguards and defensive behavioral evaluation techniques require further
[14]	Deshkar (2025)	Quantitative analysis of GenAI-facilitated social engineering threats	Mathematical and computational modeling	Proposed quantitative models for estimating the probability and effectiveness of AI-generated social engineering attacks.	Requires validation using real-world organizational behavioral datasets.
[15]	Kumar et al. (2025)	Security flaws in AI-powered Brain–Computer Interfaces	Security analysis and threat modeling	Identified multiple behavioral and AI-driven attack vectors affecting Brain–Computer Interface ecosystems.	Human-centric defense mechanisms for AI-enabled BCIs remain immature.
[16]	ENISA (2024)	European Cybersecurity Threat Landscape	Threat intelligence analysis	Reported significant growth in AI-assisted phishing, deepfakes, identity attacks, and social engineering campaigns.	Limited guidance on behavioral red teaming methodologies.
[17]	Verizon (2025)	Data Breach Investigations Report	Global incident analysis	Human error, credential theft, and phishing remain the dominant causes of organizational data breaches worldwide.	Standardized behavioral resilience metrics are lacking.
[18]	IBM X-Force (2025)	Threat Intelligence Index	Enterprise threat analysis	Identity-focused attacks and human-targeted cyber campaigns continue to increase across industries.	Few studies integrate threat intelligence with behavioral assessment frameworks.

[19]	Hadnagy (2018)	Psychology of social engineering	Behavioral security framework	Established the psychological principles underpinning successful human manipulation and social engineering attacks.	Does not address AI-driven behavioral attacks or modern adversarial technologies.
[20]	NIST Cybersecurity Framework 2.0 (2024)	Human-centric cybersecurity governance	Risk management framework	Emphasizes governance, awareness, organizational resilience, and continuous human risk management.	Does not provide a dedicated framework for behavioral red teaming or human vulnerability measurement.

Table 1. Recent Literature on Behavioral Red Teaming and Human Vulnerability Assessment

III. TAXONOMY OF BEHAVIORAL RED TEAMING

Behavioral Red Teaming (BRT) may be accomplished as several different integrated styles to address different vulnerabilities regarding humans in the scope of the organizations' CS. BRT somewhat resembles a collection of 'blokkop'-modelled drills, but it's also concerned with human conduct, human cognition, human communication, process and organisation — throwing a bit of shade, you would say. The taxonomy has also been updated with recently discovered and emerging developments in generative AI techniques, adaptive phishing, cognitive modelling techniques and behavioural analytics and the introduction of new techniques such as AI-assisted deception, using personalization in attack generation and a continuous, ongoing behavioral risk assessment. In recent years, modern BR has been combined with the concepts of cybersecurity, Psychology, social engineering, insider threat management, human-computer interaction and organization security to have a comprehensive assessment of human resilience to cyber threats that are not static [21]. Furthermore, factors other than the success rate of the attacks are relevant when calculating the human Cyber risk as proposed by Carman et al. [22] like relation of trust, decision making under uncertainty, policy compliance, security awareness and security culture. The number of strategies to consider when learning Behavioral Red Teaming (BRT) is overwhelming and when it comes to Zero Trust and Human Risk Management (HRM), the possibilities of strategy we can employ are endless, one of which is the proactive one mentioned above—Cyber Resilience. New Improvements such as Zero Trust and Human Risk Management (HRM) have added a new dimension to Behavioral Red Teaming (BRT) to anticipate some of the attacks by adversaries and uncover behavioural weaknesses [23]. In particular, having an easy-to-explain representation of BGs' support needs and of the human vulnerability assessment in the CB models can deliver better results in terms of the vulnerability assessment being accurate [24, 25].

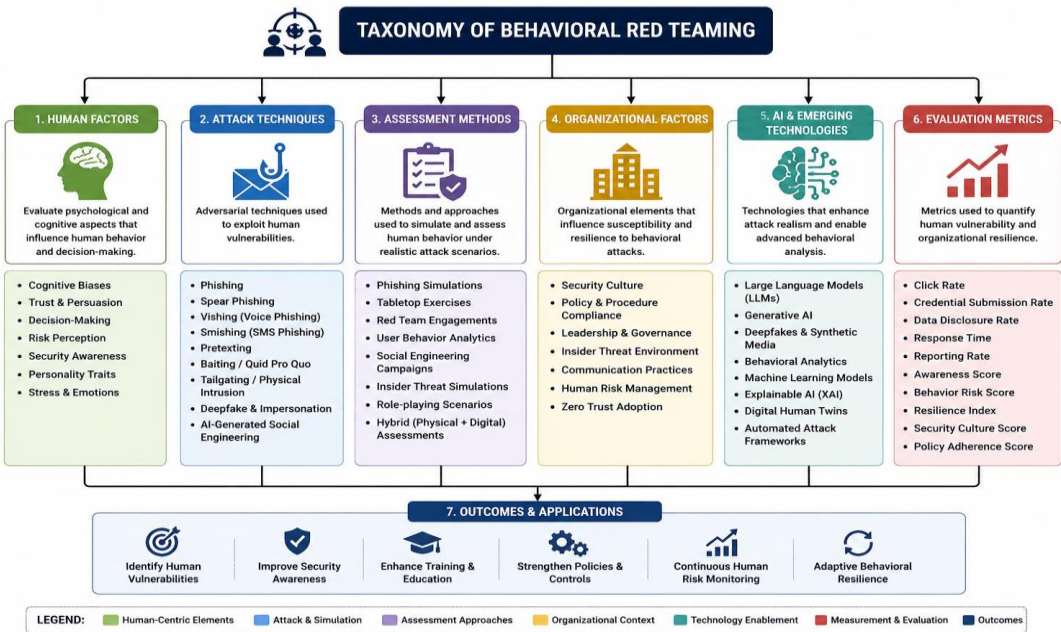


Figure 1. Taxonomy of Behavioral Red Teaming

The term "behavioural red teaming", in many cases, is the closest thing to real social engineering that acting as a "true" social engineer can come to. It is not just social engineering, but it incorporates aspects of human behavior, organizational security, attack method, AI and any measured evaluation metrics. Such numbers can be helpful to understand the whole human (cyber) threat domain and to set up a really realistic valid body of the threat/opponent to be studied or applied in practice. Besides, the taxonomy's emphasis on the effective behavioural assessment is multifaceted, and embraces the field of behavioural psychology, the cybersecurity engineering, AI based evaluation, organisational governance processes and organisational culture. The all-encompassing view will help ensure uniformity across behavior evaluations, human risk monitoring as well as set the benchmark to develop adaptive and responsive computer security training to keep pace with more sophisticated attacks fuelled by AI.

Category	Subcategories	Purpose	Representative Techniques
Human Factors	Cognitive biases, trust, decision-making, risk perception, security awareness	Assess human susceptibility to cyber threats	Behavioral profiling, cognitive assessment, awareness evaluation
Attack Techniques	Phishing, spear phishing, vishing, smishing, pretexting, baiting, deepfake attacks, AI-generated phishing	Simulate realistic adversarial scenarios	Email phishing, voice phishing, QR phishing, AI-generated deception
Assessment Methods	Phishing simulations, tabletop exercises, red team engagements, user behavior analytics, insider threat simulations	Measure organizational resilience against behavioral attacks	Controlled campaigns, behavioral observation, simulated insider attacks
Organizational Factors	Security culture, policy compliance, insider threat management, human risk management, Zero Trust	Evaluate organizational preparedness and governance	Compliance audits, security maturity assessment, HRM frameworks
AI & Emerging Technologies	Large Language Models, Generative AI, Explainable AI, behavioral analytics, digital human twins	Enhance behavioral attack simulation and risk prediction	AI-assisted phishing, adaptive attack generation, predictive analytics
Evaluation Metrics	Click rate, credential submission rate, response latency, awareness score, behavioral risk score, resilience index	Quantify employee performance and organizational cyber resilience	Risk scoring models, behavioral dashboards, resilience benchmarking

Table 2. Taxonomy of Behavioral Red Teaming

IV. HUMAN VULNERABILITY ASSESSMENT TECHNIQUES

Human Vulnerability Assessment (HVA) is one of the key pillars of Behavioral Red Teaming which involves the systematic process of identifying, quantifying and countering adversarial exploited behavioral flaws in the attack of a cyber asset. Traditionally the elements considered for vulnerability assessment are network configurations, infrastructure security, software vulnerabilities, and those examined in the human dimension of security are cognitive vulnerabilities, trust networks, decision making, communication and, when it comes to a realistic attack, the responses to security. Cyber threats are increasingly leveraging generative language models, AI, Deepfakes and even custom phishing attacks, and instead of spending their financial and time resources on one-off security awareness training and initiatives, organisations are investing in evolving and adaptable behavioural assessment methodologies, that are able to measure and track the extent to which a person's behaviour is moving toward the development of an attack [25]. Modernized assessment methods involve psychological instruments, adversarial simulation, behaviours, machine learning and algorithms to review the individual's response, perception and interpretation in deceptive cyber scenarios. Roger assesement is not only based on phishing click rates but also other behavioral components like reporting behavior, policy compliance, response latencies, disclosure of credentials, situational awareness, and risk perception [26]. Since the transactional data and patterns of users have high dimensionality, research indicates that using both Explainable Artificial Intelligence (XAI) and Entity and User Behavior Analytics (UEBA) would be essential to consistently monitor deviations in the users' behaviour, as well as identifying high-risk users before they reach the desired objective [27]. Furthermore, behavioral assessments are becoming more relevant in the assessment of security within the whole enterprise, leading to a shift from compliance-based security assessment to behavioral assessment as a key method of enterprise risk assessment, Zero Trust architecture and Human Risk Management (HRM) initiatives. This has resulted in Human Vulnerability Assessment (HVA) turning into an active and intelligence-driven mechanism to identify vulnerable individuals, deliver intelligence to inform security awareness adaptation, create vulnerability remediation strategies and continually develop the organization's cyber-resilience against growingly complex human-targeted attacks [28].

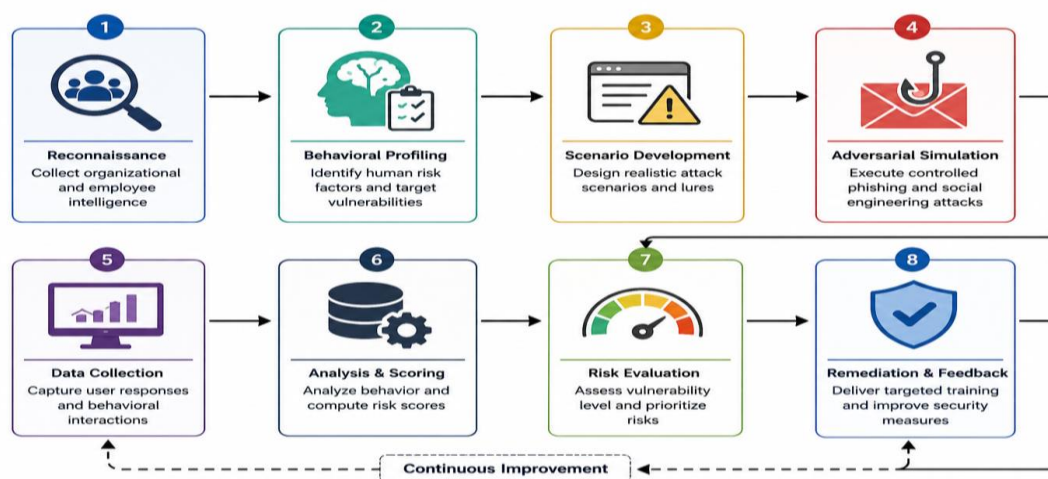


Figure 2. Human Vulnerability Assessment Workflow

The Human Vulnerability Assessment workflow is not a single event in the security workflow; it's a process. It starts with reconnaissance of the organization: workings done for the purpose of identifying the occurrences of persons and events, or the characteristics and activities of people or the physical and mechanical properties of a place that may be of interest to other organizations or the public and that can serve as valuable sources of information. This intelligence is then applied to discover the behavioral risks that are present via trust relationships, cognition patterns, work workflows and security processes in use. These observations can then be put to use in the creation of realistic attack scenarios that can accurately represent current threat landscapes such as attacks based on AI-generated phishing pages, insider threats and impersonation attacks and voice phishing, among others [29]. In controlled execution procedures, reactions are carefully monitored in order to get information about the participants' observable behavior, e.g. whether correct decisions are made, reaction times, whether they are susceptible to deception, whether they adhere to rules of the organization or not, and so on. Finally, behaviour data is analysed statistically, using behavioural analytics and AI models, to develop numerical human risk scores, and to define organisational patterns within vulnerabilities [30]. The iterative process does not rely, as typical awareness assessment does, on training results to measure awareness. Human Vulnerability Assessment provides helpful insights for organizations to adjust awareness campaigns, strengthen their security policies and strategies, build quantitative human resilience metrics relevant to their cyber goals, and stay one step ahead on tactics that adversary and intelligent automation-based social engineering campaigns use today and may use tomorrow.

Assessment Technique	Description	Primary Objective	Typical Metrics	Advantages	Limitations
Phishing Simulation	Simulated phishing emails sent to employees to evaluate response behavior.	Measure phishing susceptibility and awareness.	Click rate, credential submission rate, reporting rate.	Easy to deploy and highly scalable.	Users may become accustomed to recurring simulations.
Spear-Phishing Assessment	Personalized phishing campaigns targeting specific individuals or departments.	Evaluate resistance to targeted attacks.	Response rate, data disclosure rate, detection accuracy.	Closely mimics real-world attacks.	Requires significant preparation and ethical oversight.
Vishing & Smishing Exercises	Voice- and SMS-based social engineering assessments.	Assess verbal and mobile communication security.	Call success rate, information disclosure rate.	Evaluates multiple communication channels.	Resource-intensive and time-consuming.
User Behavior Analytics (UBA)	Continuous monitoring of user activities using AI and machine learning.	Detect anomalous or risky user behavior.	Risk score, anomaly score, behavioral deviation index.	Enables continuous monitoring and early detection.	Privacy concerns and potential false positives.

Insider Threat Simulation	Simulates malicious or negligent insider activities.	Assess organizational resilience against insider risks.	Detection time, response effectiveness, containment rate.	Evaluates operational preparedness.	Complex to design without disrupting operations.
Tabletop Security Exercises	Scenario-based discussions involving stakeholders and incident response teams.	Evaluate decision-making and coordination.	Response time, procedural compliance, communication effectiveness.	Improves organizational readiness.	Does not replicate actual attacker behavior.
Security Awareness Assessment	Structured quizzes, simulated exercises, and knowledge evaluations.	Measure employee cybersecurity awareness.	Awareness score, knowledge retention, reporting behavior.	Supports targeted awareness programs.	Knowledge may not translate into secure behavior.
AI-Assisted Behavioral Simulation	Uses Generative AI and LLMs to generate adaptive and personalized attack scenarios.	Evaluate resilience against emerging AI-enabled social engineering attacks.	Behavioral resilience score, adaptive response rate, AI attack detection rate.	Highly realistic, scalable, and continuously adaptive.	Ethical concerns, computational cost, and lack of standardized evaluation frameworks.

Table 3. Human Vulnerability Assessment Techniques in Behavioral Red Teaming

V. PSYCHOLOGICAL FOUNDATIONS

Behavioral Red Teaming relies on human behavioral psychology such as information processing, judgement formation, relationship building, and response to uncertainty in online interactions [31]. Behavioral Testing is more than a technology-driven approach to security testing, as it seeks to identify how an attacker or intruder is attempting to alter a user's behavior by leveraging his or her cognitive and emotional processes or systems. The threats today are limitless or they are crafted on a basis of perceiving user trust in authority, urgency, reciprocity, familiarity, social proof, fear and curiosity to cause a reaction putting users at risk of compromising organisation security [32]. These patterns shape our attention and perception of risk, and at times, even override rational thinking, even for security-minded people. Therefore, when implementing psychological lessons in the context of Behavioral Red Teaming, an organization can simulate an attack environment near the reality of a social engineering attack scenario, and tested human resistance in addition to technical resistance, to see how humans react to the attack scenarios [33]. In more recent years, other research has also demonstrated the potential improvements that can be made to the design of adaptive adversarial simulations by incorporating cognitive models, behavioural economic analysis, and HC theories, which can take into account the users' behaviour and awareness, and account for the variability of organizational cultures [34]. The need for gaining insight into human cognition has become even more important with the advent of generative AI and the generation of attacks in personalized fashion to identify high-risk persons before being compromised [35]. So, today's Behavioral Red Teaming is as much about the theories underlying what the attackers do successfully as it is about designing specifically effective behavioral awareness raising, security tailored interventions, explainable behavioral analytics and continued human resource management (HRM) processes to increase long-term enterprise cyberresilience capture [36].

Psychological Theory	Core Principle	Application in Behavioral Red Teaming	Common Attack Scenario	Behavioral Indicators Assessed
Protection Motivation Theory (PMT)	Individuals adopt protective behavior based on perceived threat severity and coping capability.	Evaluates how users respond to security warnings, phishing alerts, and incident notifications.	Phishing campaigns with security-related lures	Threat perception, response behavior, protective actions
Theory of Planned Behavior	Behavioral intention is influenced by	Assesses security compliance and policy adherence among employees.	Password sharing, policy violations	Compliance intention, behavioral

(TPB)	attitude, subjective norms, and perceived behavioral control.			consistency
Cognitive Load Theory (CLT)	High mental workload reduces analytical thinking and increases susceptibility to deception.	Measures employee responses under stress or information overload.	Urgent phishing emails, fake IT support requests	Decision latency, error rate, cognitive overload
Dual Process Theory	Human decisions involve both intuitive (System 1) and analytical (System 2) thinking.	Evaluates whether users rely on instinct or deliberate verification during attacks.	Executive impersonation, invoice fraud	Verification behavior, impulsive decisions
Social Influence Theory	Human behavior is shaped by authority, social proof, and peer influence.	Simulates attacks exploiting organizational hierarchy and trusted relationships.	CEO fraud, authority-based phishing	Trust level, authority compliance
Prospect Theory	Decisions are influenced by perceived gains and losses rather than objective outcomes.	Designs attack scenarios exploiting fear of financial or operational loss.	Account suspension, payroll fraud	Risk perception, emotional response
Cialdini's Principles of Persuasion	Persuasion is driven by reciprocity, scarcity, commitment, authority, liking, and social proof.	Models realistic social engineering techniques used by attackers.	Fake rewards, urgent approvals, vendor impersonation	Persuasion susceptibility, trust response
Trust Theory	Trust develops through perceived credibility, familiarity, and reliability.	Assesses how employees verify identities before sharing information.	Vendor impersonation, business email compromise	Trust validation, identity verification
Behavioral Economics	Human decisions frequently deviate from rational models due to biases and heuristics.	Identifies predictable cognitive biases exploited during cyberattacks.	Investment scams, financial phishing	Bias susceptibility, decision accuracy
Human Factors Engineering	Human performance is influenced by system design, usability, and environmental conditions.	Evaluates the impact of interface design and workflow on secure behavior.	Fake login portals, deceptive web interfaces	Usability errors, interaction patterns

Table 4. Psychological Foundations Supporting Behavioral Red Teaming

VI. AI AND BEHAVIORAL RED TEAMING

Behavioral Red Teaming has been revolutionized by Artificial Intelligence (AI), allowing defense teams to enhance the assessment process and bring the most advanced levels of automation, personalization, and scalability to their human-centric cyberattacks. Generative AI, Large Language Models (LLMs), machine learning, and multimodal AI systems have the power to create highly convincing phishing emails, voice cloning attacks, deepmedia fake videos, fake IDs, and “context-aware” social engineering campaigns that closely simulate authentic organizational communications. AI-based Behavioral Red Teaming simulates threats by reacting to user activity, user profiles and context, as opposed to traditional attack simulations which operate in a pre-set scenario. In the defense segment, AI powers long term behavioural awareness by offering User & Entity Behaviour Analytics (UEBA), Anomaly detection, Explainable AI (XAI), Predictive risk scoring and adaptive awareness, to name a few, all helping to spot problematic users before they are compromised. AI has transformed Behavioral Red Teaming from a static security evaluation process to intelligent, data-driven, and continuously evolving human risk evaluation that can handle more sophisticated cyber threats.

To develop the AI-driven Behavioral Red Teaming workflow shown in Figure 3, the literature was reviewed, summarized, and provided in this way. It “showcases the potential role that Large Language Models (LLMs), Machine Learning (ML) and Generative AI can play in the area of behavioural intelligence such as user/user profiling and behavioural attacks and risk predictions.” All these capabilities enable a more detailed and comprehensive phishing and deceptive playbacks to be realized with a sense of personalisation, and enriched with AI that adds to the personalised social engineering playback. Routine configuration and daily monitoring of users and entities behavior, anomaly detection and behavior analysis of the users are used to assess users' behavior patterns, reporting behavior and compliance with policies. Collected data is then analysed to calculate Human risk level and organizational resilience that can be leveraged to design opportunities for awareness training and continuous audit and Human Risk Management (HRM) for human assets. This figure merges together the salient features and process steps that have been found in recent publications on AI powered Behavioral Red Teaming.

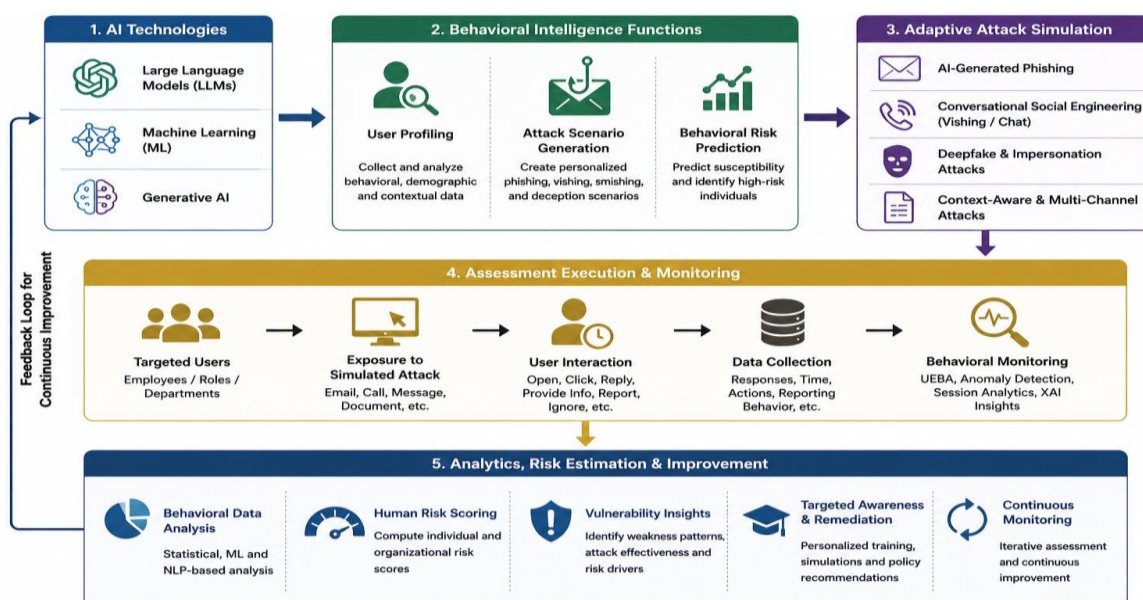


Figure 3. AI-Driven Behavioral Red Teaming Framework

AI Technology	Primary Function	Application in Behavioral Red Teaming	Technical Capabilities	Expected Benefits	Challenges
Large Language Models (LLMs)	Natural language generation	Generate realistic phishing emails, business email compromise (BEC) messages, and conversational attacks	Context-aware text generation, multilingual communication, adaptive dialogue	High realism, personalized attack simulation	Prompt manipulation, ethical concerns
Generative AI	Synthetic content generation	Create phishing emails, fake documents, impersonation	Text, image, and document synthesis	Scalable adversarial simulations	Potential misuse and governance issues

		messages, and adaptive social engineering content			
Deep Learning	Pattern recognition	Analyze behavioral responses and identify vulnerable user groups	Neural network-based behavioral classification	Improved prediction accuracy	Requires large labeled datasets
Machine Learning	Behavioral prediction	Predict phishing susceptibility and human risk levels	Classification, clustering, anomaly detection	Automated risk assessment	Model bias and false positives
User and Entity Behavior Analytics (UEBA)	Continuous monitoring	Detect anomalous user activities indicating compromised or high-risk behavior	Baseline modeling, anomaly detection, behavioral profiling	Continuous human risk monitoring	Privacy and compliance concerns
Explainable AI (XAI)	Model interpretability	Explain AI-generated behavioral risk scores and recommendations	Feature attribution, decision transparency	Increased trust and regulatory compliance	Limited maturity for complex models
Natural Language Processing (NLP)	Language understanding	Analyze communication patterns and detect phishing characteristics	Sentiment analysis, semantic similarity, entity extraction	Improved phishing detection	Language ambiguity and multilingual challenges
Reinforcement Learning	Adaptive optimization	Continuously optimize attack strategies and defensive training scenarios	Dynamic policy learning, feedback optimization	Adaptive behavioral assessments	Computational complexity
Digital Human Twins	Human behavior simulation	Model employee behavior for predictive vulnerability assessment	Behavioral modeling, scenario simulation	Predictive human risk analysis	High implementation complexity
AI Risk Scoring Engines	Human risk quantification	Compute dynamic behavioral risk indices using multiple behavioral indicators	Multi-factor scoring, predictive analytics	Continuous organizational risk visibility	Lack of standardized benchmarking metrics

Table 5. AI Technologies Supporting Behavioral Red Teaming

VII. CONCLUSION

Behavioral Red Teaming is an essential practice to learn in building cybersecurity resilience but it doesn't focus solely on finding technical flaws, rather it simulates the circumstances and observes how people respond. This SLR synthesized the available literature on HVA techniques, psychological underpinning, and improvements related to AI, cybersecurity, psychology, behavioral science and social engineering to get a nuanced, three-dimensional perspective on HVA techniques, their psychological foundations, and enhancements through AI. The report highlights the powerful cyber attacks of today, targeting the weaknesses in human nature through personalized phishing attacks and the use of deep fakes, and the importance of continual behavioural assessment and Human Risk Management (HRM). Even with all the advances, there is still need for uniform methods in measurement, uniform scales for observations, ethical issues and a longitudinal validation. The future research directions consist

of building integrated assessment frameworks, building explainable AI behavioral analytics and benchmark dataset, building adaptive assessment models, etc. Overall, it brings together existing knowledge, underlines the gaps in the field, and offers a starting point for the evolution of Behavioral Red Teaming as a vital element in today's cyber warfare.

REFERENCES

- Hadnagy, C. (2018). *Social engineering: The science of human hacking*. Wiley publ.
- Gass, R. H., & Seiter, J. S. (2022). *Persuasion: Social influence and compliance gaining*. Routledge.
- Wilson, M., & Hash, J. (2003). National Institute of Standards and Technology (NIST), Building an Information Technology Security Awareness and Training Program (NIST SP 800-50). *US Department of Commerce, Washington, DC*.
- Cichonski, P., Millar, T., Grance, T., & Scarfone, K. (2012). Computer security incident handling guide. *NIST Special Publication, 800(61)*, 1-147.
- Spring, J. M., & Illari, P. (2021). Review of human decision-making during computer security incident analysis. *Digital Threats: Research and Practice, 2(2)*, 1-47.
- RISK, C. (2021). Risk management framework. *COPRPORATE REPORT*.
- Landscape, E. T. (2023). European Union Agency for Cybersecurity (ENISA), 2023. URL: <https://www.enisa.europa.eu/topics/cyber-threats/threat-landscape> (date of access: 8.10. 2025).
- Kaczmarczyk, B. (2024). Cyberattacks/incidents and responding to them.
- Omotade, A. L., Ghimire, A., & Sultan, M. (2025, October). Cybersecurity, Data Privacy, and Information Security Management. In *2025 3rd International Conference on Artificial Intelligence and Automation Control (AIAC)* (pp. 27-33). IEEE.
- Nissinen, M. (2025). Strategy-Driven Information Security Framework for Enhancing Situational Awareness.
- Kumar, N., & Patel, N. M. (2025). Social engineering attack in the era of generative AI. *International Journal for Research in Applied Science and Engineering Technology, 13(1)*, 1737-1747.
- Kumar, N. (2026). Cyber Security Threats and Challenges in CPS. *AI and Cyber Security in Cyber-Physical Systems*, 53-78.
- Kumar, N., Deshkar, D., & Patel, N. (2025). Fine-Tuning Language Models for Social Engineering: A Technical Feasibility Study. In *Proceedings of the IEEE 7th International Conference on Computing, Communication and Automation (ICCCA)* (pp. 1-6).
- Deshkar, D. (2025). Quantitative and Computational Mathematical Analysis of Gen-AI-Facilitated Social Engineering Threats. *International Journal of Applied Mathematics, 38(10s)*, 2159-2179.
- Kumar, N., Deshkar, D., De, S., Saini, A., Kania, R. P., & Kasera, C. (2025). A Comprehensive Analysis of Security Flaws and Attack Vectors in Artificial Intelligence-Powered Brain-Computer Interfaces. *Vascular and Endovascular Review, 8(6s)*, 106-121.
- Davis, G. E. (2025). CYBER AND DATA PRIVACY INSURANCE IN 2025. *Brief, 54(2)*.
- LAKHTIONOVA, A. (2025). GLOBAL CYBERSECURITY MARKET 2017-2029: DYNAMICS, STRUCTURE, CHALLENGES AND KEY PLAYERS. *Society. Economy. Digitalization, 1(4)*, 61-75.
- Gazzola, P., Amelio, S., Pavione, E., & Marubini, F. (2026). The centrality of cybersecurity: business models, competitive strategies and sustainability practices. *Business Process Management Journal, 1-24*.
- Van Than, V. (2025). Future-Back Threat Modeling: A Foresight-Driven Security Framework. *arXiv preprint arXiv:2511.16088*.
- Konerza, C. (2025). Aviation Cybersecurity and Third-Party Software Service Providers: Do Companies Like CrowdStrike Get a Free Pass to Create Cyber Vulnerabilities?. *Journal of Air Law and Commerce, 90(4)*, 453.
- Mc Cabe, M., & Houmb, S. H. (2026). Enhancing the MITRE ATT&CK® Framework for Cyber-Physical Systems Using Insights from Advanced Persistent Threats. *Applied Sciences, 16(4)*, 1815.
- Pollini, A., Callari, T. C., Tedeschi, A., Ruscio, D., Save, L., Chiarugi, F., & Guerri, D. (2022). Leveraging human factors in cybersecurity: an integrated methodological approach. *Cognition, Technology & Work, 24(2)*, 371-390.
- Peca, L., & Tsurcanu, D. (2025). Reducing cyber risk through a human-centred approach. *Journal of Engineering Science, (1)*, 18-31.
- Aquisti, A., Brandimarte, L., & Loewenstein, G. (2015). Privacy and human behavior in the age of information. *Science, 347(6221)*, 509-514.
- Schneier, B. (2015). *Secrets and lies: digital security in a networked world*. John Wiley & Sons.
- Kahneman, D. *Thinking, Fast and Slow*. New York, NY, USA: Farrar, Straus and Giroux; 2011.
- Thaler, R. H., & Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- Furnell, S., & Thomson, K. L. (2009). From culture to disobedience: Recognising the varying user acceptance of IT security. *Computer fraud & security, 2009(2)*, 5-10.
- Workman, M. (2008). Wisecrackers: A theory-grounded investigation of phishing and pretext social engineering threats to information security. *Journal of the American society for information science and technology, 59(4)*, 662-674.
- Vishwanath, A., Herath, T., Chen, R., Wang, J., & Rao, H. R. (2011). Why do people get phished? Testing individual differences in phishing vulnerability within an integrated, information processing model. *Decision Support Systems, 51(3)*, 576-586.
- Cranor, L. F. (2008). A framework for reasoning about the human in the loop.
- Weatherald, N. A. (2022). Reinventing the blurry oval: Practitioner perceptions of deepfakes as a tool for anonymisation in documentary film and video journalism.
- Bonneau, J. (2012, May). The science of guessing: analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE symposium on security and privacy* (pp. 538-552). IEEE.
- Shostack, A. (2014). *Threat modeling: Designing for security*. John Wiley & sons.
- Sotira, N. (2018). The human factor in cyber security. *Cyber Security: A Peer-Reviewed Journal, 1(4)*, 326-330.
- Aljohani, M., Furnell, S., Carpent, X., & Gervassis, N. (2025, June). Examining the Influence of Cultural Diversity in Cybersecurity Culture. In *European Interdisciplinary Cybersecurity Conference* (pp. 151-165). Cham: Springer Nature Switzerland.