What The Phish! Effects of AI on Phishing Attacks and Defense

Shreyas Kumar, Anisha Menezes, Sarthak Giri and Srujan Kotikela
Texas A&M University, College Station, USA

shreyas.kumar@tamu.edu
anisha.menezes@tamu.edu
nb2812@tamu.edu
drsk@tamu.edu

Abstract: The rapid advancement of artificial intelligence (AI) has significantly transformed the landscape of phishing attacks, presenting new challenges for detection and defense. AI-generated phishing emails, which leverage machine learning and natural language processing (NLP), have become increasingly sophisticated, making traditional detection methods ineffective. This research analyzes the evolution and impact of AI-driven phishing attacks, comparing the distinguishing linguistic and contextual patterns of AI-generated versus human-generated phishing emails. The study utilizes a comprehensive dataset, insights from informal discussions with Chief Information Security Officers (CISOs), and an analysis of historical phishing incidents before and after the release of advanced generative models like ChatGPT. Findings reveal that AI-generated phishing emails exhibit higher success rates due to their ability to bypass conventional spam filters and mimic human communication styles. Additionally, the research identifies significant gaps in current defense strategies and recommends a multi-layered security framework that integrates AI-specific detection tools, real-time threat intelligence, and machine learning-based anomaly detection to mitigate these evolving threats. This study emphasizes the need for organizations to proactively adapt to the growing sophistication of AI-powered phishing by implementing advanced defenses that are capable of keeping pace with the rapidly changing cyber threat landscape.

Keywords: AI-generated phishing, Cybersecurity, Machine learning, Data breaches, Defensive strategies

1. Introduction

The rise of artificial intelligence (AI) has significantly heightened the sophistication and effectiveness of phishing attacks, making them increasingly difficult to detect using traditional cybersecurity measures. AI-generated phishing emails utilize advanced natural language processing (NLP) and machine learning techniques to craft highly personalized and contextually relevant messages that can evade conventional spam filters and security systems. These AI-driven attacks pose a serious threat to critical industries like finance, healthcare, and agriculture, where interconnected systems are particularly vulnerable to disruption. For example, a targeted attack on the beef supply chain could manipulate logistics data, disrupt operations, and cause significant economic and public health repercussions. This research analyzes the unique features of AI-generated phishing emails, evaluates current defense mechanisms, and reveals significant gaps in existing security frameworks based on a survey of CISOs. To address these challenges, it proposes a multi-layered security framework incorporating AI-driven threat detection models, real-time threat intelligence, and improved traceability through RFID enhancements. Organizations must act swiftly to implement AI-based defenses, as failure to adapt could lead to severe data breaches, financial losses, and compromised digital infrastructures, underscoring the need for a proactive and integrated cybersecurity approach.

Research Questions

RQ1) How do AI-generated phishing emails compare to traditional phishing emails in terms of detection difficulty and effectiveness in bypassing current email security systems?

RQ2) Is there a growth of AI-based solutions to counter phishing attacks?

2. Background

Phishing attacks, which use deceptive emails or instant messages to trick users into divulging private information, have seen a significant rise in volume and complexity over the past few decades. Initially coined in the mid-1990s, “phishing” described a method used by hackers posing as AOL representatives to steal user credentials (Steves, Greene, and Theofanos, 2019). These early attacks laid the groundwork for what would become one of the most prevalent cyber threats today. As social media gained traction, attackers began exploiting personal information to craft targeted messages in what is now known as spear phishing. The latest evolution in phishing involves the use of artificial intelligence (AI) to generate convincing emails that can easily bypass traditional detection systems (Barrientos, Jacobs, and Dawkins, 2021).

With AI tools like natural language processing and machine learning, cybercriminals can now automate the creation of sophisticated phishing emails at unprecedented speed and scale (Canham, Dawkins, and Jacobs, 2024). This has contributed to a 94% increase in phishing attacks since 2020, with a 27.8% surge in global phishing activities from 2022 to 2023 alone (Proofpoint, 2024; Bolster, 2024). The release of AI tools like ChatGPT in November 2022 has made it easier for attackers to generate AI-phishing emails that are difficult for both humans and current models to detect. Technologically less advanced adversaries, such as smaller nations or organizations with limited resources, that had already been facing distinct difficulties in defending against cyberattacks, now face even greater challenges (Kumar, Nagar, 2024). As a result, cybersecurity defenders must urgently develop and implement AI-specific detection software (Greene, Steves, and Theofanos, 2018).

This research aims to analyze the evolving techniques used in AI-powered phishing and assess the effectiveness of existing countermeasures. By evaluating changes in the accuracy of leading phishing detection software and conducting interviews with Chief Information Security Officers (CISOs), this study provides a comprehensive overview of how organizations are adapting to the rise of AI-phishing and identifies areas for improvement in defensive strategies. The findings will contribute to developing advanced security measures that can effectively combat this growing threat (NIST, 2024; Schwartz, 2022).

3. Related Work

Research in the field of phishing detection and defense has grown significantly in recent years, particularly with the advent of AI-generated phishing attacks. Steves et al. (2019) introduced the "Phish Scale," a novel tool for assessing the difficulty of detecting phishing messages, which has set the foundation for evaluating phishing susceptibility in users. This work was further extended by Barrientos et al. (2021), who developed a framework for scaling the difficulty of detecting AI-generated phishing emails. Canham et al. (2024) expanded on these studies by examining differential phishing susceptibility among various user demographics, highlighting that not all victims react the same way to similar phishing cues. Similarly, Neil et al. (2023) provided an analysis of real-world phishing indicators, revealing how user behavior is influenced by specific email characteristics.

From a machine learning perspective, the National Institute of Standards and Technology (NIST) has contributed significantly to understanding and managing AI risks. In particular, NIST (2024) published a comprehensive taxonomy of adversarial machine learning attacks, detailing common vulnerabilities in AI-based systems and suggesting mitigation strategies. This publication complements the work of Schwartz (2022), which addresses bias in AI systems and its implications for cybersecurity, underscoring the need for balanced datasets and fair algorithmic practices. Additionally, Greene et al. (2018) explored the effectiveness of various cybersecurity strategies against sophisticated phishing attempts, emphasizing the challenges in detecting attacks that exploit advanced natural language processing capabilities.

3.1 Pre AI Attack and Defense

Before the advent of AI, phishing attacks were still on the rise as people became increasingly reliant on the Internet. During the COVID-19 pandemic, many services transitioned to online formats and people used instant messaging and emails as their primary means of communication. This gave attackers greater opportunity to deceive unsuspecting people. In fact, over the course of the pandemic, phishing attacks increased more than 600% (Verizon 2024).

However, defenders were also developing detection models at a rapid pace. Jain & Gupta (2016) suggest using an auto-updated white-list of legitimate sites accessed by the user. When the user tries to visit a website not in the white-list, the browser warns the user not to disclose their sensitive information. This model also checks the legitimacy of a website by extracting the hyperlinks from the source code of the webpage and applying them to a phishing detection algorithm (Jain & Gupta 2016). However, this model only had 86.02% accuracy rate in 2016, and phishing attacks have only grown more advanced since then. Yasin & Abuhasan (2016) also introduced the concept of weighting phishing terms in each email during the analysis phase. Their research tested 5 popular classification algorithms and found that the Random Forest algorithm worked the best with 99.1% accuracy.

More recently, Ho et al. (2019) proposed a large-scale characterization of lateral phishing attacks. Lateral phishing attacks are where the adversary leverages a compromised account to email other users, using the implicit trust between users and information from the compromised account. Ho et al. (2019) developed a classifier that found both identified lateral phishing attacks and those that went undetected. The model had an accuracy rate of 87.3% on their dataset of 87,413,431 emails. Unfortunately, Jain & Gupta's (2016) and Ho et al.'s (2019) models may become obsolete in the face of AI-phishing attacks.

3.2 Attacks Using AI

Phishing emails generated by AI often lack the poor grammar and overly suspicious requests that characterize traditional phishing emails. Additionally, AI-phishing emails can be tailored to each individual recipient and can even mimic the linguistic style of a close acquaintance (Heiding et al. 2024). Many research studies have concluded that the proliferation of AI and machine learning threatens the cybersecurity community (Dash & Sharma 2023; Putra et al. 2024; Vassilev et al. 2024). Heiding et al. (2024) posits that AI-phishing emails have a higher click percentage (30-44%) than traditional phishing emails (19-28%). Additionally, the authors note that "AI significantly increases the incentives to launch phishing attacks by reducing their cost and required revenue." (Heiding et al. 2024). As such, there has been a significant spike in phishing emails since the release of ChatGPT. According to Bolster's *2024 State of Phishing and Online Scam Report*, there's been a 94% increase in phishing attacks since 2020, and there was a 27.8% increase in global phishing activities just from 2022 to 2023 (Bolster 2024). As predicted by Heiding et al. (2024), the reduced costs of AI-phishing have led to an influx of illegitimate emails across the globe. The combination of AI and phishing can be used in scenarios described by Kumar et al. (2024) where a technologically less dependent attacker wants to target a more cyber dependent nation state.

Unfortunately, this increase in AI-phishing emails may have more consequences than just bypassing traditional phishing detection software. Sarno & Neider (2021) analyzed how task factors impacted the accuracy of human identification of phishing emails. The researchers discovered that increasing email load decreased classification accuracy when 50% of the emails were phishing attacks and 50% were legitimate. Participants also had false confidence in their ability to identify phishing emails, and were very "liberal" in their classifications, identifying more emails as legitimate than phishing (Sarno & Neider 2021). Given that Heiding et al. (2024) have observed a spike in phishing emails, AI-phishing emails may not only be hard for humans to classify based on their targeted language; the sheer volume of emails may also decrease the accuracy of human identification.

3.3 Defense Using AI

Given the issues stated above, it is critical that cybersecurity professionals implement AI-phishing specific detection in their products to avoid increased data breaches. Fortunately, some researchers have begun fighting fire with fire by implementing AI in their phishing detection models. Koide et al. (2024) found that LLMs can detect AI-phishing attacks to 98% accuracy. Koide et al.'s (2024) model, called "ChatPhishDetector", utilizes a web crawler to gather information from phishing websites and then generates prompts for LLMs based on the crawled data. Then, the detection model scans for emails that are similar to the responses generated by the LLMs. This system is able to detect multilingual phishing sites by "identifying impersonated brands and social engineering techniques in the context of the entire website," which removes the need to train machine learning models. Similarly, Misra and Rayz (2022) have developed a system that integrates both a multilayer Perceptron (MLP) model and two pretrained Natural Language Processing (NLP) models to detect traditional phishing emails. Misra and Rayz's (2022) model can detect phishing attacks with 97.18% accuracy. Çolhak et al. (2024) also developed a model using NLP, but their research focused on detecting phishing websites rather than phishing emails. Sameen et al. (2020) created a machine learning-based detection system called PhishHaven to identify AI-generated and human-crafted phishing URLs as a response to the deep neural network-based phishing URL generating system called DeepPhish. This model has a 98% success rate when detecting either AI generated or human made phishing urls (Sameen et al 2020). The models discussed here show that ML can be used to detect both AI-generated and traditional phishing attacks to a high degree of accuracy. However, just because these models exist in the literature doesn't mean companies are using AI in their detection models currently. That is what this study aims to evaluate.

4. Methodology

The methodology employed in this study involved three primary approaches: dataset analysis, pilot interviews, and a comparative examination of AI-phishing incidents pre- and post-release of advanced generative models like ChatGPT.

Dataset Analysis: A comprehensive dataset of AI-generated and human-generated phishing emails was collected and analyzed to identify distinguishing linguistic and structural features. Techniques such as Natural Language Processing (NLP) were used to extract patterns and anomalies, which were then compared to conventional phishing indicators. The analysis focused on variations in sentence complexity, tone, and content relevance, aiming to pinpoint attributes unique to AI-generated attacks.

Pilot Interviews: Semi-structured interviews were conducted with nine CISOs across various industries to gain insights into their experiences with AI-driven phishing attacks. The interviews explored perceived levels of

sophistication in AI-phishing emails, current defense strategies, and the effectiveness of AI in their organizations' cybersecurity practices. The qualitative data from these discussions was categorized to highlight common challenges and security gaps.

Comparative Analysis: A historical comparison of AI-phishing incidents before and after the release of advanced models such as ChatGPT was performed. The comparison focused on metrics such as the volume of AI-phishing attempts, success rates, and time-to-detection by traditional systems. This analysis provided a quantifiable measure of how AI models have changed the threat landscape and how defenses need to adapt accordingly.

These combined methods enabled a thorough evaluation of AI-driven phishing threats and provided actionable insights for developing AI-specific defensive strategies.

4.1 Limitations

Human Research approval: The Human research approval takes time and our IRB application is still in process. Hence formal surveys or interviews were not possible. We conducted informal discussions as a pilot study, making sure no personal data was collected and the summary of these discussions to not disclose the identity of the experts we talked to.

Due to ethical reasons, there is a limitation on ChatGPT to create phishing emails. We were limited to small examples of emails, generated by ChatGPT in the context of learning.

5. Findings

Following findings are based on information gathered from journal articles, cybersecurity analysis reports and informal discussions with three Security leaders. For detailed discussions with CISOs we are waiting on IRB approval.

5.1 Attack scenarios

According to the latest State of Phishing assessment report by SlashNext Security (2024), the volume of malicious emails has skyrocketed by a whopping 856% in the last 12 months, with an even more alarming 4151% increase since the launch of ChatGPT in November 2022.

Additionally, the significant differences in Verizon's annual security reports from 2022 and 2024 make the upsurge of phishing attacks and data breaches conspicuous, which can partly be associated with the advancement in AI and introduction of LLMs.

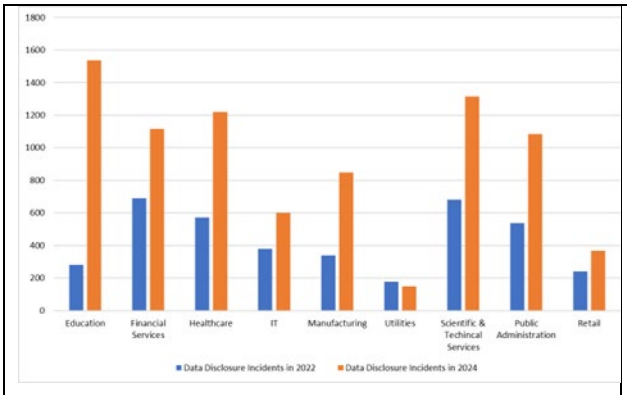

Figure 1: Number of data breaches in major industries between 2022 and 2024

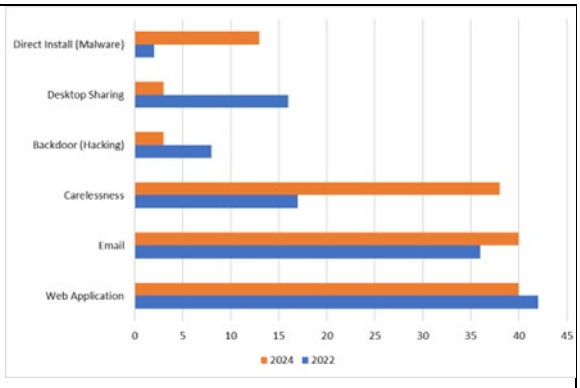

Figure 2: Major Action Vectors in breaches by percentage

(Adapted from Verizon 2022 DBIR and Verizon 2024 DBIR)

Most industries, except Utilities (mining, quarrying, oil extraction etc.) have reported a spike in information leaks, most of them involving system intrusion through hacking and malware. Overall, the percentage of cyber attack incidents covered by Verizon that resulted in data breaches went up from 21.8% in 2022 to 34.9% in 2024, with the contribution of phishing emails rising from 36% in 2022 to about 40% in 2024, highlighting the magnitude of amplification phishing has had when combined with AI.

The biggest phishing attacks from 2022 and 2024 further reinforce the claim. On August 5, 2022, Living Innovations, a global technology firm, fell victim to a social engineering-based phishing attack, one of the top phishing attacks of 2022 (PhishProtection, 2022). Hackers targeted multiple employees by sending phishing emails that tricked them into revealing login credentials. Once inside the network, the attackers gained access to sensitive company information, including intellectual property and operational details.

On August 9, just 4 days later, another phishing data breach occurred when Acorn Financial Services was targeted through an email-based cyber attack, which led to an unauthorized party gaining access to consumer data contained on Acorn's network (JD Supra, 2022). Although both organizations suffered financial losses in millions related to intellectual property compromise, loss of operational data etc., the combined damages paled in comparison to the Pepco Social Engineering attack of February 2024.

The Pepco Group, a major European retailer, faced a phishing attack of unprecedented scale and complexity, reported as a "sophisticated fraudulent phishing attack" by the company. Pepco's breach possibly leveraged AI to craft sophisticated and personalized emails that convincingly targeted the high-level employees. The scope of exposed data was much broader, including not only detailed personal data of the employees and customers but critical infrastructure information as well. As a result, Pepco lost approximately €15.5 million in cash ($17.3 million) and witnessed a decrease in its stock market value (Help Net Security, 2024). The breach was of such extent that it was declared to pose a national security risk. This serves as testament to the fact that AI has enhanced modern phishing into a far more dangerous threat.

5.2 Informal Discussion with CISOs

Based on our pilot study, talking to nine security experts, organizations are currently facing an increased number of sophisticated phishing attacks, particularly Business Email Compromise (BEC), fake login pages, and SMS-based phishing. AI-generated phishing emails are more well-tailored than traditional phishing emails, leveraging personal information such as job titles, trusted contacts, and social media activity to create highly targeted and convincing messages. The respondents also mentioned that AI tends to mimic senior management requests, understand context, and adjust tactics to evade detection.

The emergence of AI has led to the use of deep fakes, advanced social engineering, and more personalized content in phishing campaigns. While the volume of AI-driven phishing emails has not yet significantly increased, it is expected to scale rapidly. The future of AI-driven phishing is likely to involve integration with other cybercrimes, resulting in more damaging and widespread attacks. The automation of phishing campaigns, combined with AI's ability to evade detection, makes these threats more dangerous and accessible to cybercriminals, posing a significant challenge to cybersecurity defenses.

5.3 Defense Scenarios

Organizations are at varying stages of implementing AI to detect and prevent AI-generated phishing attacks. While some have not yet adopted AI in their detection models, there is a strong sense of urgency to do so, given AI's potential to enhance threat detection and response capabilities. For now, most companies are focusing on employee training and awareness campaigns as a first line of defense.

In response to the rise of AI, organizations are encouraging employees to understand its implications and are building internal capabilities. Some have established Centers of Excellence (CoE) for AI and developed in-house tools like AI-based code generation and log parsers. AI-driven defenses are expected to revolutionize cybersecurity by enabling predictive analysis of threats, automated detection and response, and continuous vulnerability assessments. When implemented at scale, these solutions can mitigate risks and minimize damages from AI-powered attacks, providing a proactive approach to securing the organization's digital assets.

In order to evaluate how current email security software has fared against the onslaught of AI-phishing attacks, this paper compared the accuracy rates of the most prominent email security companies from 2022 to 2023. The data was aggregated from the annual security reports by SE Labs, a UK based company that specializes in testing and evaluating cyber security solutions. Since ChatGPT was released in November of 2022 and the SE Labs 2022 report was released in June, this data gives a good snapshot of the height of email security software before AI became mainstream. SE Labs had not yet released their 2024 Email Security Report as of the writing of this paper, so this research uses the 2023 report.

These security reports evaluated how the most popular email security softwares fared against different types of phishing attacks. Additionally, SE Labs calculated the accuracy of each software to detect legitimate emails, to protect against phishing attacks, and then the total accuracy with all factors accounted for.

Table 1: Average Accuracy Rates of the Most Popular Email Security Companies in 2022 and 2023

	2023 Average	2022 Average	2023 Avg.- 2022 Avg.
Total Accuracy	67.2%	77.75%	-10.55%
Protection Accuracy	61%	70.75%	-9.75%
Legitimate Accuracy	93.8%	91.5%	+2.3%
Business Email Compromise (BEC)	49.8%	69%	-19.2%
Social Engineering	31.8%	80.75%	-48.95%
Malware	70.2%	78.75%	-8.55%

(Adapted from SE Labs 2022 Email Security Services Report and SE Labs 2023 Email Security Services Report)

In all categories except for legitimate accuracy, on average, email security software has performed worse from 2022 to 2023. Overall, the total accuracy shows that prominent email security software has decreased in accuracy by 10.55%. When it comes to the percentage of illegitimate emails that these softwares identify (Protection Accuracy), they performed worse against AI-phishing emails than traditional emails. However, these softwares have gotten marginally better at identifying legitimate emails and have a reduced amount of false positives (Legitimate Accuracy).

In terms of the specific types of phishing attacks, the popular softwares perform worse against all versions when it comes to AI-phishing emails. Particularly, they decreased in accuracy by 48.95% on Business Email Compromise (BEC) phishing attacks. This sentiment was echoed in this paper's informal discussions with CISOs, which is discussed further below.

5.3.1 AI offerings for email defense products

We analyzed a range of email security tools, synthesizing data from multiple lists of the most prominent email security tools. Since our aim is not to focus on any particular brand or endorse it, we randomly selected two from an averaged list of top 10 products for deeper analysis: Proofpoint and Mimecast. Mimecast and Proofpoint are both leading cybersecurity companies leveraging artificial intelligence (AI) to enhance protection against sophisticated cyber threats.

5.3.2 case study: Mimecast

Mimecast focuses on AI-powered threat detection and response systems designed to intercept the most sophisticated cyber threats, such as Business Email Compromise (BEC), phishing, and data breaches. The company's approach integrates natural language processing (NLP), social graphing, and computer vision to identify anomalies and detect even the most subtle phishing attempts. By using AI at scale, Mimecast's solutions are able to dynamically learn from new threats and continuously improve the detection accuracy over time. Their AI is layered onto existing security measures, enhancing overall efficacy without relying entirely on automation. This multi-faceted approach ensures a high level of threat protection while maintaining the adaptability needed to counter emerging cyber risks. Mimecast's AI technology is particularly well-suited for organizations looking for comprehensive email and communication security solutions.

5.3.3 Case study: Proofpoint

Proofpoint utilizes AI technologies to protect sensitive data across various platforms, including email, cloud applications, and social media. Their AI models analyze behavioral patterns and contextual signals to detect suspicious activities and prevent data exfiltration. The company emphasizes the use of machine learning

algorithms to identify potential threats before they can cause damage, and also provides real-time monitoring and incident response capabilities. Proofpoint's AI technologies are specifically designed to integrate seamlessly with existing security infrastructures, making them a practical choice for organizations aiming to bolster their defense mechanisms against AI-driven cyber attacks.

Mimecast and Proofpoint are both leveraging AI technologies to provide robust cybersecurity solutions, but their approaches and focus areas vary significantly. Mimecast's AI-driven solutions emphasize comprehensive threat detection for email and communication systems. By using advanced techniques like natural language processing (NLP), social graphing, and machine learning, Mimecast's AI models can detect subtle anomalies that indicate sophisticated phishing attempts or Business Email Compromise (BEC). The company's focus on integrating AI with traditional cybersecurity measures enables a layered security approach, which helps reduce false positives and provides a balanced, adaptive defense against a broad spectrum of cyber threats.

In contrast, Proofpoint's AI technologies focus on protecting sensitive data across multiple platforms, including email, cloud services, and social media. Proofpoint utilizes machine learning algorithms to analyze user behaviors and detect anomalous activities indicative of cyber threats. The company's AI solutions are built to prevent data exfiltration and to provide real-time monitoring, making it particularly strong in protecting against data breaches and insider threats. Proofpoint's integration of AI into existing security infrastructures makes its solutions flexible and easily adaptable for organizations looking to enhance data protection without overhauling their current systems.

In summary, while both Mimecast and Proofpoint use AI to enhance cybersecurity, Mimecast focuses more on threat detection and communication security, whereas Proofpoint specializes in data protection and preventing data loss across various platforms. Both companies highlight that AI should complement existing security measures, providing organizations with a multi-layered defense against evolving cyber threats.

5.4 Informal Discussions with CISOs

The CISOs agreed that AI-driven defense could be very promising; however, multiple warned that such models would require lots of data to work on a large scale. This lack of data seems to be the underlying reason behind why most companies are emphasizing on phishing awareness training programs, while simultaneously trying to implement AI into their current detection softwares.

Organizations are at varying stages of implementing AI to detect and prevent AI-generated phishing attacks. While some have not yet adopted AI in their detection models, there is a strong sense of urgency to do so, given AI's potential to enhance threat detection and response capabilities. For now, most companies are focusing on employee training and awareness campaigns as a first line of defense.

In response to the rise of AI, organizations are encouraging employees to understand its implications and are building internal capabilities. Some have established Centers of Excellence (CoE) for AI and developed in-house tools like AI-based code generation and log parsers. AI-driven defenses are expected to revolutionize cybersecurity by enabling predictive analysis of threats, automated detection and response, and continuous vulnerability assessments. When implemented at scale, these solutions can mitigate risks and minimize damages from AI-powered attacks, providing a proactive approach to securing the organization's digital assets.

Our findings indicate that AI-generated phishing emails are significantly more challenging to detect and are more effective at evading traditional security mechanisms compared to human-generated phishing attempts. And the analysis of security products indicates a trend towards strong AI-based offerings.

6. Future Work

Future research will focus on refining AI-based defense strategies to improve the detection and prevention of AI-generated phishing attacks. This includes developing predictive models that can identify emerging phishing patterns and integrating AI with real-time threat intelligence to enhance response capabilities. Additionally, the use of large language models (LLMs) will be explored to automate the identification of phishing emails, analyze attacker behaviors, and provide adaptive learning systems that can evolve with the threat landscape. Expanding the research to include more diverse datasets and conducting extensive evaluations across different sectors will provide a more comprehensive understanding of AI's role in both offensive and defensive cybersecurity applications. Once we receive the human research approval, we plan to conduct formal and detailed interviews with security experts, practitioners and product designers.

7. Conclusion

The findings of this research indicate that AI-generated phishing and cyber threats are becoming increasingly sophisticated, making traditional cybersecurity defenses insufficient. The study demonstrates that AI-powered attacks can easily evade detection by conventional security mechanisms, thereby escalating the risk of data breaches, financial losses, and organizational disruptions. This evolving threat landscape underscores the necessity for organizations to proactively adapt their cybersecurity strategies to include AI-driven defenses that are capable of identifying, analyzing, and mitigating these advanced threats in real time.

To address this challenge, it is recommended that organizations adopt a multi-layered approach incorporating AI-based threat detection and response systems alongside traditional security measures. Implementing such advanced technologies will enable organizations to more effectively anticipate and counteract AI-enabled attacks. This approach not only enhances detection capabilities but also strengthens overall resilience by enabling rapid adaptation to new and emerging threats.

While integrating AI-driven defenses may require significant initial investment and strategic planning, the long-term benefits in terms of reduced risk and enhanced security far outweigh the costs. Therefore, it is imperative that industry leaders, policymakers, and cybersecurity professionals prioritize the adoption of AI-based solutions to safeguard their digital assets and ensure the stability and integrity of their operations in an increasingly complex cyber environment. A failure to act promptly could leave organizations exposed and vulnerable to the next generation of AI-powered cyber attacks, making immediate action essential to maintaining robust cybersecurity postures.

References

Barrientos, F., Jacobs, J., and Dawkins, S. (2021). S*caling the phish: Advancing the NIST phish scale*. In HCI International 2021-Posters: 23rd HCI International Conference, HCII 2021, Virtual Event, July 24–29, 2021, Proceedings, Part II 23 (pp. 383-390). Springer International Publishing.

Canham, M., Dawkins, S., and Jacobs, J. (2024, June). *Not All Victims Are Created Equal: Investigating Differential Phishing Susceptibility*. In International Conference on Human-Computer Interaction (pp. 3-21). Cham: Springer Nature Switzerland.

Çolhak, F., Ecevit, M.İ., Uçar, B.E., Creutzburg, R., and Dağ, H. (2024). *Phishing Website Detection through Multi-Model Analysis of HTML Content*. arXiv preprint arXiv:2401.04820.

Dash, B., and Sharma, P. (2023). *Are ChatGPT and deepfake algorithms endangering the cybersecurity industry? A review*. International Journal of Engineering and Applied Sciences, 10(1).

Dawkins, S., Jacobs, J., and Neil, L. (2023). *Peering into the phish bowl: An analysis of real-world phishing cues*. Indicator, 43, pp.72-88.

Greene, K., Steves, M., and Theofanos, M. (2018). *No phishing beyond this point.* Computer, 51.

Greene, K.K., Tamborello, F.P. (2017). *Exploratory lens model of decision-making in a potential phishing attack scenario*. US Department of Commerce, National Institute of Standards and Technology.

Heiding, F. et al. (2024). *Devising and detecting phishing emails using large language models*, IEEE Access, 12. doi:10.1109/access.2024.3375882.

Ho, G., Cidon, A., Gavish, L., Schweighauser, M., Paxson, V., Savage, S., Voelker, G.M., and Wagner, D. (2019). *Detecting and characterizing lateral phishing at scale*. In 28th USENIX security symposium (USENIX security 19).

Jain, A.K., and Gupta, B.B. (2016). *A novel approach to protect against phishing attacks at client side using auto-updated white-list*, EURASIP Journal on Information Security, 2016(1). doi:10.1186/s13635-016-0034-3.

Koide, T., Fukushi, N., Nakano, H., and Chiba, D. (2023). *Detecting phishing sites using chatgpt*. arXiv preprint arXiv:2306.05816.

Kumar, S., Kocian, L. and Loo, L. 2024. *Blockchain Applications for Cyber Liability Insurance*. 2nd International Conference on Blockchain, Cybersecurity and Internet of Things, BCYIoT 2024.

Kumar, S., and Nagar, G. (2024). *Threat Modeling for Cyber Warfare against Less Cyber-dependent Adversaries*. 23rd European Conference on Cyber Warfare and Security (ECCWS), 27-28 June, Jyväskylä, Finland.

Mimecast (2024). *AI-powered security organizations can trust*, *Ai Cybersecurity*. Available at: https://www.mimecast.com/ai-cybersecurity/ (Accessed: 29 September 2024).

Misra, K. and Rayz, J.T. (2022). *LMS go phishing: Adapting pre-trained language models to detect phishing emails*. 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) 3, pp. 135–142. doi:10.1109/wi-iat55865.2022.00028.

Proofpoint (2023) *Proofpoint AI technologies that protect sensitive data*, *Proofpoint US*. Available at: https://www.proofpoint.com/us/resources/solution-briefs/proofpoint-ai-technologies-protect-sensitive-data (Accessed: 29 September 2024).

Putra, F.P.E., Zulfikri, A., Arifin, G., and Ilhamsyah, R.M. (2024). *Analysis of Phishing Attack Trends, Impacts and Prevention Methods: Literature Study*. Brilliance: Research of Artificial Intelligence, 4(1), pp.413-421.

Sameen, M., Han, K., and Hwang, S.O. (2020). *Phishhaven—an efficient real-time AI phishing urls detection system*, IEEE Access, 8. doi:10.1109/access.2020.2991403.

Sarno, D.M., and Neider, M.B. (2021). *So many phish, so little time: Exploring email task factors and phishing susceptibility*, Human Factors: The Journal of the Human Factors and Ergonomics Society, 64(8), pp. 1379–1403. doi:10.1177/0018720821999174.

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., and Hall, P. (2022). *Towards a standard for identifying and managing bias in artificial intelligence* (Vol. 3, p. 00). US Department of Commerce, National Institute of Standards and Technology.

Steves, M.P., Greene, K.K., and Theofanos, M.F. (2019). *A phish scale: rating human phishing message detection difficulty*. In Workshop on usable security (USEC).

Vassilev, A., Oprea, A., Fordyce, A., and Anderson, H. (2024). *Adversarial machine learning: A taxonomy and terminology of attacks and mitigations* (No. NIST Artificial Intelligence (AI) 100-2 E2023). National Institute of Standards and Technology.

Yasin, A., and Abuhasan, A. (2016). *An intelligent classification model for phishing email detection*. arXiv preprint arXiv:1608.02196.

Verizon (2022) *2022 Data Breach Statistics by industry*, *Verizon Business*. Available at: https://www.verizon.com/business/resources/reports/dbir/2022/intro-to-industries/ (Accessed: 27 September 2024).

Verizon (2024) *DBIR report 2024: Industries - introduction*, *Verizon Business*. Available at: https://www.verizon.com/business/resources/reports/dbir/2024/industries-intro/ (Accessed: 27 September 2024).