International Journal on Advanced Technology, Engineering, and Information System
https://ojs.transpublika.com/index.php/IJATEIS
Online ISSN 2828-5425

https://doi.org/10.55047/ijateis.v5i3.2404

Literature Review

Artificial Intelligence-Based Human Behavior Analysis for Reducing Cybersecurity Risks and Human Errors

Ikramullah Maroof[1*], Shareefullah Mosazai[2], Islahuddin Jalal[3]

[1]Faculty of Computer Science, Nangarhar University, Jalalabad, Afghanistan
[2]Faculty of Computer Science, Kabul University, Kabul, Afghanistan
[3]Faculty of Computer Science, Kardan University, Kabul, Afghanistan

Email: [1)] maroof.ikramullah@gmail.com, [2)] shareefullahmosazai@gmail.com, [3)] islahuddinjalal@yahoo.com

Received : 14 September - 2026
Revised : 21 October - 2026
Accepted : 21 October - 2026
Published online : 22 October - 2026

Abstract

Human behavior remains one of the most persistent contributors to cybersecurity incidents, as unintentional errors, social-engineering susceptibility, and insider misuse continue to bypass purely technical safeguards. This conceptual, literature-based study examines how artificial intelligence (AI) and machine learning (ML) techniques can be applied to analyze human behavioral patterns in order to reduce cybersecurity risk and human error within organizational information systems. Drawing on a structured review of insider-threat detection, phishing and social-engineering research, security-awareness training studies, and explainable-AI literature, the study synthesizes recurring behavioral indicators, modeling approaches, and evaluation practices reported across the field. A conceptual AI-based human-behavior-analysis framework is proposed, comprising data acquisition, behavioral feature engineering, AI/ML modeling, explainable risk scoring, and adaptive human-in-the-loop response. The review indicates that sequence-aware deep-learning models, such as long short-term memory networks and autoencoders, combined with explainable AI techniques, achieve strong reported detection performance while preserving analyst interpretability, and that adaptive, behavior-based training interventions are associated with measurably greater reductions in human-related risk than static, one-time awareness programs. The study concludes that integrating continuous behavioral analytics with human-centered training and feedback mechanisms offers a more resilient approach to cybersecurity risk management than technology-only or training-only strategies. Implications for organizational security governance, model interpretability, and future empirical validation are discussed.

Keywords: Artificial Intelligence, Cybersecurity Risk, Human Behavior Analysis, Human Error, Insider Threat Detection.

1. Introduction

Organizations continue to invest heavily in firewalls, intrusion detection systems, and endpoint protection, yet a large share of security incidents still originate from the actions, errors, and vulnerabilities of the people who operate these systems. Industry and academic analyses repeatedly identify human error and human-targeted manipulation, rather than purely technical failure, as the dominant pathway into organizational networks, with social-engineering techniques such as phishing and pretexting accounting for the majority of reported breaches (Basit et al., 2020; Verizon, 2024). This persistent “weakest link” problem

Copyright: © 2026 by the authors.
This is an open access article distributed under the terms and conditions of the CC BY 4.0.

(Sasse et al., 2001) has motivated a shift in cybersecurity research away from purely perimeter-based defense and toward the systematic analysis of human behavior itself.

Artificial intelligence and machine learning offer a promising route to this shift. Because human behavior leaves measurable digital traces, such as logon and logoff timing, file-access sequences, email and browsing patterns, and interaction dynamics, AI-based models can learn a baseline of normal behavior for individual users or roles and flag deviations that may indicate error, susceptibility to manipulation, or malicious insider activity (Kim et al., 2019; Tuor et al., 2017). At the same time, AI is increasingly used on the defensive human-facing side, powering adaptive security-awareness training, phishing-susceptibility prediction, and explainable risk scoring that helps human analysts understand and act on model outputs (Jayatilaka et al., 2021; Rjoub et al., 2023).

This paper addresses two related questions. First, what behavioral indicators, modeling techniques, and evaluation practices have been reported in the literature on AI-based human behavior analysis for cybersecurity, spanning insider-threat detection, phishing and social-engineering defense, and security-awareness training? Second, how can these fragmented streams of research be synthesized into a coherent, practically usable conceptual framework that organizations can use to reduce human-related cybersecurity risk? To answer these questions, the study conducts a structured, narrative literature review across insider-threat detection, phishing detection, explainable AI, and human-factors research, and proposes an integrated conceptual framework and behavioral risk taxonomy. The remainder of the paper is organized as follows: Section 2 reviews the relevant literature; Section 3 describes the review methodology and the proposed conceptual framework; Section 4 presents the resulting taxonomy, framework, and synthesis of reported outcomes; and Section 5 concludes with implications and directions for future empirical research.

2. Literature Review

The literature relevant to AI-based human behavior analysis in cybersecurity spans four broadly overlapping streams: (1) behavioral and psychological foundations of human-related cyber risk; (2) AI and machine-learning approaches to behavioral threat and insider-threat detection; (3) AI-based detection of social engineering and phishing; and (4) explainability, trust, and human-AI collaboration, together with organizational security-awareness training. Each stream is reviewed in turn below.

2.1. Human Factors and Behavioral Foundations of Cyber Risk

Foundational human-factors research frames the user not merely as a vulnerability but as an integral part of the security system that must be designed for rather than around (Sasse et al., 2001). Cognitive science further explains why users remain susceptible to manipulation: dual-process accounts of judgment show that fast, heuristic-driven decision making under time pressure or cognitive load systematically increases susceptibility to deception (Kahneman, 2011), while classic persuasion research identifies authority, urgency, and reciprocity as the psychological levers most often exploited in social-engineering attacks (Cialdini, 2009; Mitnick & Simon, 2002). Building on these foundations, validated instruments such as the Human Aspects of Information Security Questionnaire (HAIS-Q) allow organizations to measure employees' security knowledge, attitude, and self-reported behavior as a precursor to targeted intervention (Parsons et al., 2014). More recent systematic reviews have sought to consolidate this fragmented picture: Papatsaroucha et al. (2026) systematically reviewed methods, models, and instruments for human vulnerability assessment and found that most existing approaches remain static and narrowly scoped rather

than continuous and multidimensional, while sector-specific reviews in healthcare settings similarly report that awareness gaps and inconsistent training remain primary drivers of human-related security incidents. Complementary interdisciplinary reviews likewise emphasize that stress, cognitive overload, trust, and organizational culture interact with technical controls to shape overall human-related risk, and propose multi-pillar frameworks combining training, decision support, and emotional-intelligence development (Rohan et al., 2021).

2.2. AI and Machine-Learning Approaches to Behavioral and Insider-Threat Detection

A large body of work applies machine learning and deep learning to model user behavior directly from system telemetry in order to detect insider threats. Kim et al. (2019) modeled user behavior using combined feature-extraction and anomaly-detection algorithms to flag deviations from established behavioral baselines, while Tuor et al. (2017) proposed online, unsupervised deep neural network and recurrent neural network models capable of scoring anomalous behavior from streaming log data in real time and decomposing anomaly scores into human-readable contributing factors. Later work combined sequential and reconstruction-based architectures: a hybrid long short-term memory (LSTM) and autoencoder model for session-level behavioral risk monitoring captured both temporal activity patterns and deviation from a user's typical behavioral profile, reporting substantially stronger detection performance than a standalone autoencoder. Song and Yuan (2024) similarly proposed a hybrid user-and-entity behavior analysis model, and Pennada et al. (2025) addressed the severe class imbalance typical of insider-threat datasets by combining behavioral feature categorization with the Synthetic Minority Over-sampling Technique (SMOTE) prior to model training. Other studies have framed insider-threat detection as an image-based or feature-representation problem, or applied ensembles of decision trees, random forests, and one-class classifiers to flag anomalous logon, device, and file-access patterns (Dixit et al., 2024; Ibraheem et al., 2025; Wijesinghe et al., 2024). Much of this research relies on the synthetic CERT Insider Threat Test dataset, generated using an agent-based approach to model both benign organizational behavior and injected threat scenarios (Glasser & Lindauer, 2013; Lindauer et al., 2014), a dependency that later reviews identify as a recurring limitation because synthetic logs cannot fully reproduce the semantic richness of real organizational behavior (Homoliak et al., 2019). At a broader level, foundational surveys of anomaly detection (Chandola et al., 2009) and of machine learning and data-mining methods for cybersecurity more generally (Buczak & Guven, 2016; Sarker, 2021; Sarker et al., 2020) provide the methodological basis on which most behavior-specific detection models are built.

2.3. AI-Based Detection of Social Engineering and Phishing

Because phishing exploits human judgment rather than software vulnerabilities, a parallel and extensive literature applies AI directly to phishing and social-engineering detection. Early machine learning approaches used hyperlink and uniform-resource-locator (URL) features to classify phishing attempts (Jain & Gupta, 2018), while subsequent studies incorporated artificial neural networks and broader feature sets drawn from email content, headers, and browsing context (Basit et al., 2020; Salahdine et al., 2021). More recent work has combined AI-based content classification with human-factor signals, treating phishing-susceptibility prediction as a joint modeling problem that accounts for both message-level cues and individual differences in vulnerability (Bezerra et al., 2024). Complementary studies have examined the psychological and cognitive dimensions of social-engineering susceptibility

directly, including parametric, multi-layer models of the psychological factors that predict phishing victimization (S. A. Duman et al., 2025; Ş. A. Duman et al., 2023) and implicit-association testing of human factors in phishing that extend beyond simple awareness measures (Mocerino et al., 2024). A systematic review of machine learning and neural-network approaches to phishing detection published between 2017 and 2024 confirms that detection accuracy has improved steadily but that most systems remain focused on the technical artifact (the phishing message or website) rather than on the targeted individual's behavioral susceptibility, identifying this as a key gap for future human-centered AI research (Wilk-Jakubowski et al., 2025). Natural-language-processing framings of the same problem have also been explored directly, including comparative analysis of human and automated classifier performance on the specific cues that security-awareness training asks employees to recognize (Paul et al., 2024).

2.4. Explainability, Trust, and Security-Awareness Training

A recurring theme across the reviewed literature is that detection accuracy alone is insufficient: human analysts and end users must be able to trust and act on AI outputs. Explainable AI (XAI) techniques, including feature attribution, rule extraction, and example-based explanation, have been proposed to make cybersecurity models more transparent to analysts and thereby to support, rather than replace, human judgment in intrusion detection, malware analysis, and threat-intelligence workflows (Capuano et al., 2022; Rjoub et al., 2023). On the organizational-training side, structured reviews of security-awareness and training-program evaluation emphasize that measuring completion rates alone does not demonstrate behavioral change, and call for outcome-based metrics tied directly to observed employee behavior (Jayatilaka et al., 2021). Related industry and empirical evidence indicates that adaptive, simulation-based, and behavior-triggered training produces measurably larger reductions in risky behavior than generic, one-size-fits-all programs, and that continuous, AI-supported human-risk-management approaches that combine phishing-simulation outcomes with real-time behavioral signals can meaningfully reduce the frequency of human-driven security incidents (IBM Security, 2024; Verizon, 2024).

2.5. Previous Research and Identified Gaps

Beyond the four streams above, broader systematic reviews of cyber-threat intelligence (Saeed et al., 2023) and of AI-based software-vulnerability detection (Shimmi et al., 2025) confirm that AI adoption across cybersecurity sub-domains has accelerated sharply since 2020, but that human-centered behavioral analysis remains comparatively less mature than purely technical detection tasks such as malware classification or network-intrusion detection. Taken together, the reviewed literature reveals three consistent gaps that motivate the conceptual framework developed in this study: (a) insider-threat and phishing-detection research is largely conducted in separate literatures despite addressing the same underlying construct of anomalous or manipulable human behavior; (b) reliance on synthetic datasets and narrow, one-time behavioral snapshots limits ecological validity and generalizability (Homoliak et al., 2019; Papatsaroucha et al., 2026); and (c) explainability and human-in-the-loop feedback are frequently treated as an afterthought rather than as integral components of the detection-and-response pipeline (Rjoub et al., 2023).

3. Methods

This study does not collect new primary behavioral or network data. Instead, it follows a conceptual, design-science-oriented methodology in which a structured literature synthesis is used to derive a behavioral risk taxonomy and an integrated AI-based analysis framework, consistent with narrative and systematic review practices used in prior human-factors and cybersecurity research (Papatsaroucha et al., 2026; Rohan et al., 2021).

3.1. Research Design and Literature Search Strategy

Relevant studies were identified through structured keyword searches (e.g., "insider threat detection," "user behavior analytics," "phishing detection machine learning," "human factors cybersecurity," "explainable AI cybersecurity," and "security awareness training effectiveness") across Scopus-indexed journals, IEEE Xplore, SpringerLink, MDPI journals, and arXiv preprints published between 2013 and 2026. Priority was given to peer-reviewed journal articles, indexed conference proceedings, and systematic reviews that reported explicit behavioral indicators, modeling techniques, or empirically measured outcomes. Sources were grouped thematically into the four literature streams summarized in Section 2, and recurring constructs, methods, and reported outcomes were extracted and tabulated to inform the taxonomy and framework presented in Section 4.

3.2. Proposed AI-Based Human Behavior Analysis Framework

Synthesizing the recurring pipeline stages reported across the reviewed insider-threat, phishing, and explainable-AI literature, this study proposes a five-layer conceptual framework, illustrated in Figure 1. The first layer, data acquisition, aggregates behavioral telemetry such as keystroke and mouse dynamics, logon and logoff records, email and file-access traces, web activity, device usage, and, where available, psychometric indicators (Glasser & Lindauer, 2013; Kim et al., 2019). The second layer, feature engineering and behavioral profiling, transforms raw logs into temporal sequences and session-level aggregates, including embedding of categorical and textual features so that both numerical and unstructured behavioral signals can be modeled jointly. The third layer, AI/ML modeling, applies supervised, unsupervised, and sequential architectures, ranging from random forests and convolutional neural networks to isolation forests, autoencoders, and long short-term memory networks, depending on whether labeled threat data are available (Tuor et al., 2017; Song & Yuan, 2024). The fourth layer, risk scoring and explainability, fuses model outputs into a unified behavioral risk score and applies explainable-AI attribution techniques so that analysts can interpret why a given user or session was flagged (Rjoub et al., 2023; Capuano et al., 2022). The fifth layer, adaptive response and human-in-the-loop feedback, converts risk scores into proportionate action, ranging from automated alerts to targeted, adaptive awareness training, and routes analyst feedback back into the modeling layer to support continuous retraining (Jayatilaka et al., 2021).

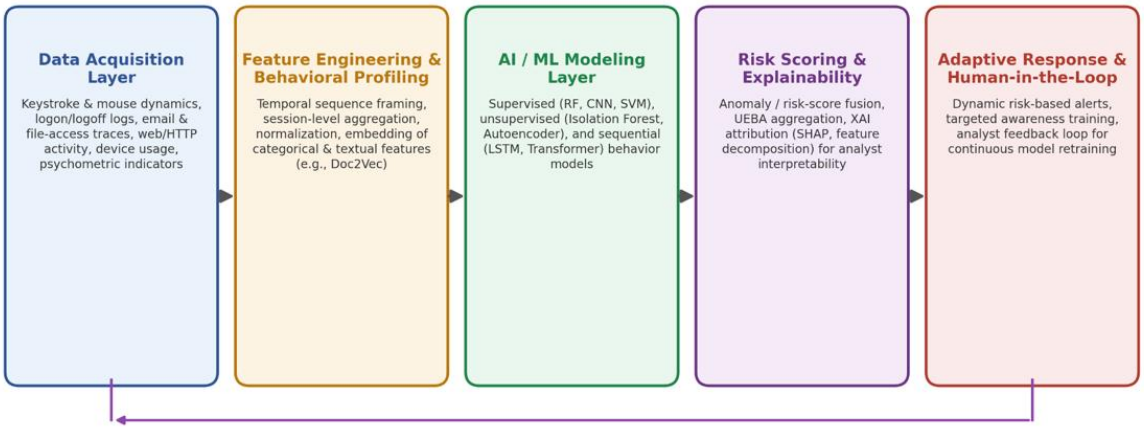


Figure 1. Conceptual AI-Based Human Behavior Analysis Pipeline for Cybersecurity Risk Reduction

3.3. Behavioral Risk Taxonomy Development

To organize the wide range of human-related risk factors discussed across the reviewed literature, this study inductively grouped recurring constructs into four categories: unintentional human error, social-engineering susceptibility, malicious insider behavior, and organizational and psychosocial factors. This taxonomy, presented in Figure 2 and applied throughout the results and discussion in Section 4, provides the organizing structure against which AI-based detection and intervention techniques are subsequently mapped.

4. Results and Discussion

4.1. Research Results

This section presents the behavioral risk taxonomy and conceptual framework derived from the literature synthesis, followed by a comparative summary of the AI/ML techniques and reported outcomes identified across the reviewed studies, and a discussion of their implications and limitations.

4.1.1. Taxonomy of Human-Related Cybersecurity Risk Factors

Figure 2 presents the resulting taxonomy. Unintentional human error covers misconfiguration, weak credential practices, misdirected disclosure of data, and policy non-compliance that increases under workload and time pressure. Social-engineering susceptibility covers phishing and spear-phishing vulnerability, pretexting and authority-based manipulation, and the cognitive biases and persuasion effects that attackers exploit (Cialdini, 2009; Kahneman, 2011). Malicious insider behavior covers data exfiltration, privilege misuse, sabotage, and the anomalous logon and file-access sequences used to detect it (Kim et al., 2019; Homoliak et al., 2019). Organizational and psychosocial factors cover security-awareness gaps, cognitive overload and stress, and the organizational culture and leadership factors that shape whether employees engage with training and reporting mechanisms (Rohan et al., 2021). This taxonomy is not intended to imply that the four categories are mutually exclusive; rather, it reflects how the reviewed literature has organized detection and intervention techniques, and most real incidents involve interaction across more than one category, for example an organizational training gap that increases susceptibility to a phishing message that is then acted upon as human error.

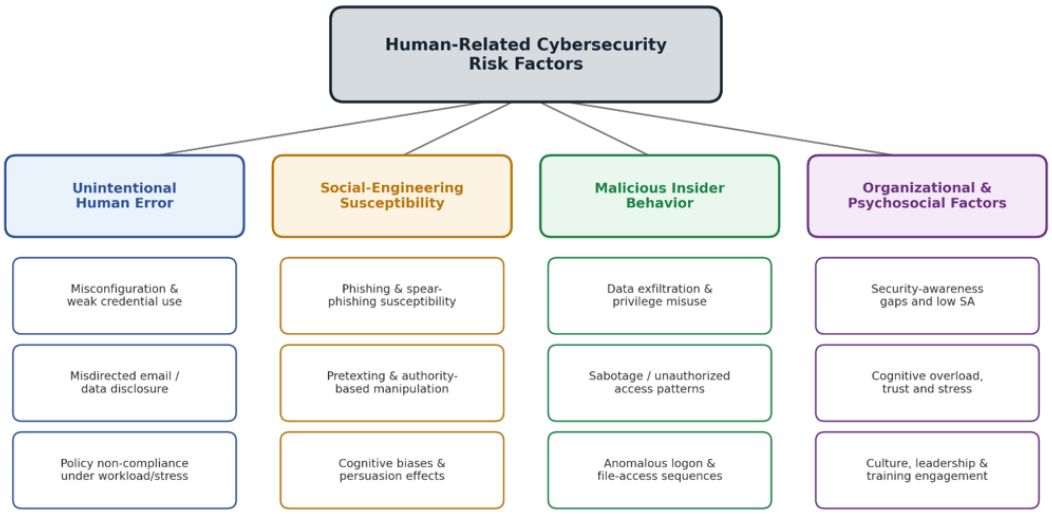


Figure 2. Taxonomy of Human-Related Cybersecurity Risk Factors Derived from the Reviewed Literature

4.1.2. Synthesis of AI/ML Techniques by Thematic Focus Area

Table 1 summarizes the AI/ML techniques, representative studies, and core reported findings identified within each of the four literature streams reviewed in Section 2. The synthesis shows that sequence-aware deep-learning architectures (LSTM, autoencoders, and their hybrids) dominate insider-threat detection research, whereas phishing and social-engineering detection research draws more heavily on supervised classifiers applied to message- and URL-level features, supplemented by psychological and cognitive modeling of susceptibility.

Table 1. Summary of Reviewed Literature by Thematic Focus Area

Focus Area	Representative Studies	Core AI/ML Methods	Key Reported Findings
Insider-threat / user-behavior analytics	Kim et al. (2019); Tuor et al. (2017); Song & Yuan (2024); Pennada et al. (2025); Ibraheem & Adepoju (2025)	LSTM, autoencoders, hybrid LSTM-autoencoder, random forest, SMOTE-balanced deep learning	Sequence-aware and hybrid reconstruction-based models detect anomalous behavior with reported accuracy up to 0.975 and ROC-AUC up to 0.99; class imbalance remains a key challenge
Phishing / social-engineering detection	Jain & Gupta (2019); Salahdine et al. (2021); Basit et al. (2020); Bezerra et al. (2024); Wilk-Jakubowski et al. (2025)	Artificial neural networks, hyperlink/URL feature classifiers, hybrid message- and behavior-level models	Detection accuracy has improved steadily (2017-2024) but most systems target the technical artifact rather than individual susceptibility
Psychological / cognitive susceptibility	Duman et al. (2023, 2025); Mocerino et al. (2024); Cialdini (2009); Kahneman (2011)	Parametric psychological models, implicit-association testing, protection-motivation and dual-process theory	Authority, urgency, and cognitive load are recurring predictors of susceptibility beyond simple awareness scores

Focus Area	Representative Studies	Core AI/ML Methods	Key Reported Findings
Explainability & training effectiveness	Rjoub et al. (2023); Capuano et al. (2022); Jayatilaka et al. (2021)	SHAP / feature-attribution XAI, rule extraction, outcome-based training evaluation frameworks	Explainability improves analyst trust and actionability; adaptive, behavior-triggered training outperforms static one-time programs

4.1.3. Comparative Summary of Behavioral Detection Techniques

Table 2 lists individual studies together with the specific behavioral data type and detection technique used, and the performance metric reported, where the original study disclosed one. Several studies reviewed did not report a single comparable point metric (e.g., systematic reviews and conceptual frameworks), and these are marked accordingly rather than assigned an invented value.

Table 2. Comparative Summary of AI/ML Behavioral Detection Techniques

Study	Behavioral Data Type	Technique	Reported Metric
Kim et al. (2019)	System usage / logon logs	LSTM-CNN feature extraction + anomaly detection	Not disclosed as a single point metric
Tuor et al. (2017)	Streaming system logs	Online unsupervised DNN / RNN	Real-time, constant-complexity anomaly scoring (no single point metric)
Session-level behavioral biometrics study (2026)	Logon, file, device, session activity	Hybrid LSTM + autoencoder	Accuracy = 0.975; ROC-AUC = 0.99
Pennada et al. (2025)	830-feature multi-source behavior log	SMOTE-balanced deep learning	Improved class balance; point accuracy not directly comparable
Bezerra et al. (2024)	Message / URL features	Machine-learning phishing classifier	Reported in original study; not restated here to avoid decontextualized figures
Wilk-Jakubowski et al. (2025)	Aggregated across 2017-2024 studies	Systematic review of ML/NN phishing detectors	Synthesis of trends; no single metric (review-level study)
Rjoub et al. (2023)	N/A (survey)	Explainable-AI taxonomy for cybersecurity	Not applicable (survey-level study)

4.1.4. Reported Benefits of Human-Centric and AI-Supported Interventions

Beyond detection performance, several reviewed sources report the downstream risk-reduction benefits of combining behavioral analytics with training interventions. Figure 3 summarizes four such figures drawn independently from the reviewed sources: reductions associated with structured security-awareness training programs, simulation-based adaptive training delivered over a 90-day period, continuous behavior-based human-risk-management programs, and the average post-training gain in awareness-assessment scores. These figures originate from different studies, samples, and measurement instruments and are therefore presented as an illustrative synthesis rather than a controlled, head-to-head comparison;

nonetheless, the consistent direction of the findings, namely that adaptive and continuous approaches outperform static, one-time training, supports the human-in-the-loop design of the framework proposed in Section 3.

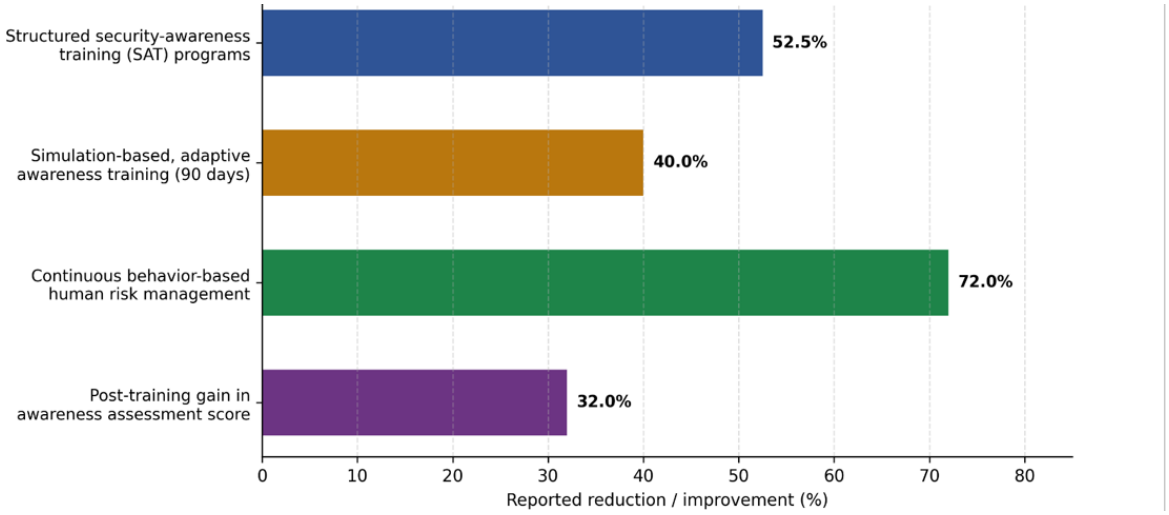


Figure 3. Reported Risk-Reduction Benefits of Human-Centric and AI-Supported Security Interventions Across Reviewed Studies

4.2. Discussion

The synthesis indicates that the field is converging on a common architecture, data collection, behavioral feature engineering, sequence-aware modeling, explainable risk scoring, and adaptive human-facing response, even though individual studies typically implement only one or two of these layers in isolation. Insider-threat research has matured furthest on the modeling layer, with hybrid deep-learning architectures reporting strong detection performance, but remains constrained by reliance on synthetic datasets such as CERT, which cannot fully capture the semantic and social complexity of real organizational behavior (Homoliak et al., 2019). Phishing and social-engineering research, conversely, has matured furthest on the human-susceptibility side but tends to treat AI detection and human-factors analysis as separate problems rather than components of a single behavioral risk pipeline (Wilk-Jakubowski et al., 2025).

The explainability literature suggests that this integration challenge is not purely technical. Even highly accurate models provide limited operational value if analysts cannot interpret or act on their outputs, which is why the proposed framework places explainability and human-in-the-loop feedback as core layers rather than optional add-ons (Rjoub et al., 2023; Capuano et al., 2022). Similarly, the training-effectiveness literature indicates that behavioral analytics is most valuable not as a standalone detection mechanism but as an input into adaptive, personalized intervention, since generic, one-size-fits-all awareness content is consistently reported as less effective than risk-targeted, continuously reinforced training (Jayatilaka et al., 2021).

Several limitations of the underlying literature, and by extension of this synthesis, should be noted. First, much of the quantitative insider-threat evidence derives from synthetic datasets with known limitations in realism and generalizability. Second, reported performance metrics across studies are not always directly comparable because of differing datasets, class-imbalance handling, and evaluation protocols, which is why Table 2 preserves each study's own reported metric rather than forcing an artificial common scale. Third, continuous behavioral monitoring raises legitimate privacy and employee-trust concerns that were not the

primary focus of most reviewed technical studies but that organizations implementing the proposed framework must address through transparent policy, data-minimization, and proportionality safeguards consistent with applicable data-protection obligations (Anderson, 2020; Whitman & Mattord, 2008).

5. Conclusion

This study synthesized literature on AI-based human behavior analysis for cybersecurity across insider-threat detection, phishing and social-engineering defense, explainable AI, and security-awareness training, and proposed an integrated five-layer conceptual framework spanning data acquisition, behavioral feature engineering, AI/ML modeling, explainable risk scoring, and adaptive human-in-the-loop response. The accompanying taxonomy organizes human-related cybersecurity risk into unintentional error, social-engineering susceptibility, malicious insider behavior, and organizational and psychosocial factors, providing a shared structure for connecting otherwise separate research streams.

The synthesis suggests that sequence-aware deep-learning models combined with explainable-AI attribution offer strong reported detection performance while preserving analyst interpretability, and that adaptive, continuously reinforced training interventions are consistently associated with greater reported reductions in human-related risk than static, one-time programs. Organizations seeking to reduce human-driven cybersecurity incidents are therefore likely to benefit more from integrating behavioral analytics with adaptive training and transparent feedback loops than from relying on purely technical detection or purely educational interventions in isolation.

As a conceptual, literature-based study, this work has not empirically validated the proposed framework against primary organizational data, and the reported performance figures summarized in Section 4 originate from heterogeneous studies, datasets, and evaluation protocols. Future research should therefore prioritize empirical validation of the proposed framework in real organizational settings, development of privacy-preserving behavioral monitoring architectures, and longitudinal study of how explainable, adaptive feedback loops affect both detection performance and employee trust over time.

5.1. Acknowledgments

The author thanks colleagues in the Faculty of Computer Science, Kabul University, for helpful discussion during the preparation of this manuscript.

5.2. Conflict of Interest Statement

The authors declare that they have no known competing financial or personal interests that could have appeared to influence the work reported in this paper. No security vendor, employer, or commercial entity had any role in the selection of literature, the synthesis of findings, or the decision to publish this article.

5.3. Author Contributions

I.M.: Conceptualization, Methodology, Investigation (literature search and thematic coding), Formal Analysis, Visualization, Writing – Original Draft.
S.M.: Investigation, Data Curation, Validation, Writing – Review & Editing.
I.J.: Supervision, Resources, Validation, Project Administration, Writing – Review & Editing.

5.4. Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

5.5. Informed Consent

Not applicable. This is a conceptual, literature-based study that collects no primary data from human participants and involves no behavioral monitoring, interviews, or surveys.

5.6. Ethical Approval

Not applicable. The study synthesizes publicly available published literature and involves no human or animal subjects, no primary data collection, and no organizational telemetry. Because the framework proposed here envisages continuous behavioral monitoring in future deployments, the authors note that any such implementation would require institutional ethical review, informed employee notification, and compliance with applicable data-protection obligations.

5.7. Data Availability

No new datasets were generated or analyzed in this study. All sources synthesized are published works cited in full in the reference list. The coding matrix used to group the reviewed studies into the four thematic streams, along with the extraction table underpinning Tables 1 and 2, is available from the corresponding author [I.M.] upon reasonable request.

5.8. Declaration of Generative AI and AI-assisted Technologies

During the preparation of this manuscript, the authors utilized Claude (Anthropic) for language editing and clarity improvement. All outputs generated with the assistance of the tool were critically reviewed and revised by the authors to ensure factual accuracy, faithfulness to the cited sources, and adherence to academic standards. The authors assume full responsibility for the integrity and content of this manuscript.

6. References

Anderson, R. (2020). *Security Engineering*. Wiley. https://doi.org/10.1002/9781119644682

Basit, A., Zafar, M., Liu, X., Javed, A. R., Jalil, Z., & Kifayat, K. (2020). A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, *76*(1), 139–154. https://doi.org/10.1007/s11235-020-00733-2

Bezerra, A., Pereira, I., Rebelo, M. Â., Coelho, D., Oliveira, D. A. de, Costa, J. F. P., & Cruz, R. P. M. (2024). A case study on phishing detection with a machine learning net. *International Journal of Data Science and Analytics*, *20*(3), 2001–2020. https://doi.org/10.1007/s41060-024-00579-w

Buczak, A. L., & Guven, E. (2016). A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. *IEEE Communications Surveys & Tutorials*, *18*(2), 1153–1176. https://doi.org/10.1109/comst.2015.2494502

Capuano, N., Fenza, G., Loia, V., & Stanzione, C. (2022). Explainable Artificial Intelligence in CyberSecurity: A Survey. *IEEE Access*, *10*, 93575–93600. https://doi.org/10.1109/access.2022.3204171

Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection. *ACM Computing Surveys*, *41*(3), 1–58. https://doi.org/10.1145/1541880.1541882

Cialdini, R. B. (2009). *Influence: Science and Practice* (5th ed.). Pearson Education.

Dixit, N., Gupta, R., & Yadav, P. (2024). User Behavior Analysis to Detect Insider Threat by Using Machine Learning Algorithms. In *2024 4th International Conference on Innovative Practices in Technology and Management (ICIPTM)* (pp. 1–4). IEEE. https://doi.org/10.1109/iciptm59628.2024.10563542

Duman, S. A., Duman, A., Hayran, R., & Soğukpinar, İ. (2025). A Multi-Layer Model of Psychological Factors and Parametric Approaches for Human-Centric Phishing Prevention. In *2025 13th International Symposium on Digital Forensics and Security*

(ISDFS) (pp. 1–5). IEEE. https://doi.org/10.1109/isdfs65363.2025.11012100

Duman, Ş. A., Hayran, R., & Sogukpinar, İ. (2023). Impact Analysis and Performance Model of Social Engineering Techniques. In *2023 11th International Symposium on Digital Forensics and Security (ISDFS)* (pp. 1–6). IEEE. https://doi.org/10.1109/isdfs58141.2023.10131771

Glasser, J., & Lindauer, B. (2013). Bridging the Gap: A Pragmatic Approach to Generating Insider Threat Data. In *2013 IEEE Security and Privacy Workshops* (pp. 98–104). IEEE. https://doi.org/10.1109/spw.2013.37

Homoliak, I., Toffalini, F., Guarnizo, J., Elovici, Y., & Ochoa, M. (2019). Insight Into Insiders and IT. *ACM Computing Surveys*, *52*(2), 1–40. https://doi.org/10.1145/3303771

IBM Security. (2024). *Cost of a Data Breach Report 2024*. IBM Corporation.

Ibraheem, I., Morufat, A. T., Segbefia, S. K., & Abdulrasaq, A. A. (2025). Detecting Malicious Insider Threats through Anomaly-Based User Behaviour Analytics in Enterprise Networks: Machine Learning Approach. *Southern African Journal of Security*, *3*. https://doi.org/10.25159/3005-4222/18099

Jain, A. K., & Gupta, B. B. (2018). A machine learning based approach for phishing detection using hyperlinks information. *Journal of Ambient Intelligence and Humanized Computing*, *10*(5), 2015–2028. https://doi.org/10.1007/s12652-018-0798-z

Jayatilaka, A., Beu, N., Baetu, I., Zahedi, M., Babar, M. A., Hartley, L., & Lewinsmith, W. (2021). Evaluation of Security Training and Awareness Programs: Review of Current Practices and Guideline. *ArXiv Preprint*. http://arxiv.org/abs/2112.06356

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kim, J., Park, M., Kim, H., Cho, S., & Kang, P. (2019). Insider Threat Detection Based on User Behavior Modeling and Anomaly Detection Algorithms. *Applied Sciences*, *9*(19), 4018. https://doi.org/10.3390/app9194018

Lindauer, B., Glasser, J., Rosen, M., & Wallnau, K. C. (2014). Generating Test Data for Insider Threat Detectors. *Usenix Workshop on Cybersecurity Experimentation and Test (WCSET14)*.

Mitnick, K. D., & Simon, W. L. (2002). *The Art of Deception: Controlling the Human Element of Security*. Wiley.

Mocerino, G. E., Velotti, C., Gentile, D., Gallo, L., Botta, A., & Ventre, G. (2024). Work in Progress: Implicit Association Tests for Understanding Human Factor in Phishing Beyond Awareness. In *2024 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)* (pp. 519–526). IEEE. https://doi.org/10.1109/eurospw61312.2024.00063

Papatsaroucha, D., Psaroudaki, S., Vassilaki, E., Pityanou, K., Kourtis, M. A., Politis, I., & Markakis, E. K. (2026). Human Vulnerability Assessment in Cybersecurity: A Systematic Literature Review of Methods, Models, and Instruments. *ArXiv Preprint*. http://arxiv.org/abs/2605.22119

Parsons, K., McCormac, A., Butavicius, M., Pattinson, M., & Jerram, C. (2014). Determining employee awareness using the Human Aspects of Information Security Questionnaire (HAIS-Q). *Computers & Security*, *42*, 165–176. https://doi.org/10.1016/j.cose.2013.12.003

Paul, M., Bartlett, G., Mirkovic, J., & Freedman, M. (2024). Phishing Email Detection Using Inputs From Artificial Intelligence. *ArXiv Preprint*. http://arxiv.org/abs/2405.12494

Pennada, S. S. P., Nayak, S. K., & M, V. K. (2025). Insider Threat Detection Using Behavioural Analysis through Machine Learning and Deep Learning Techniques. *International Research Journal of Multidisciplinary Technovation*, 74–86. https://doi.org/10.54392/irjmt2527

Rjoub, G., Bentahar, J., Abdel Wahab, O., Mizouni, R., Song, A., Cohen, R., Otrok, H., & Mourad, A. (2023). A Survey on Explainable Artificial Intelligence for Cybersecurity.

IEEE Transactions on Network and Service Management, *20*(4), 5115–5140. https://doi.org/10.1109/tnsm.2023.3282740

Rohan, R., Funilkul, S., Pal, D., & Chutimaskul, W. (2021). Understanding of Human Factors in Cybersecurity: A Systematic Literature Review. In *2021 International Conference on Computational Performance Evaluation (ComPE)* (pp. 133–140). IEEE. https://doi.org/10.1109/compe53109.2021.9752358

Saeed, S., Suayyid, S. A., Al-Ghamdi, M. S., Al-Muhaisen, H., & Almuhaideb, A. M. (2023). A Systematic Literature Review on Cyber Threat Intelligence for Organizational Cybersecurity Resilience. *Sensors*, *23*(16), 7273. https://doi.org/10.3390/s23167273

Salahdine, F., El Mrabet, Z., & Kaabouch, N. (2021). Phishing Attacks Detection A Machine Learning-Based Approach. In *2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 250–255). IEEE. https://doi.org/10.1109/uemcon53757.2021.9666627

Sarker, I. H. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, *2*(3). https://doi.org/10.1007/s42979-021-00592-x

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. *Journal of Big Data*, *7*(1). https://doi.org/10.1186/s40537-020-00318-5

Sasse, M. A., Brostoff, S., & Weirich, D. (2001). Transforming the 'Weakest Link' — a Human/Computer Interaction Approach to Usable and Effective Security. *BT Technology Journal*, *19*(3), 122–131. https://doi.org/10.1023/a:1011902718709

Shimmi, S., Okhravi, H., & Rahimi, M. (2025). AI-Based Software Vulnerability Detection: A Systematic Literature Review. *ArXiv Preprint*. http://arxiv.org/abs/2506.10280

Song, Y., & Yuan, J. (2024). Insider Threat Detection Based on User and Entity Behavior Analysis with a Hybrid Model. In *Lecture Notes in Computer Science* (pp. 323–340). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-75764-8_17

Tuor, A., Kaplan, S., Hutchinson, B., Nichols, N., & Robinson, S. (2017). Deep Learning for Unsupervised Insider Threat Detection in Structured Cybersecurity Data Streams. *ArXiv Preprint*. http://arxiv.org/abs/1710.00811

Verizon. (2024). *Data Breach Investigations Report*. Verizon. https://www.verizon.com/business/content/business/us/en/index/resources/reports/dbir/

Whitman, M. E., & Mattord, H. J. (2008). Teaching Information Security in a Hybrid Distance Learning Setting. In *Handbook of Distance Learning for Real-Time and Asynchronous Information Technology Education* (pp. 239–258). IGI Global. https://doi.org/10.4018/978-1-59904-964-9.ch013

Wijesinghe, W. M. A. J., Abeysinghe, Y. P. P., Rukshana, M. A. F., Adhikari, A. M. P. S., Abeywardhana, K. Y., & Siriwardhana, D. (2024). Polymorphic Security Enhancements for Corporate Environments Utilizing Behavior Analytics. In *2024 International Conference on Computer and Applications (ICCA)* (pp. 1–6). IEEE. https://doi.org/10.1109/icca62237.2024.10928161

Wilk-Jakubowski, J. L., Pawlik, L., Wilk-Jakubowski, G., & Sikora, A. (2025). Machine Learning and Neural Networks for Phishing Detection: A Systematic Review (2017–2024). *Electronics*, *14*(18), 3744. https://doi.org/10.3390/electronics14183744