



International Journal of Research Publication and Reviews

Journal Homepage: www.ijrpr.com ISSN 2582-7421

The Rise of Automated Cybercrime: How AI is Redefining Cyber Threats

Harshita Gautam¹, Ritu Dagar², and Dr. Manju Rani³

¹ Student, School of Engineering and Technology, Apeejay Styta University, Gurugram, Haryana, India

² Student, School of Engineering and Technology, Apeejay Styta University, Gurugram, Haryana, India

³ Assistant Professor, School of Behavioural Sciences, Apeejay Styta University, Gurugram, Haryana, India

ABSTRACT

Artificial intelligence (AI) is changing the economics and operational possibilities of cybercrime, but it is important to distinguish automated activity from activity specifically enabled by AI. This paper examines AI-assisted and automated cyber threats, with attention to phishing and social engineering, credential theft, ransomware, reconnaissance, and vulnerability exploitation. It uses an exploratory mixed-source design: a questionnaire completed by 36 university students is interpreted alongside published evidence from IBM X-Force, the World Economic Forum (WEF), the European Union Agency for Cybersecurity (ENISA), Verizon, and peer-reviewed research. In the student sample, 32 respondents (88.9%) reported awareness of AI-based cyber threats; 19 (52.8%) reported receiving suspicious or potentially phishing-related messages; 22 (61.1%) believed AI makes cybercrime more dangerous; and the most common self-rated cybersecurity-awareness score was 3 out of 5 (47.2%). External reports describe an 84% year-on-year increase in emails delivering infostealer malware observed by IBM in 2024, while WEF reported that 66% of surveyed organisations expected AI to have a significant cybersecurity impact and 37% had processes for assessing AI-tool security before deployment. These figures come from different samples and measurement frameworks and are not directly comparable. The paper argues that AI is best understood as an amplifier of selected attack capabilities rather than a universal or sole cause of cybercrime. It recommends layered controls, identity protection, secure AI adoption, verification procedures, and practical cybersecurity education. Because the student sample is small and self-reported, the primary findings are descriptive and exploratory rather than generalisable or causal.

Keywords: artificial intelligence, automated cybercrime, generative AI, phishing, social engineering, cybersecurity awareness, credential theft

1. Introduction

Cybercrime has evolved alongside the growth of networked computing, cloud services, digital identity systems, and online commerce. Automation has long enabled repetitive actions at scale, including scanning, credential testing, malware distribution, and botnet coordination. Recent generative AI systems add capabilities for producing fluent text, synthesising audio and video, summarising publicly available information, and assisting with coding and analysis. These capabilities can be misused, but their presence does not establish that an attack is fully autonomous or that AI caused its success.

This paper distinguishes three concepts. Automated cybercrime uses scripts or workflows to execute repeated tasks with limited manual intervention. AI-assisted cybercrime uses machine-learning or generative systems to support one or more tasks, such as content creation, prioritisation, or analysis. Autonomous cyber operations imply a greater degree of independent decision-making across a sequence of actions; claims of autonomy therefore require stronger evidence than a report that an attacker used an AI tool. This distinction is consistent with the need to assess actual observed behaviours rather than infer AI involvement from an incident's sophistication alone.

Evidence from industry and public-sector reporting points to persistent identity and social-engineering risks. IBM X-Force reported that emails delivering infostealer malware increased by 84% in 2024 compared with 2023 in its observed telemetry; it also identified identity abuse as a prominent initial-access pattern (IBM, 2025). Verizon's 2025 Data Breach Investigations Report reported compromised credentials as an initial access vector in 22% of breaches reviewed in that analysis (Verizon, 2025). These statistics reflect different datasets and definitions and should not be combined into a single global estimate.

Research is also beginning to systematise how generative AI may amplify social engineering. Schmitt and Flechais (2024), in a systematic review, identify realistic content creation, advanced targeting and personalisation, and automated attack infrastructure as three principal areas in which generative AI may amplify social-engineering activity. ENISA (2024) similarly discusses AI-assisted tailored phishing and synthetic media as emerging concerns. Yet AI can also strengthen defensive capabilities, including anomaly detection, phishing classification, and security operations. The central question is therefore not whether AI is inherently offensive or defensive, but how it changes capability, scale, risk, and organisational readiness.

The present study contributes an exploratory student survey and places its results alongside published cybersecurity evidence. It does not claim to measure national cybercrime prevalence, prove that reported suspicious messages were AI-generated, or establish that AI increases attack success rates.

A further problem is evidential attribution. In public discussion, the presence of generative AI is sometimes treated as proof that an incident was AI-driven. Yet incident reports often describe outcomes, initial-access vectors, or attacker behaviours without establishing which tools were used to produce them. A rigorous analysis must therefore separate (a) observed incident patterns, (b) plausible AI-enabled mechanisms, and (c) demonstrated causal effects. This distinction is central to the interpretation of both the published evidence and the present survey.

The paper makes three bounded contributions. First, it distinguishes conventional automation, AI assistance, and autonomous operation as analytically different concepts. Second, it triangulates a small student survey with selected peer-reviewed literature and institutional or industry reports, while explicitly retaining each source's denominator and scope. Third, it translates the synthesis into a layered response model that connects technical controls, governance, and human behaviour. These contributions are exploratory and integrative; they should not be read as a claim that the survey is representative or that the paper offers a comprehensive systematic review.

2. Literature Review

2.1 *From automation to AI-assisted cybercrime*

The distinction also matters for measurement. An organisation may observe a large number of phishing emails without knowing whether they were generated by a language model, copied from a template, or written manually. Conversely, an AI tool may be used during reconnaissance without leaving a clear artefact in the final incident record. For research purposes, AI attribution should therefore be based on direct evidence, transparent analyst methodology, or clearly labelled uncertainty rather than inference from linguistic polish or attack sophistication.

The broader literature on cybercrime-as-a-service shows that specialisation and the division of labour can lower barriers to participation. AI may add another service layer—such as content generation, translation, or analysis—without replacing existing criminal infrastructure. This suggests a socio-technical interpretation: the effects of AI depend not only on model capability but also on criminal markets, access to compromised accounts, payment systems, hosting, and the ability to evade defensive controls.

Traditional automation reduces the labour needed for repetitive tasks, while AI can additionally support pattern recognition, language generation, classification, and summarisation. These categories overlap but should not be treated as interchangeable. A botnet may be automated without AI; a phishing email may be written by a generative model and then sent manually; and an AI-enabled security product may use machine learning defensively. The systematic review by the authors of “The Weaponization of Artificial Intelligence in Cybersecurity” synthesises 21 studies and identifies scalability, social-engineering enhancement, evasion, and vulnerability-related activity among reported themes, while also noting the need for further research (The Weaponization of Artificial Intelligence in Cybersecurity, 2024).

A useful analytical distinction is between capability and outcome. The capability to generate many plausible messages does not itself demonstrate that more recipients click, disclose credentials, or transfer money. Outcome depends on targeting, the recipient's context, security controls, reporting behaviour, and the attacker's operational choices. Consequently, empirical claims about effectiveness require experiments, incident data, or carefully defined operational metrics.

2.2 *Generative AI, phishing, and social engineering*

From a human-factors perspective, phishing effectiveness can be understood through the interaction of message cues, situational pressure, trust, and the recipient's opportunity to verify. A highly fluent message may remove one warning sign but leave other indicators, such as a mismatched domain, unusual payment instruction, unexpected attachment, or request to bypass established procedure. Training should consequently teach a verification process rather than a brittle checklist of grammatical errors.

A useful distinction is between persuasion quality and operational success. An attacker can create a persuasive message yet fail because a secure email gateway blocks it, multifactor authentication prevents account takeover, or the recipient reports it. Research evaluating AI-enabled social engineering should measure the full chain of outcomes—delivery, opening, clicking, credential submission, reporting, and confirmed compromise—where ethically and legally appropriate, rather than treating message quality as a proxy for harm.

Phishing attempts to persuade recipients to disclose information, open a malicious attachment, visit a deceptive website, or perform another action that benefits an attacker. Social engineering is broader: it exploits trust, authority, urgency, fear, reciprocity, or familiarity. Generative AI may reduce the effort required to create grammatically fluent messages, adapt tone or language, and personalise content using publicly available information. Synthetic audio and video can also support impersonation. These developments may weaken older detection heuristics that rely on spelling errors or obviously generic wording.

Schmitt and Flechais (2024) organise the potential amplification of social engineering around realistic content, advanced targeting, and automated infrastructure. ENISA (2024) notes that generative AI can support tailored phishing and deepfake-based deception. These sources establish credible threat pathways, not a universal rate of AI-driven attack success. The existence of a synthetic voice or polished email should prompt verification, but appearance alone cannot reliably establish whether a communication is authentic or AI-generated.

Human factors remain central. People often make rapid decisions under time pressure, divided attention, or perceived authority. Security education should therefore avoid framing phishing as simply a failure of intelligence or care. It should create usable verification routines, clear reporting routes, and organisational norms that make it acceptable to pause and confirm unusual requests. A systematic review of phishing-mitigation strategies covering 248 articles concluded that human abilities and behaviour must be considered alongside technical controls; technology-only approaches are insufficient (Naqvi et al., 2023).

2.3 Credential theft and identity abuse

Identity-centric defence is particularly important because authentication events can connect otherwise separate incidents. Monitoring for impossible travel, unusual device or session characteristics, repeated failed sign-ins, abnormal mailbox rules, and unexpected privilege changes can help detect misuse after credentials have been obtained. Controls should be proportionate and privacy-aware: monitoring needs defined purposes, restricted access, retention limits, and escalation procedures.

Credentials provide access to email, cloud services, institutional platforms, and business systems. Attackers may obtain them through phishing, infostealer malware, password reuse, credential stuffing, or compromised third-party services. Once valid credentials are obtained, malicious activity may resemble legitimate user behaviour and evade controls that focus only on malware signatures. IBM X-Force reported an 84% year-on-year increase in emails delivering infostealers during 2024 and highlighted identity abuse in its incident dataset (IBM, 2025). Verizon reported compromised credentials as an initial access vector in 22% of breaches reviewed in its 2025 DBIR analysis (Verizon, 2025).

These data support prioritising phishing-resistant multi-factor authentication where feasible, unique passwords managed through approved password managers, conditional access, least privilege, and monitoring for anomalous sign-in patterns. They do not show that AI caused all credential theft, nor do they measure the same underlying population.

2.4 Ransomware and cybercrime-as-a-service

Resilience also requires treating recovery as a tested capability rather than the mere existence of backup files. Organisations should verify that backups are isolated from ordinary administrative credentials, restoration procedures work under realistic conditions, and business owners understand recovery priorities. Incident exercises should include communication, legal and regulatory decision-making, third-party dependencies, and the possibility that data exfiltration accompanies encryption.

Ransomware can disrupt access to systems and data, interrupt services, and expose confidential information. Ransomware-as-a-service distributes roles across developers, affiliates, and supporting service providers, lowering the need for each participant to build an entire operation independently. Automation can support repetitive delivery and execution tasks, but incident-level workflows vary. It is inaccurate to assume that every ransomware incident uses AI or is fully autonomous.

ENISA's Threat Landscape 2024 analysed more than 11,000 publicly reported incidents and identified threats to availability, ransomware, and threats to data among its principal categories (ENISA, 2024). This illustrates the continued importance of operational resilience. Defensive priorities include offline or immutable backups, tested restoration, network segmentation, timely patching, endpoint monitoring, privileged-access controls, and rehearsed incident-response plans.

2.5 Reconnaissance and vulnerability exploitation

For vulnerability management, risk-based prioritisation is more useful than a raw count of discovered weaknesses. Relevant factors include whether a weakness is exposed to the internet, whether exploitation is known or observed, the criticality of the affected asset, compensating controls, and the potential impact of compromise. AI-generated summaries may assist analysts, but the underlying advisory, asset configuration, and patch status should be checked against authoritative sources.

Reconnaissance involves collecting information about potential targets; vulnerability exploitation involves taking advantage of a weakness in software, configuration, or infrastructure. Automated scanning predates modern generative AI. AI may help summarise technical information, correlate findings, or prioritise potential weaknesses, but generated assessments can be incomplete and require validation. A vulnerability's presence does not mean that it is exploitable in a particular environment.

Google Cloud's M-Trends 2025 reported that exploits represented 33% of initial infection vectors in Mandiant's investigations of targeted attack activity during 2024, while stolen credentials represented 16% (Kutscher, 2025). This is a vendor incident-response dataset, not a census of all attacks. The finding reinforces the need for vulnerability management as well as identity security.

2.6 AI for cyber defence

AI-supported defence should be evaluated against operational baselines. Useful measures include precision, recall, false-positive rate, detection latency, analyst workload, and performance on previously unseen data. Accuracy alone may be misleading when malicious events are rare, because a model can

appear accurate by predicting the majority class. Human analysts should understand the limits of model output and retain a route to challenge or override automated decisions.

AI can assist defenders with alert triage, anomaly detection, malware classification, threat-intelligence summarisation, and phishing detection. A systematic review of AI-based algorithms for detecting social-engineering attacks screened 891 records and selected 32 studies; it reported high accuracy for some evaluated algorithms, including Random Forest and XGBoost (Atuncar Flores et al., 2024). Such figures should not be interpreted as universal real-world performance: accuracy depends on dataset composition, class balance, evaluation design, and how well a model generalises to new attacks.

The defensive use of AI also creates risks. Models may produce false positives, miss novel behaviours, expose sensitive data, or be manipulated through adversarial inputs. WEF reported that 66% of surveyed organisations expected AI to have a significant cybersecurity impact in the coming year, but only 37% reported having processes to assess the security of AI tools before deployment (World Economic Forum [WEF], 2025). This gap supports governance, testing, access control, data minimisation, and human oversight.

2.7 Conceptual synthesis

The conceptual model developed here can be expressed as a sequence of conditional relationships: AI capability may affect the cost, speed, realism, or personalisation of a task; automation may increase repetition or scale; exposure and defensive controls shape whether the activity reaches a target; human and organisational responses shape the probability and impact of harm. This is a conceptual synthesis derived from the literature, not a causal model tested by the 36-person survey.

The model also implies a feedback loop. Defensive systems may use AI to identify patterns, while attackers may adapt messages or workflows in response to detection. This creates a dynamic environment in which static awareness materials and one-time tool deployments can become outdated. Periodic review, incident learning, and controlled evaluation are therefore essential.

The reviewed literature supports a conditional model: AI may amplify selected attack capabilities through content generation, personalisation, analysis, and workflow assistance; automation may then increase repetition or scale. Whether this results in harm depends on access to data and infrastructure, attacker objectives, target exposure, human decisions, and defensive controls. The same broad AI capabilities can also support detection and response. This paper therefore treats AI as a capability multiplier whose effect is context-dependent, not as a single independent cause of cybercrime.

Table 1. Distinguishing automation, AI assistance, and autonomy

Concept	Operational meaning	Illustrative example	Evidence caution
Automation	Predefined actions executed repeatedly by scripts or workflows	Automated scanning or bulk message delivery	Does not imply machine learning or generative AI
AI-assisted activity	AI supports one or more tasks, with human or conventional workflow control	Generating message variants or summarising technical information	Tool use alone does not prove attack success or full autonomy
Autonomous operation	A system independently selects and executes multiple actions within a defined objective	An agentic workflow operating with limited human intervention	Requires evidence about decision authority, scope, and human oversight
AI-enabled defence	AI supports detection, classification, prioritisation, or response	Flagging suspicious messages for analyst review	Performance depends on data quality, evaluation, and deployment context

2.8 Research Gap and Contribution

The literature establishes credible mechanisms through which generative AI may support social engineering and other cyber tasks, while threat reports document continuing problems involving credentials, ransomware, exploitation, and organisational preparedness. However, three gaps remain relevant to this exploratory study. First, public reports do not always identify whether AI was actually used in a given incident, limiting causal attribution. Second, awareness surveys often measure perceived familiarity rather than demonstrated response performance. Third, organisational incident metrics and student self-reports operate at different levels of analysis and should not be merged into a single estimate. The present paper addresses these issues by explicitly separating evidence types, presenting the student findings descriptively, and deriving practical recommendations without claiming representativeness or causality.

3. Methodology

The methodology is intentionally described as an exploratory mixed-source study rather than a systematic review. The manuscript supplied for revision did not contain a documented database search strategy, date range for literature retrieval, screening flow, duplicate-removal procedure, or formal quality-appraisal protocol. It would be methodologically inaccurate to label the literature component a systematic review without those records.

3.1 Objectives

To explain the relationship between automation, generative AI, and contemporary cybercrime.

To review evidence concerning AI-assisted phishing, social engineering, ransomware, credential theft, and vulnerability exploitation.

To describe the participating students' awareness, reported exposure to suspicious messages, perceived risk, and self-rated cybersecurity awareness.

To integrate the student survey with external threat reporting while preserving differences in samples, definitions, and measurement.

To propose evidence-informed defensive and educational measures.

3.2 Research questions

RQ1: How can AI and automation support selected stages of cybercriminal activity?

RQ2: What awareness and self-reported exposure patterns appear in the participating student sample?

RQ3: What defensive priorities are supported by the reviewed research and threat reports?

3.3 Research Design

The study uses an exploratory mixed-source design. The primary component is a descriptive questionnaire survey of 36 university students. The secondary component is a focused review of selected peer-reviewed studies and published reports from IBM, WEF, ENISA, Verizon, Google Cloud/Mandiant, and the National Institute of Standards and Technology (NIST). The survey provides local descriptive information; the external sources provide contextual evidence about broader threat patterns.

3.4 Participants and instrument

A total of 36 respondents completed a Google Forms questionnaire. The available study record reports questions on awareness of AI-based cyber threats, experience with suspicious or potentially phishing-related messages, perceived impact of AI on cybercrime, and self-rated cybersecurity awareness on a five-point scale. Demographic breakdowns, recruitment details, response rate, questionnaire validation, and item-level raw data were not available in the manuscript supplied for revision. They are therefore not inferred or fabricated.

3.5 Analysis

Frequencies and percentages were retained from the original survey summary. Percentages were checked against $n = 36$ and are reported to one decimal place. The analysis is descriptive; no inferential tests, associations, or causal conclusions are reported because individual-level response data and cross-tabulations were not supplied. External statistics are attributed to the organisations and datasets that produced them.

3.6 Ethics and data protection

The supplied manuscript does not document formal ethics approval, participant age, consent wording, or data-retention procedures. Before publication, the authors should verify and report the applicable institutional ethics requirements, informed-consent procedure, voluntary participation, confidentiality protections, and handling of survey data. This revision does not assume that approval was obtained.

3.7 Limitations of the evidence base

To improve reproducibility in a future version, the authors should preserve the original questionnaire, export anonymised item-level data, document recruitment and inclusion criteria, record the survey dates, and maintain a source-extraction sheet for each external report. If raw data become available, counts and percentages should be recalculated from the original records and checked independently. No individual-level or demographic findings should be inferred from the summary figures alone.

The primary sample is small and cannot represent all university students or internet users. Self-reported awareness may differ from demonstrated skill, and reported suspicious messages were not independently verified. The secondary sources use different methods, populations, time periods, and

definitions. Their percentages must not be pooled or presented as a single trend line. The literature discussion is a focused synthesis rather than a preregistered systematic review.

4. Results

4.1 Student survey findings

Table 2. Summary of primary survey results (n = 36)

Survey indicator	Response	Frequency	Percentage
Awareness of AI-based cyber threats	Yes	32	88.9%
Awareness of AI-based cyber threats	No	4	11.1%
Received suspicious/potentially phishing-related messages	Yes	19	52.8%
Received suspicious/potentially phishing-related messages	No	17	47.2%
Believes AI makes cybercrime more dangerous	Yes	22	61.1%
Believes AI makes cybercrime more dangerous	Unsure	13	36.1%
Believes AI makes cybercrime more dangerous	No significant effect	1	2.8%
Self-rated cybersecurity awareness	1	0	0.0%
Self-rated cybersecurity awareness	2	7	19.4%
Self-rated cybersecurity awareness	3	17	47.2%
Self-rated cybersecurity awareness	4	10	27.8%
Self-rated cybersecurity awareness	5	2	5.6%

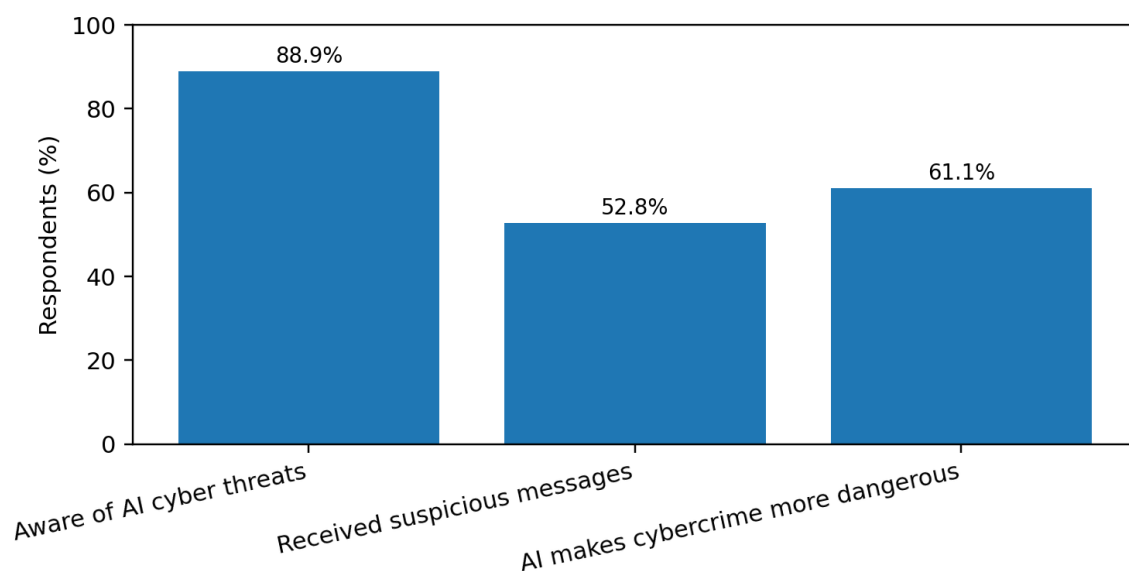


Figure 1. Three key indicators from the student survey. Percentages are descriptive and based on n = 36.

Awareness was high within the sample: 32 of 36 respondents (88.9%) said they were aware of AI-based cyber threats. This should not be equated with the ability to recognise a real phishing message or respond safely to an incident. Nearly half the sample reported no suspicious messages, while 19 respondents (52.8%) reported receiving messages they considered suspicious or potentially phishing-related. Because the messages were not collected or independently assessed, this is a report of perceived exposure rather than a verified phishing incidence rate.

4.2 Perceived impact of AI

Twenty-two respondents (61.1%) believed AI makes cybercrime more dangerous; 13 (36.1%) were unsure; and one respondent (2.8%) believed AI has no significant effect. The large unsure group is meaningful for educational planning; learners may recognise the issue but lack a clear evidence-based model of how AI changes threat capabilities. These results measure beliefs, not the actual effect of AI on attack frequency or success.

4.3 Self-rated cybersecurity awareness

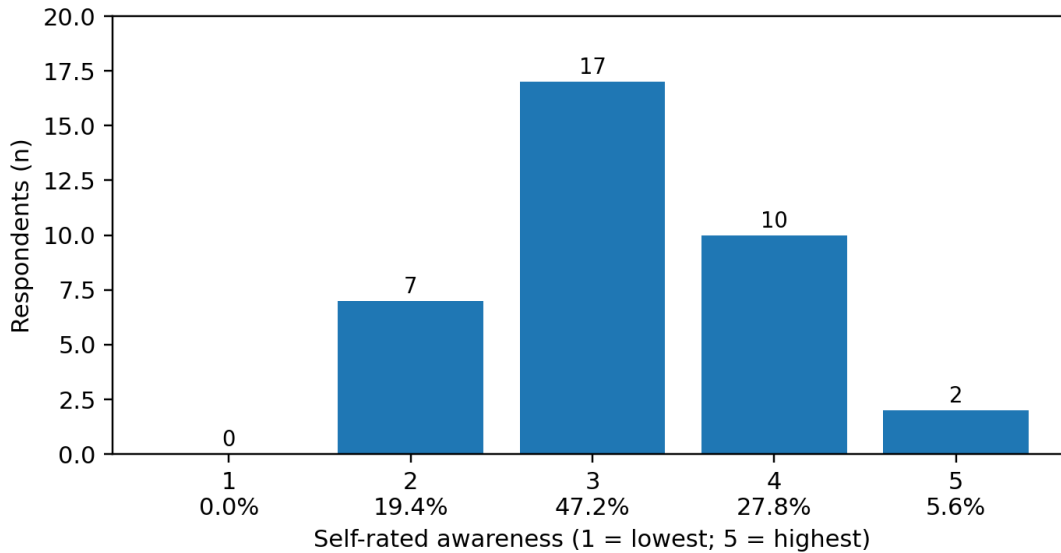


Figure 2. Distribution of self-rated cybersecurity awareness (1 = lowest, 5 = highest; n = 36).

The most common rating was 3 out of 5, selected by 17 participants (47.2%). Ten respondents (27.8%) selected 4, seven (19.4%) selected 2, and two (5.6%) selected 5; none selected 1. The distribution suggests that respondents most commonly placed their awareness in the middle of the scale. Since self-assessment can be affected by confidence and differing interpretations of the scale, an objective scenario-based assessment would be needed to evaluate practical competence.

4.4 External evidence on the contemporary threat environment

Table 3. Selected published cyber-risk indicators

Source and period	Reported indicator	Interpretation and boundary
IBM X-Force (2025 report; 2024 observations)	84% year-on-year increase in emails delivering infostealer malware	Change in IBM-observed email activity; not an estimate of all phishing globally or proof AI caused the increase
IBM X-Force (2025 report; 2024 cases)	Identity abuse was the initial entry point in 30% of cases examined	IBM's observed case dataset; not a population-wide rate
WEF (2025 outlook survey)	66% expected AI to have a significant cybersecurity impact in the coming year	Organisational expectations, not observed attack incidence
WEF (2025 outlook survey)	37% reported processes for assessing AI-tool security before deployment	Self-reported organisational preparedness
Verizon (2025 DBIR)	Compromised credentials were an initial access vector in 22% of breaches reviewed	Verizon DBIR dataset and definitions
Mandiant M-Trends (2025; 2024 investigations)	Exploits were 33% and stolen credentials 16% of initial infection vectors	Mandiant investigation dataset; not directly comparable with Verizon or IBM

ENISA (2024 threat landscape)	Analysed more than 11,000 publicly reported incidents; availability threats, ransomware, and data threats were leading categories	Publicly reported incidents within ENISA's scope
-------------------------------	---	--

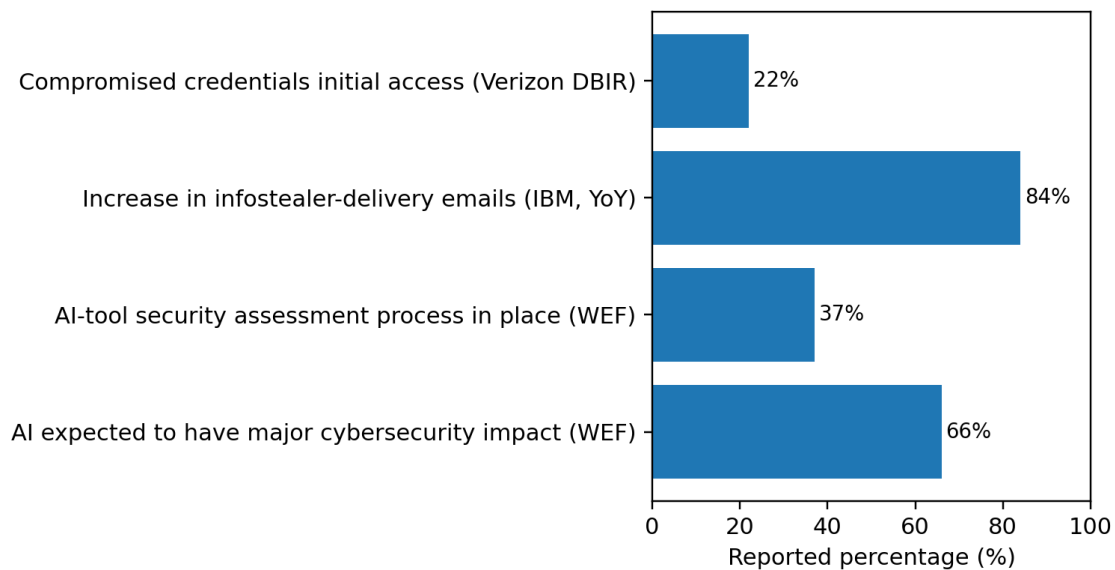


Figure 3. Selected reported percentages from separate industry datasets and surveys. These metrics have different denominators and meanings; the chart is contextual, not a direct comparison.

Figure 3 is intentionally presented as a set of contextual indicators, not as a ranking of equivalent risks. The WEF values are survey responses about expectations and organisational processes; IBM's 84% is a year-on-year change in observed email activity; Verizon's 22% is a share of reviewed breaches; and Mandiant's values concern initial infection vectors in its investigations. The differing measures should not be pooled, averaged, or interpreted as one common prevalence estimate.

A central finding from the triangulation is that the strongest defensible conclusion concerns risk priorities, not a single measured effect of AI. Across the sources, identity protection, phishing resilience, vulnerability management, and organisational preparedness recur as important concerns. The sources do not share a common sampling frame, and the student questionnaire did not measure AI attribution. Consequently, the external data contextualise the survey rather than validate or generalise its percentages.

The primary survey is also too small to support subgroup analysis or robust inferential claims. For example, the aggregate results do not reveal whether students who reported greater awareness were more likely to recognise suspicious messages. Such a relationship would require participant-level paired responses and a prespecified analysis plan. This distinction prevents the discussion from treating plausible associations as established findings.

4.5 Evidence synthesis

Table 4. How the evidence informs the research questions

Evidence stream	What it supports	What it does not support
Student questionnaire	Description of awareness, perceived risk, self-rating, and reported suspicious-message exposure among 36 participants	Population prevalence, verified phishing incidence, AI attribution, or causal inference
Peer-reviewed reviews	Conceptual mechanisms by which generative AI may amplify social engineering; current defensive approaches	A single universal success rate for AI-enabled attacks
Industry threat reports	Observed patterns within each publisher's incident, telemetry, or survey dataset	Direct comparisons across publishers without harmonised definitions
Standards and frameworks	Structured risk-management and governance actions for organisations	A guarantee that adopting a framework alone prevents incidents

5. Discussion

The student survey indicates a distinction between awareness and confidence. Although 88.9% reported awareness of AI-based cyber threats, the most frequent self-rating was the midpoint of the scale, and 36.1% were unsure whether AI makes cybercrime more dangerous. This pattern should not be overinterpreted, but it suggests that awareness campaigns should move beyond general warnings and teach observable behaviours: checking sender identity through a separate channel, avoiding unverified links, reporting suspicious messages, and using multifactor authentication.

The 52.8% who reported receiving suspicious messages provide a relevant educational signal, but not a verified exposure rate. Some messages may have been legitimate communications; others may have been phishing; and the questionnaire did not determine whether AI was involved. Future surveys should collect non-sensitive examples only with consent and safe handling, use validated phishing scenarios, and measure correct identification and response behaviour rather than relying solely on self-report.

The external evidence aligns with the emphasis on identity and social engineering. IBM's observed increase in infostealer-delivery emails and Verizon's reporting on compromised credentials both reinforce the importance of protecting identities. WEF's reported gap between expected AI impact and AI-tool security assessment indicates that organisational governance may lag adoption. Together, these findings support a layered response rather than reliance on a single AI detector or awareness lecture.

The literature also cautions against technological determinism. Generative AI can reduce the effort required to produce plausible content, but it does not eliminate the need for access, infrastructure, targeting, and operational decisions. AI-assisted attacks may fail, be detected, or be blocked. Conversely, ordinary credential abuse or exploitation can be highly damaging without generative AI. The appropriate research question is therefore which tasks AI changes, under what conditions, and with what measurable outcomes.

For universities, this is both a technical and behavioural challenge. Institutions hold identity data, research records, financial information, and learning-management accounts, often across diverse user groups and devices. A proportionate programme should combine secure account defaults, accessible reporting channels, repeated low-stakes practice, incident communication, and clear policies for the safe use of AI tools. Training should be designed around realistic decisions rather than blame: learners should know how to pause, verify, and report without fear of embarrassment.

5.1 Practical implications

Individuals: use unique passwords and an approved password manager; enable multifactor authentication; verify unexpected payment, credential, or file-sharing requests through a separate trusted channel; report suspicious messages.

Universities and organisations: enforce least privilege, patch exposed systems, monitor authentication events, maintain tested backups, segment critical systems, and rehearse incident response.

AI governance: inventory AI tools, assess data exposure and vendor controls, test systems before deployment, restrict sensitive data in prompts, log high-risk use, and retain human review for consequential security decisions.

Cybersecurity education: assess practical performance through safe simulations, teach verification routines, provide clear reporting mechanisms, and repeat training over time.

Research: distinguish AI-generated content from AI-enabled delivery and measure outcomes such as identification accuracy, reporting rates, false positives, and time to response.

5.2 Recommended defence framework

Table 5. Layered response to automated and AI-assisted cybercrime

Layer	Recommended measures	Rationale
Identity	Phishing-resistant MFA where available; unique passwords; least privilege; conditional access	Reduces the value of stolen credentials and limits account compromise
Email and messaging	Domain authentication; attachment/link controls; external-sender indicators; easy reporting; out-of-band verification	Addresses deceptive communication without assuming AI can be detected reliably from style alone
Endpoint and network	Timely patching; endpoint detection; segmentation; secure configuration; centralised logging	Reduces opportunities for exploitation and lateral movement

Layer	Recommended measures	Rationale
Resilience	Offline/immutable backups; tested restoration; incident-response playbooks; tabletop exercises	Limits operational impact and improves recovery
AI system governance	Inventory; security assessment before deployment; data minimisation; access controls; monitoring; human oversight	Addresses risks introduced by organisational AI adoption
Human factors	Scenario-based training; just-in-time prompts; no-blame reporting; repeated practice	Builds usable habits under realistic conditions
Measurement	Track reporting, detection precision, response time, repeat errors, and recovery outcomes	Allows programmes to be evaluated rather than relying only on attendance or self-confidence

Implementation should be prioritised according to risk and feasibility. Institutions with limited resources can begin with high-impact baseline controls—multifactor authentication, timely patching of internet-facing systems, protected backups, and a clear suspicious-message reporting route—before investing in more complex AI-driven detection. More mature organisations can add telemetry correlation, automated enrichment, and model-assisted triage, provided that they evaluate false positives, data exposure, and analyst oversight.

Evaluation should include leading and lagging indicators. Leading indicators may include multifactor-authentication coverage, patch compliance, participation in simulations, and the proportion of staff who know how to report an incident. Lagging indicators may include confirmed account compromises, time to contain incidents, recovery time, and recurrence of similar events. These measures should be interpreted alongside operational context; a rise in reports can indicate improved detection and trust in the reporting process rather than a deterioration in security.

5.3 Framework alignment

NIST Cybersecurity Framework 2.0 provides a risk-management structure organised around Govern, Identify, Protect, Detect, Respond, and Recover (Pascoe et al., 2024). This structure can be used to assign ownership and assess gaps across the measures in Table 5. It is a framework for managing risk, not a product or a guarantee of security. Organisations should tailor implementation to their size, mission, threat exposure, and available resources.

Table 6. Suggested indicators for evaluating a university cybersecurity programme

6. Limitations and Future Research

Several limitations restrict the conclusions. First, the survey includes only 36 participants, and the sampling approach and response rate are not documented in the supplied manuscript. Second, responses are self-reported. Third, suspicious messages were not verified, and the study did not identify whether AI was used in any reported communication. Fourth, individual-level data were unavailable, preventing cross-tabulations or inferential analysis. Fifth, external reports use different sampling frames and metrics. Finally, the literature synthesis is focused rather than a systematic review with a reproducible database search protocol.

Future research should recruit a larger and more diverse sample, document recruitment and participant characteristics, use a validated questionnaire, and add a consent-based scenario assessment. A pre/post design could evaluate whether training improves recognition and reporting. If raw responses are available, future analyses could examine whether self-rated awareness corresponds to scenario accuracy; such a relationship must be tested rather than assumed. Longitudinal studies and incident datasets with clear attribution criteria are needed to estimate changes in AI use and attack outcomes.

7. Conclusion

AI and automation are reshaping selected parts of the cyber threat landscape, particularly content generation, personalisation, analysis, and the scaling of repetitive workflows. The reviewed evidence supports concern about AI-assisted social engineering and the continuing importance of credential protection, vulnerability management, and organisational readiness. It does not support the claim that every modern attack uses AI or that AI alone explains changes in cybercrime.

In the primary survey, 88.9% of participants reported awareness of AI-based cyber threats, 52.8% reported receiving suspicious or potentially phishing-related messages, and 61.1% believed AI makes cybercrime more dangerous. Self-rated awareness most commonly fell at the midpoint. These exploratory findings point to a need for practical, measurable cybersecurity education, while their small sample and self-report design preclude broad generalisation.

The most defensible response is layered and evidence-led: strengthen identity controls, patch and monitor systems, maintain recoverable backups, assess AI tools before deployment, and build reliable human verification and reporting practices. Further research should measure actual performance and incident outcomes, distinguish AI assistance from automation and autonomy, and avoid attributing cyber incidents to AI without evidence.

References

- Atuncar Flores, E. J., Chuan García, A. F., Ráez Martínez, H. T., & Pachas Quispe, G. H. (2024). Algorithms based on artificial intelligence for the detection and prevention of social engineering attacks: Systematic review. **Proceedings of the LACCEI International Multi-Conference for Engineering, Education, and Technology, 1*(10)*. <https://doi.org/10.18687/LACCEI2024.1.1.1026>
- European Union Agency for Cybersecurity. (2024). **ENISA threat landscape 2024**. <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2024>
- IBM. (2025). **X-Force threat intelligence index 2025**. IBM Security. <https://www.ibm.com/think/x-force/x-force-threat-intelligence-index-2025-attackers-steal-sell-user-identities>
- Kutscher, J. (2025, April 23). **M-Trends 2025: Data, insights, and recommendations from the frontlines**. Google Cloud. <https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2025/>
- Naqvi, B., Perova, K., Farooq, A., et al. (2023). Mitigation strategies against the phishing attacks: A systematic literature review. **Computers & Security, 132**, 103387. <https://doi.org/10.1016/j.cose.2023.103387>
- Pascoe, C., Quinn, S., & Scarfone, K. (2024). **The NIST Cybersecurity Framework (CSF) 2.0** (NIST Cybersecurity White Paper CSWP 29). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.CSWP.29>
- Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. **Artificial Intelligence Review, 57**, Article 324. <https://doi.org/10.1007/s10462-024-10973-2>
- The weaponization of artificial intelligence in cybersecurity: A systematic review. (2024). **Procedia Computer Science, 239**, 547–555. <https://doi.org/10.1016/j.procs.2024.06.206>
- Verizon. (2025). **2025 data breach investigations report**. <https://www.verizon.com/business/resources/reports/dbir/>
- World Economic Forum. (2025). **Global cybersecurity outlook 2025**. <https://www.weforum.org/publications/global-cybersecurity-outlook-2025/>
- Primary survey data. (2026). **Google Forms survey on AI-based cybercrime awareness** [Unpublished raw survey summary; n = 36].

Evaluation domain	Example indicator	Interpretation caution
Preparedness	Percentage of accounts protected by MFA	Coverage should be interpreted alongside account type and authentication strength
Human response	Correct identification and reporting rate in safe simulations	Simulation performance does not perfectly predict behaviour during real incidents
Detection	Median time from alert to triage	Lower time is useful only if accuracy and workload remain acceptable
Resilience	Successful restoration in scheduled recovery tests	A backup is not a proven recovery capability until restoration is tested
Learning	Change in scenario-based assessment from pre-test to post-test	Use comparable instruments and report attrition and uncertainty

The indicators in Table 6 are proposed evaluation measures, not outcomes observed in the present study. They should be selected in advance, measured consistently, and reviewed for unintended effects such as excessive false-positive alerts or privacy intrusion.