International Journal for Innovative Engineering and Management Research
PEER REVIEWED OPEN ACCESS INTERNATIONAL JOURNAL
www.ijiemr.org

When AI Becomes the Hacker: How Artificial Intelligence Is Transforming Cybercrime.

Ms. Ameesha Firdous Mohammed, Mr. EN. Ganesh Nanda, Mr. Mirza Qanitullah Baig.

Students, BS-MS (Computer Science), Vishwa Vishwani Institute of Systems and Management, Hyderabad, Telangana, India.
Email: ameeshafirdous@gmail.com | endlaganeshnanda123@gmail.com
mirzaqanitullahbaig2007@gmail.com

Abstract

AI serves as a double-edged sword in cybersecurity. It is used by organizations as defensive tool and by cybercriminals as a powerful offensive weapon. Cybercriminals and cybersecurity professionals increasingly use the same AI technologies for opposing purposes. Key challenges are adversarial AI (attacks that manipulate AI models), data poisoning (corrupting training data) and automation complacency (missed context by over-relying on AI). Defenders want to check if there is any threat and then defend it. Meanwhile, attackers use AI to increase their efficiency, causing a breach. In defensive methodology, defenders use AI to analyze behavioral patterns that traditional systems miss. In offensive methodology, cybercriminals use AI to attack and adapt tactics. The result of the dual role of cybersecurity includes increased attack size and speed, democratization of cybercrime and increased fraud losses. To help people be more aware of cybersecurity related apps and build trust among people, we propose designing an app that helps people detect frauds and make sure that more people are aware of it. We also conducted a survey to understand people's awareness of AI-related cybercrime, their experiences with online fraud, and their knowledge of how to identify and respond to such incidents. We have received 36 responses from the survey that we conducted, The responses helped us understand the importance of cybersecurity awareness and the need for an application that helps people detect potential frauds and stay safe online. With the increase in the usage of AI, cybercrime rates are increasing.

Keywords: Artificial intelligence; cybersecurity awareness; cybercrime; fraud detection; user survey; attackers; defenders; technology; application software.

1. Introduction

Artificial intelligence (AI) is entering daily life and transforming communication, work and online service provision. In cybersecurity, it is a double-edged sword: it enables organizations to identify suspicious activity, analyze patterns and recognize threats, but it also empowers cyber-criminals to pursue their own ends. Understanding the ways in which it can be used for both is important for the entities and individuals that rely on the digital world.

Cyber-criminals can use it to generate phishing messages, impersonate others using deepfakes and exploit weaknesses in a target’s defenses with automated attacks. At the same time, people can place blind trust in security applications and automatically accept all alerts as potential threats. NIST emphasizes the danger of believing in the reliability of AI in all situations and the paper’s

main challenge is related to the interplay between evolving attack techniques and people's understanding of them.

This research was conducted to separate the ways in which AI-assisted cyber-attacks and protective measures co-exist. It seeks to review the protective applications and explore their strengths and weaknesses in terms of user experience and threat identification and response. We are interested in the ways in which users can trust such software and benefit from it without being misled by apparent automation and false positive claims. This paper aims to contribute to the understanding of the ways in which AI-driven crime can be combated through common applications and procedures.

Its primary objectives are to identify the role of AI in cyber-crime and review the protective measures taken by applications. It also endeavors to discuss the strengths and weaknesses of such programs and elaborate on the opportunities for human factor intervention. We conducted a survey to gather responses about people's experience with online fraud, awareness of AI-related risks and willingness to use applications to detect and report cyber-security issues. We also analyzed six cybersecurity applications to identify and evaluate their protective mechanisms and user-friendly interfaces and document the known risks and weaknesses. The paper also aims to benefit from published research and documented cases of AI-assisted crime and attempts to combat it.

It focuses on phishing, impersonation and online fraud, as well as AI's role in protection. By reviewing the literature, analyzing cases and compiling survey results, we attempted to devise a concept of an application that users can use to detect potential fraud and expand their knowledge of the issue. Our paper focuses on one potential solution, but it is only a concept that requires further validation. Our survey only reflected the perspectives of our respondents and cannot be generalized to the broader population.

2. Literature Review

2.1 Evolution of cybercrime and AI

Cybercrime is evolving from traditional malware and phishing to include attacks facilitated by generative artificial intelligence (AI), including synthetic voices and video, and automation. Such tools build on existing methods, complementing and automating some aspects, but not replacing, manual effort. National Cyber Security Centre (NCSC, 2025) notes that attackers are already using AI for reconnaissance, social engineering and malware design, with increased adoption set to make cybercrime more efficient and accessible. Its horizon scan makes clear that it considers fully autonomous sophisticated attacks to be unlikely prior to 2027. In other words, AI tends to be used to make labour-intensive processes more repeatable and scalable, but not necessarily to replace skilled operators entirely.

2.2 Existing research on AI-assisted cyber-attacks

Existing research demonstrates the potential of AI to enable increasingly realistic and targeted phishing. Heiding et al.'s (2024) preprint, based on an experiment with 101 participants, found

that phishing emails generated by AI had a 54% click-through rate, comparable to the performance of humans. Notably, this says nothing about the financial outcome, as clicking an email is not the same as transferring funds, and a single experiment is not necessarily generalizable across populations. Microsoft and OpenAI's (2024) report on the abuse of AI found that threat actors were already using language models for research, scripting and phishing emails, but saw no significant attacks beyond known methods. Similarly, while deepfake-enabled impersonation is also a concern, Arup's CIO notes that the company fell victim to a $25M scam due to a deepfake of a senior executive, not a breach of their IT infrastructure.

2.3 Existing research on AI-powered cyber-security

But attackers are not the only ones using the technology. Generative and other forms of AI can also be trained to detect behavioural patterns characteristic of cyber-attacks, thus supporting organisational defences. Knowledge and Information Systems journal's 2025 review on machine learning for intrusion detection, malware classification and other cyber-security tasks discusses advantages and shortcomings of various approaches. The review highlights the value of using machine learning to identify patterns and correlations across large datasets, as well as response speed, but also notes the risks of inadequate training data and reduced performance against novel attacks. In other words, attackers can employ tactics that the system does not recognize, leading to either false negatives (missed attacks) or false positives (alerts when there are none). Moreover, threat actors might also target the system itself, for example, by manipulating input data or poisoning machine learning models. In practice, as Microsoft's 2025 Cyber Signals report demonstrates, AI-driven fraud prevention systems rely on combinations of different signals to detect, for example, a fraudulent attempt to reconnect with a previously blocked business. In none of the cases reviewed does the use of AI appear to guarantee protection against attacks.

2.4 Research gaps and relevance to the current study

The existing research highlights the importance of continued vigilance and the need for further investigations into the use of AI for cyber-security. The main contribution of the current study is to add a participant-focused perspective, highlighting awareness and recognition challenges through the analysis of survey results and the proposed comparison of six cybersecurity applications. The limitations of these tools in terms of user understanding and documented attack vectors contribute to the argument for the development of a fraud-awareness application. Its value, however, should not be overstated: the study only provides context for the need for such software, without demonstrating its efficacy.

3. AI as a tool for cybercrime

3.1 AI-generated phishing and social engineering

Traditional phishing schemes involve mass mailing, whereas targeted phishing requires research and writing skills. Generative AI reduces the effort needed for the latter by providing professional language, detecting and correcting grammar mistakes, and applying relevant terms for each target.

Criminals can use this technology to impersonate respectable companies and base the message on the target's job role or hobbies. The FBI warns that this method makes the phishing attempt more credible and allows reaching thousands of scammers.

The working principle of phishing remains social engineering since users are persuaded to perform particular actions, provide information, or fall victim to scams. In some cases, attackers use urgency, scare tactics, or appeal to authority, whereas trustworthy-looking messages promote products or services without being suspicious. AI increases the quality of phishing emails and text messages, which makes them harder to detect, but their success still depends on the target's reaction and security measures. Overall, the main value of using AI for cybercriminals is creating personalized phishing campaigns with smaller effort and distributing them at a larger scale.

3.2 Deepfakes, voice cloning and identity fraud

Deepfake technology employs AI to create videos or audio recordings of people saying or doing things they never did. The FBI warns about phone scams with cloned voices of relatives having an emergency, fake video conferences of executives authorizing payments, and altered photos or videos of recognizable people. Fake identification documents and passports also rely on deepfake portraits. In general, authentic voices, images, and videos are no longer a guarantee of trustworthiness.

An example of a deepfake scam is the $25 million fraud from engineering company Arup in early 2024. The criminals used deepfakes of a senior executive and other colleagues on a video call to trick an employee into transferring the money to scammers. Arup's CIO said that the incident was not a technical breach but social engineering with AI involvement. In the case of fake voices, a call to the supposed victim can help clarify the situation since contacting the same person via phone is harder for an attacker. In general, any suspicious message or call should be double-checked, and sensitive information or large transfers require additional verification.

3.3 AI-assisted malware and attack automation

Malware creation and security detection rely on extensive programming knowledge and are complicated by various protection mechanisms. AI tools can assist attackers by suggesting code alterations and explaining the existing code structure, and Microsoft's 2024 report confirms such attacks' occurrence. Cybersecurity analysts notice that threat actors ask AI to provide scripting advice and research topics, such as methods to avoid detection.

Attacks using such AI assistance still require manual input and do not demonstrate sophisticated capabilities. The NCSC published a technical note in March 2026, claiming that the tested models successfully launched attacks but were unable to complete the task of simulating a ransomware attack across an enterprise IT environment. In general, the researchers note discrepancies and flaws in the code suggested by AI, which limits its immediate use. However, such tools are helpful for both cybercriminals and defenders since the former try to find system weaknesses, and the latter attempt to beat them by testing suggested opportunities.

3.4 Automated reconnaissance and vulnerability discovery

Reconnaissance is the process of gathering information about the target to find potential weaknesses for attack. Malware developers use this intelligence to identify the systems to launch attacks, victims' personal data to extort money, or company employees to bribe for data and access. AI tools help in this process by summarizing public sources, listing target-related topics, or explaining the information on a webpage. Microsoft's report on AI-driven security threats mentions that threat actors use language models to research potential victims, collaborators, and various companies in the supply chain. AI is instrumental in reconnaissance since it can also find publicly available exploits or test whether the discovered vulnerability can benefit the attacker.

Such tasks are time-consuming and require significant technical expertise, which is why automated tools are employed. Note that vulnerability discovery does not mean that the weakness can be used immediately; this requires separate assessment and exploitation research, which is complicated by protective measures. However, defenders use the same opportunities to check if their systems are affected and recommend technical improvements for better security. Both sides use reconnaissance data, but the attackers' efforts are limited by the need for manual input and the possibility of error.

4. Methodology

4.1 Survey design and participants

We conducted a questionnaire survey, the Cybersecurity Awareness & Trust Survey, to gather responses about people's experience with online fraud, awareness of AI-related risks and willingness to use applications to detect and report cyber-security issues. The questionnaire also asked respondents to confirm consent and age eligibility (18 years or older). The survey was created as a Google Form and shared with our college group, seniors and other contacts, as well as on LinkedIn, Instagram and WhatsApp, so respondents were reached through our own networks. It was shared on 15 October 2026 and received 36 responses over five days, which are summarized descriptively in Section 5 using counts, percentages and mean ratings out of 5.

4.2 Application review

We reviewed six cybersecurity applications (Norton, McAfee, CyBx, CyberTor, Mkavach and Mobi Armor) to identify and evaluate their protective mechanisms and user-friendly interfaces and to document known risks and weaknesses. The review considered protective features together with affordability, guidance and simplicity. Each application was downloaded and checked in turn by one of our team members, who compared them and recorded her notes. Our comparison is based on these notes rather than a formal or objective assessment of each application. The review was carried out on 14 and 15 October 2026, using the latest version of each application available on the Google Play Store.

4.3 Literature and case review

We also drew on published research, institutional guidance and documented cases of AI-assisted crime and attempts to combat it (see Sections 2 and 3). Sources were found through online searches, and we studied the most reliable ones.

5. Cybersecurity Awareness & Trust Survey: Summary Report

Table 1 summarizes the key figures reported in the subsections below.

Table 1. Key figures from the Cybersecurity Awareness & Trust Survey (authors' survey).

Measure	Result
Total responses	36
Respondents aged 18 to 24	30
Gender	20 female, 14 male
Comfort with technology (mean, out of 5)	3.8
Personal concern about security (mean, out of 5)	4.0
Confidence in spotting a scam (mean, out of 5)	3.4
Personally targeted by a scam	16 (44%)
Knew someone who had been targeted	22 (61%)
Did not report the incident (of those who answered)	15
Reported to the police or an official portal	6
Unfamiliar with organizations that handle cybercrime	19
Did not know the fraud-reporting number	19 of 31
Rated interest in a scam-news app at 4 or 5	24
Preferred Design A (clean and minimal, padlock or shield)	27 of 36

5.1 Sample

The survey collected 36 responses, skewed toward young adults: 30 respondents were aged 18 to 24. Twenty were female and 14 male. Comfort with technology was fairly high (average 3.8 out of 5). One respondent declined consent and four said they are not 18 or older, so those entries may need review before formal analysis.

5.2 Concern versus confidence

Respondents care about security, averaging 4.0 out of 5 for personal concern, but are less sure of their own skills. Average confidence in spotting a scam was 3.4, and a quarter rated themselves 1 or 2.

5.3 Experience with scams

Sixteen respondents (44%) had personally been targeted, and 22 (61%) knew someone who had. Reporting was weak. Of those who answered, 15 did not report the incident, most often because they believed it would not help, the loss seemed small, or they felt embarrassed. Only six reported to the police or an official portal.

5.4 Knowledge gaps

Over half (19) were unfamiliar with organizations that handle cybercrime, and 19 of 31 respondents did not know the fraud-reporting number.

5.5 Behavior in scenarios

Responses were strongest for financial fraud: 30 would call their bank and the fraud helpline immediately, though two would wait a few days and four were unsure. For a "digital arrest" call, 20 would hang up and verify, but 11 would stay on the line and verify the caller directly, and one would follow instructions. For a scam SMS, 13 would simply ignore it and only seven would report it. After a hacked social media account, answers split between reporting to a cybercrime portal (15), using account recovery (10), and changing the password (6).

5.6 Trust in apps

Good ratings (23), a well-known brand (22), and a clear privacy policy (18) build trust most. Red flags include early requests for money or card details (29), excessive permissions (26), a missing privacy policy or contact details (22), and installs from outside the official store (21). Design A, clean and minimal with a padlock or shield, was preferred by 27 of 36.

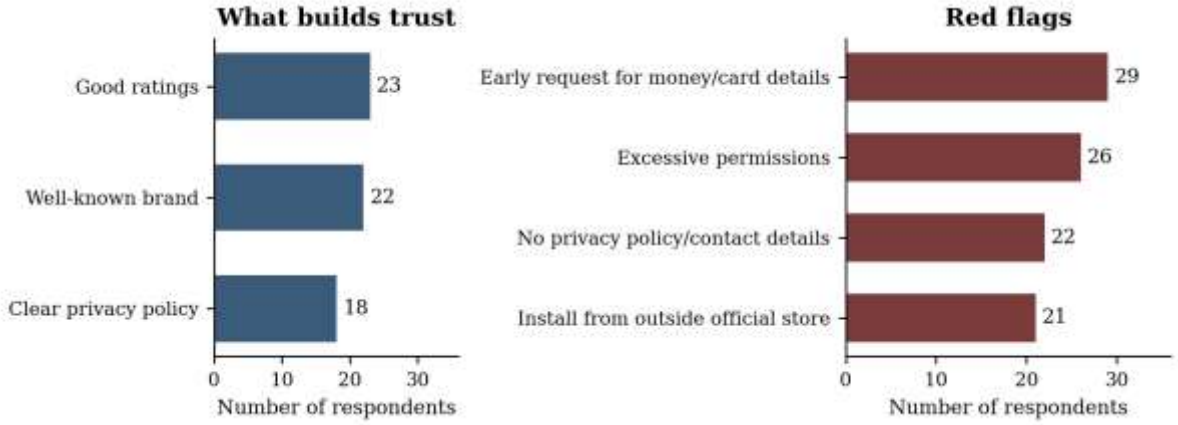


Figure 1. Factors that build trust in an app and red flags, as selected by survey respondents (number of respondents; 36 responses in total).

5.7 Appetite for a safety app

Two-thirds (24) rated interest in a scam-news app at 4 or 5. Step-by-step guides were the most wanted content (24), followed by local news (17) and direct reporting links (16). Half (18) preferred a mix of assistant types. For alerts about vulnerable loved ones, 20 said yes and 13 said maybe, depending on privacy. The top comfort factors were a clear explanation of what is monitored (19) and data staying on the device (14).

6. Cybersecurity Applications: Advantages, Limitations and Trust

Cybersecurity applications assist people in recognizing threats and protecting their devices. However, our review notes show that the applications emphasize distinct needs. The usefulness is determined not only by the protective features but also by such factors as affordability, guidance and simplicity. The below comparison reflects our notes rather than providing an objective assessment of every application.

6.1 Norton

Norton stands out in our comparison due to the VPN and dark web scanning features. However, the application's strongest points are paid, which may limit the usage for those who rely on free applications. Thus, its advantages should be weighed against the potential inability to use them.

6.2 McAfee

Our notes highlight that McAfee's access to advanced protection is a limitation due to its potential inaccessibility. In turn, the usefulness of the suggested tools depends on whether the user can afford and comprehend them. As our notes do not allow for assessing individuality, a more significant comparison requires further research.

6.3 CyBx

CyBx offers call protection and safety score, which may help users in understanding the current state of their devices. Nevertheless, a numeric value may not be always informative, as it does not specify the issues. The context of each warning should be explained in details to make every user, including novices, benefit from them.

6.4 CyberTor

Similarly to CyBx, CyberTor suggests call protection to mitigate suspicious calls. However, our notes do not specify whether this application adequately detects threats and guides users after they are detected. Thus, the application's main strength is the call security, whereas others should be considered with care.

6.5 Mkavach

Mkavach highlights hidden-app detection and fraud reporting features, which counter threats through recognition and response. However, our review shows that the interface is cluttered, making the functions unavailable to less experienced users. As a result, the novice-friendly utility may lack functions due to the inadequate interface.

6.6 Mobi Armor

Mobi Armor introduces expert settings that can be overwhelming to some users. However, this application can be recommended to those interested in advanced protection, which is not hindered by the interface. To make this tool beneficial for everyone, the essential and additional settings should be differentiated with simpler explanations.

7. The Kind of Application We Propose

The strongest concept is an application that is novice-friendly, offers fraud awareness, response guides and direct reporting. The users should be provided with informative warnings that explain what issues may be detected instead of focusing on a score.

Our survey results support this concept as 24 respondents opted for step-by-step guides, whereas 26 indicated excessive permissions as a red flag. Thus, the basic application should offer simple guidance, features and response actions. The extended version can suggest scam news, alerts, and awareness lessons. This approach will ensure that most users can recognize threats without being overwhelmed by the comprehensive device scans.

The interface should be designed with usability in mind, using a trustworthy color scheme and privacy-focused policy. The battery consumption, accuracy, and response rate should be mentioned for transparency, whereas the permissions will indicate whether the application is trustworthy. Finally, the awareness about frauds should be the main point of the discussed application, which should be achievable with reasonable expectations.

8. Comparative Discussion and Key Findings

8.1 Offensive AI vs. Defensive AI

AI is dual-use in cybersecurity: criminals apply it to generate phishing messages, impersonate humans and automate attacks while defenders utilize it for detecting anomalies, analyzing malware and preventing fraud. However, having access to the same technology does not provide the same outcome: practical results depend on the involved skills, available resources, data and systems. Furthermore, while attackers only worry about achieving their aims, defenders should consider privacy and the impact on innocent activity, requiring additional checks and regulation at every stage. Both sides may have advantages depending on the situation, but neither can act with a guaranteed outcome: attack scripts can fail, and defensive systems can generate false alarms. One cannot claim based on the evidence that one side is superior or has more value: the circumstances matter.

8.2 Findings from the Survey

The survey had 36 filled questions, but the exact number of valid entries is unknown since one participant refused consent and another four claimed to be under 18, possibly overlapping with the first one. Thus, the following findings apply to the entire sample of 36, possibly as overstatements. On average, the mean of "Level of concern about cybersecurity issues" was 4.0, while "Confidence in one's ability to spot scams" scored 3.4. Furthermore, 19 (out of 31) respondents did not know the number to report fraud, and 24 (out of 31) desired step-by-step guides as security content. Therefore, the general level of awareness seems low, but it is unclear if it translates into susceptibility. The findings might be utilized to recommend further research or report features, such as the need for guides, privacy protection mechanisms or easier access to reports and

regulation. However, it cannot be concluded that the respondents can be easily hacked or that their answers reflect general trust levels.

8.3 Advantages, Limitations and Risks

The application comparison shows that the protection is situational since every tool has its weaknesses while affordability, discoverability and guidance are questionable. At the same time, the literature review demonstrated that faster analysis alone is insufficient: most attention should be paid to false positives, requiring human validation before any reaction. That is why NIST's draft Cyber AI Profile prioritizes the involvement of people in the loop to ensure reasonable risk assessment.

Similarly, better detection capabilities should be supplemented with permission management and data minimization to avoid excessive surveillance. As the survey highlights people's concerns about privacy, the automation of protection measures should be context-specific and limited by the ability to stop, review and comment on every action. Everyone should have access to threat assessments, receive reasonable explanations and utilize data-driven decision-making support tools.

8.4 Overall Interpretation

Overall, the literature, survey and application review contribute to the understanding that AI influences both sides of the conflict, but the best results are achieved when detection tools are supported by guidance. The analysis of the application showed the importance of human validation, especially for false positives, but it does not diminish the value of the opposite side's readiness to embrace new threats. Furthermore, the survey results need to be considered with care due to the small sample size and possible age bias. Therefore, organizations should pay attention to the tools that prioritize transparency and human validation while individuals should request detailed explanations, guides and reports. The findings do not support the notion of universal protection, but they can become the basis for building the protection concept around the application of these principles. The subsequent section will provide specific recommendations based on the literature, survey and application review.

9. Recommendations and Future Directions

9.1 Human Oversight and Security Awareness

Survey participants advice that will help them learn about tailored phishing and deepfake impersonation, using realistic examples. We recommend verifying suspicious requests by other means and not being convinced by a message alone, even if it appeared to come from a known sender. High value transfers should always require additional approvals and multifactor authentication, and password managers should be encouraged. Decisions made automatically by security tools that could have negative consequences for individuals or organizations should have human oversight, particularly if mistakes by the system could cause disruptions. We recommend

that human judgment and discretion should be part of security decision making, and that technology should not create the illusion that social engineering or human error are no longer a problem.

9.2 Secure AI Development and Testing

We recommend that developers build in security at all stages of design, deployment and maintenance, rather than responding to incidents after they occur. Testing should include adversarial examples and prompt injection when relevant, and training data and proprietary models should be guarded against unauthorized access. We recommend that developers identify and limit access to functionality that could be abused, and that they monitor their systems in production, including maintaining logs that can help with auditing. They should analyze false positives, false negatives, and shifts in performance, and address any weaknesses that may emerge when new data, models, prompts, or other elements are added or changed. Testing should continue after initial release since updates and changes to the system can introduce new flaws, and this allows developers to identify risks before users encounter them.

9.3 Responsible AI Governance and Regulation

We recommend that organizations establish accountability for decisions made by AI and that they provide clear explanations for how personal information may be used. The NIST framework for voluntary use by organizations is not a regulation or guarantee, but rather guidance they can use to understand and manage risks at each stage. We recommend that governments and industry promote responsible governance, including appropriate levels of oversight for different applications, standardized responsible disclosure practices, and collaboration among researchers, companies, and government in addressing emerging risks. Effective governance should balance regulatory and technical approaches, without assuming that one alone will address all risks.

9.4 Future Challenges of Autonomous AI

The ability of AI assistants to complete complex tasks with limited direction is growing, but successful demonstrations of such capabilities should not be mistaken for true autonomy. Analysis of known attacks shows that attackers have had mixed results when attempting to use such methods. Researchers should consider the risks posed by AI agents that find and exploit vulnerabilities, adapt their approach when their initial efforts are countered, or appear to make mistakes or be misled by human operators. We recommend that researchers evaluate safe operating envelopes, requirements for human approval or interruption, and mechanisms to halt or disable autonomous systems. The guidance from NCSC recommends that organizations use sandboxes and limited resources when testing software, and actively monitor systems so that they are not vulnerable to unexpected behaviors of autonomous AI.

10. Conclusion

The article concludes by stating that AI is a double-edged sword that may be used both to protect organizations from cyber threats and to facilitate dishonest activities and scams. Thus, based on the application analysis and survey, one can state that protection should include not only detection but also explainable AI, easy reporting, and transparency of privacy procedures. However, these measures cannot guarantee that the proposed application will safeguard the users from all kinds of fraud and fulfill all expectations.

Modernizing protection means ensuring users' trust through human supervision, security awareness, and transaction monitoring. At the same time, developers should guarantee data privacy, implement detection tools, and evaluate the application continuously to ensure that it is not vulnerable to attacks. Finally, organizations and government should provide accountability, responsible breach disclosure, and appropriate security measures. Autonomous artificial intelligence development needs more analysis to determine whether it can help cybersecurity and what threats it can bring. As a result, protecting users from fraud and dishonest activities requires not only technological advances but also human expertise in choosing the necessary means of protection.

References

Federal Bureau of Investigation, Internet Crime Complaint Center. (2024). Criminal activity uses generative artificial intelligence to support financial fraud (Public Service Announcement I-120324-PSA). https://www.ic3.gov/PSA/2024/PSA241203

Heiding, F., Lermen, S., Kao, A., Schneier, B., & Vishwanath, A. (2024). Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects [Preprint]. arXiv. https://arxiv.org/abs/2412.00586

Microsoft. (2024). Staying ahead of threat actors in the age of AI. Microsoft Security Blog. https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/

Microsoft. (2025, April 16). Cyber Signals Issue 9: AI-powered deception: Emerging fraud threats and countermeasures. Microsoft Security Blog. https://www.microsoft.com/en-us/security/blog/2025/04/16/cyber-signals-issue-9-ai-powered-deception-emerging-fraud-threats-and-countermeasures/

Mohamed, N. (2025). Artificial intelligence and machine learning in cybersecurity: A deep dive into state-of-the-art techniques and future paradigms. Knowledge and Information Systems, 67(8), 6969–7055. https://doi.org/10.1007/s10115-025-02429-y

National Cyber Security Centre. (2023). Guidelines for secure AI system development. https://www.ncsc.gov.uk/collection/guidelines-secure-ai-system-development

National Cyber Security Centre. (2024). AI and cyber security: What you need to know. https://www.ncsc.gov.uk/guidance/ai-and-cyber-security-what-you-need-to-know

National Cyber Security Centre. (2025). Impact of AI and cyber threat from now to 2027. https://www.ncsc.gov.uk/report/impact-ai-cyber-threat-now-2027

National Cyber Security Centre. (2026a). Managing the cyber risk of agentic AI. https://www.ncsc.gov.uk/blogs/managing-the-cyber-risk-of-agentic-ai

National Cyber Security Centre. (2026b). Why cyber defenders need to be ready for frontier AI. https://www.ncsc.gov.uk/blogs/why-cyber-defenders-need-to-be-ready-for-frontier-ai

National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

National Institute of Standards and Technology. (2025). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (NIST AI 100-2 E2025). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf

World Economic Forum. (2024). This engineering firm was hit by a deepfake fraud. Here's what it learned [Video]. https://www.weforum.org/videos/arup-deepfake-fraud