

# AI-Driven Phishing: Techniques, Threats, and Defence Strategies

Liza Shrestha<sup>1</sup>, Hamed Balogun<sup>1</sup>, and Suleman Khan<sup>2</sup>

<sup>1</sup> University of Central Lancashire, United Kingdom  
{lshrestha,hbalogun1}@uclan.ac.uk

<sup>2</sup> University of Bradford, United Kingdom  
s.khan387@bradford.ac.uk

**Abstract.** Phishing attacks are one of the most challenging threats in the cyberspace. Recently, rapid advancements of (generative) AI and its wide applicability has been exploited by attackers to perform phishing attacks. However, limited studies exist regarding AI-based phishing and apt defence strategies. In this work, we explore a wide variety of AI-leveraging techniques, adopted by attackers to conduct successful phishing campaigns. Additionally, we highlight the negative impact of AI-driven phishing in real-world and the attendant challenges that it has on cybersecurity. We also examine various factors and features of AI-powered phishing which makes these difficult to identify and complicated to defend. Consequently, we evaluate the crucial aspects of phishing attacks and discuss its defence strategies, human-centred preventive measures, and ethical considerations for enhancing security against AI-based phishing threats. Our findings provide valuable insights to the evolving cybersecurity threats and effective approaches to defend against these sophisticated AI-driven attacks.

**Keywords:** AI · Phishing · Detection · Techniques · Defence · Security

## 1 Introduction to AI-Powered Phishing

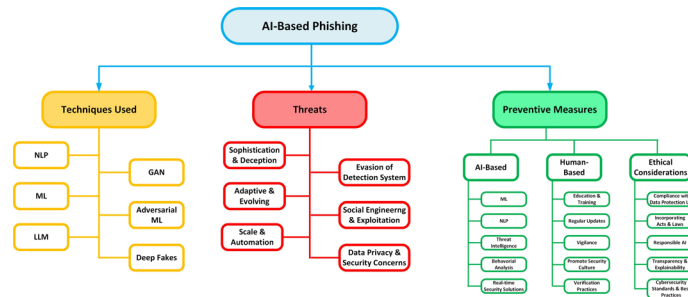
Phishing is a prominent social engineering attack where an individual or a targeted group are deceived and manipulated into revealing sensitive information [1]. The means and methods by which attackers initiate phishing has been evolving rapidly such that existing solutions against phishing are unable to keep up with the advanced technologies used by attackers, thereby making them less effective [2]. One of the notable ascents regarding phishing and other cybercrimes is the use of Artificial Intelligence (AI). From the past trends of cybercrimes, it is evident that attackers are drawn towards utilising new technologies making AI as a priority tool for commencing their mischievous activities [3]. AI has transformed the patterns of phishing attacks in terms of sophistication and highly personalised attacks [4]. Automation is another advantage provided by AI which impose less work on attackers, allowing them to conduct large scale attacks effortlessly in short time frame [5]. The popularity and functions provided by AI

has made it a favourite tool for cyberattackers. Machine learning (ML) algorithms, Natural Language Processing (NLP), Generative Adversarial Network (GAN) are some of the AI techniques that have been widely used for notorious activities. Compared to other cybercrimes, attackers are more influenced by AI in committing social engineering attacks like phishing. Generative AI has brought a tremendous upgrade in the magnitude of phishing activities [6]. The dual nature of AI pertinently contributes to it being a choice tool for attackers [7].

### 1.1 Scope of the work

This paper focuses on exploring and analysing the AI-techniques used for conducting AI-driven phishing attacks, the threat it poses and effective defence strategies as shown in Fig.1. The aim is to provide an understanding of evolving AI-driven phishing attacks and highlight its impacts on cybersecurity. The challenges involved in preventing these attacks have been analysed and AI-based multi-modal detection framework have been proposed along with human-focused countermeasures by evaluating the existing preventive measures and emerging technologies. Key contributions of this paper are:

- Highlighting the growing use of AI techniques for sophisticated phishing attacks and its impact.
- Illuminating the urgent need for smart solutions for AI-driven phishing.
- Proposal of a novel AI-based multi-modal detection framework to efficiently combat the adverse effects of AI-based phishing.



**Fig. 1.** Techniques, threats and preventive measures of AI-driven phishing.

## 2 Motivation

According to Anti-Phishing Work Group (APWG), in 2023 approximately 5 million phishing attacks were recorded, and this is expected to increase each year [8]. The increasing attack rates indicates that the existing prevention tools

are not fully effective towards the evolving phishing attacks. With the popularity of generative AI and other AI-based tools, their misuse has also been dramatically increased [9]. However, inadequate attention has been provided to AI-based attacks [4], specifically AI-based phishing attacks, which is the main motivation for this paper. It is crucial to understand the scale of damage AI-based phishing can impose on cybersecurity. Generative AI and other AI techniques not only leverage the intensity of existing phishing attacks but also help in implementing innovative zero-day phishing attacks [10]. Hence, posing great threat to the overall cybersecurity landscape. This paper aims to provide insights to the underlying misuse of AI to conduct hostile phishing attacks. It covers currently used AI-techniques and explores other plausible misuse of AI-techniques that can enhance the effectiveness of phishing attacks. Moreover, AI-features and challenges which makes AI-powered phishing less susceptible to detect and prevent are also discussed.

### 3 Background/ Related Work

The history of phishing attacks is recorded from early 1990s when American Online (AOL) users were tricked by hackers who were pretending to be AOL employees [11]. The users were deceived into revealing their personal information, login credentials and credit card details. In early 2000s, email phishing was introduced for stealing personal information [12]. This trend took a leap in mid 2000s through spear phishing which was more targeted and personalised form of phishing [13]. And in the late 2000s, whaling and clone phishing were conducted which were more sophisticated compared to previous attacks.

The advancement of technology and automation were reflected in mid to late 2010s. New forms of phishing attacks such as pharming which redirected victims from legitimate websites to fraudulent ones were conducted. These types of phishing were executed by compromising DNS servers or with the use of malicious software. One of the other fraudulent attacks initiated in that time is man-in-the-middle attacks using automated phishing tools. Automated phishing tools comprised of sophisticated tools and scripts which made the large-scale campaigns easier to accomplish. The data records from 2010s shows a drastic increase in the statistics of successful phishing attacks each year. Since then, attackers have been using advanced tools to conduct phishing attacks with various motives. Some of the popular tools were URL shorteners and obfuscation which kept the malicious links hidden in the emails. Similarly, various other techniques were used to evade spam filters and email security software.

There is no evidence for the exact date when the use of AI in phishing began. It was a gradual process and with time large numbers of AI-driven phishing were encountered, where attackers started to experiment with AI and machine learning algorithms to enhance phishing attacks [14]. Focus was seen on improving email content and target selection in the initial use by cybercriminals. As the growth of AI in other fields increased exponentially, it also highly influenced the attackers. The rise of AI in the early 2020s brought a revolutionary leap in the

field of technology. AI not only transformed the digital world with extraordinary applications, but it also provided support for cyberattackers. The misuse of AI by attackers for various cybercrimes has been one of the major challenges for cybersecurity professionals [14]. Matured social engineering techniques, COVID pandemic and deep learning techniques are the reasons for massive increase in phishing attack. From cybercrimes like intrusion, hacking, malware attack and phishing attacks, AI have aided attackers to commence them in more sophisticated way [15].

In [16], misuse of AI in cyber-attacks was discussed in 3 categories which are synthetic data generation, data analysis, and data misclassification, with significant implications for phishing. Natural Language Processing (NLP), GANs, and ML are some of the widely used AI techniques. NLP helps in generating highly convincing phishing emails by accurately mimicking legitimate communications to [17]. For example, attackers used AI to impersonate an employee of MGM Resorts in a call to IT service desk to trick the IT team into resetting employee's password. They were successful in gaining network access which led to a massive ransomware attack. Similarly, GANs have been used to generate synthetic phishing emails both by by cybersecurity researchers and attackers [18]. Various ML algorithms have also been used for crafting personalised phishing emails that have high success rates.

With a rapid digital transformation, Generative AI (GenAI) has brought massive enhancement in the utilisation of technology and its applicability to perform various tasks which includes both use and misuse. Researchers have discussed the potential of generative models such as ChatGPT to cause a negative impact on cybersecurity. For instance, authors in [19], discussed the use of ChatGPT by attackers to perform social engineering/phishing attacks ('spear phishing'). Thus, the use of AI tools and techniques provides an easy path for 'unsophisticated' attackers to draft convincing phishing emails [20] and potentially carry out effective attacks.

Previously, phishing attacks had numerous weakness which made their detection easier. Some of the obvious features of traditional phishing attacks are poor grammar, unprofessional formatting and spelling mistakes [21]. Other distinctive features were lack of personalisation and links to newly created or suspicious domains. All these features made the detection of phishing attacks simple. It could be easily detected with the defensive tools which were designed to identify these obvious patterns. However, this has substantially changed because AI has altered the characteristics of phishing attacks by advancing its features and overcoming most of these weaknesses [22]. It has become more difficult to identify phishing attacks looking at the features, for example, because AI-based phishing attacks are grammatically correct with zero spelling errors along with high personalisation [23]. Publicly available LLM models with the ability to generate human-like texts have made phishing attack scenarios worse than ever [24]. Attackers with poor writing skills can also easily draft high quality phishing emails with ease. LLMs such as ChatGPT and Microsoft Copilot can generate highly convincing phishing emails with just one simple instruction. A stimulated experiment

conducted by Singapore's Government Technology Agency where spear phishing emails were sent by the security team resulted in the internet users clicking the links in AI-generated emails rather than in human-written ones [25].

It is important to understand how these magnificent AI tools affect, positively as well as negatively, our overall cyberspace and mankind. The more details that we acquire about how AI can be utilised adversely, the easier it will be to come up with effective countermeasures against them. In identifying effective countermeasures, considering each AI technique has unique features, it is important to study the adverse use of individual techniques in phishing attacks rather than generalizing AI as a whole.

## 4 AI Techniques Used in Phishing

The use of AI-techniques by attackers to commence phishing attacks varies vastly and are constantly evolving. From the wide range of techniques that have been implemented for malicious purpose, attackers select the technique that help them conduct more sophisticated and personalised phishing attempts.

### 4.1 Natural Language Processing

NLP helps computers to understand and interpret human language using various machine learning and can act as a mediator between computers and human linguistics which includes both text and speech [26]. It can be used to refine traditional phishing emails for generating more convincing human generated emails with less errors [27]. NLP has also shown its ability to analyse text from large volume of social media data [28] which can be adversely used by attackers for commencing targeted phishing attack. For a successful phishing, sentimentally and contextually analysed emails and conversations play a vital role as it can manipulate victim more effectively. NLP can be used to conduct sentimental analysis in dialogues based on various context to determine emotional tone during a conversation [29] making victims more vulnerable in being tricked. For instance, in an article [30], NLP was used to extract numerous contextual information from social media. Considering the volume of data that an individual shares in social media platform, attackers can easily extract them to analyse and design a customised spear phishing with high success rate.

### 4.2 GANs

GAN, a combination of two neural networks (generator and discriminator), can create synthetic data taking references from training data. There are studies showing the potential of GAN in generating different types of data including text, images, and more. For example, in an experiment by Mahiuddin, Md, et al. [31], a real-life like synthetic images of human faces were generated with the use of GAN. This proficiency of GAN to generate realistic data has not just contributed to numerous fruitful aspects but has also led to increase in its

utilisation for commencing various cyber-attacks such as phishing. The ability of GAN to generate realistic sentences have been explored widely especially TextGAN [32]. It is not limited to generating sentences but larger text generation like essays can be achieved with proper trainings [33]. This depicts that it be used for drafting realistic emails in more efficient and automated form for targeted phishing attacks.

### **4.3 Machine Learning**

The scale of damage is much higher in targeted attacks conducted according to target's preference and behaviour compared to a generalised attack. Machine learning algorithms have made the analysis of target's behaviour simple and robust. For example, Activision, the company behind Call of Duty games became a victim of targeted phishing campaign in which attackers used AI to create highly convincing SMS messages [34]. This led one of the staff member to fall for the bait and hackers were able to gain complete access to employee database. User profiling, which is a form of understanding user's behaviour, interests and preferences, plays a vital role in targeted attacks [35]. AI-based user profiling has been utilised for digital marketing by focusing on the customer's need [36]. However, the application of this process is not limited to digital marketing but can also be a systematic approach for targeted adaptive social engineering. Both supervised and unsupervised ML can be used for conducting user profiling. Some researchers [37] have demonstrated the generation of fake text documents with the use of genetic algorithm depicting the capability of ML in making fraud text which can be utilised by attackers for phishing.

### **4.4 Adversarial Machine Learning**

There has been increase in the preferences of AI tools over traditional methods for detecting the phishing attacks which have simultaneously increased the ML-based adversarial attacks on these AI tools. No system can be full proofed with no vulnerabilities. Systems may be vulnerable to evasion attacks which have been seen in an experiment by [38], where 12 evasion attacks were conducted against a phishing website detector and the results demonstrated that it was affected by few attacks. Similar experiments have also been performed to bypass phishing website classifier using three mutation-based attacks [39]. These experiments were conducted to analyse the effectiveness and accuracy of phishing detector tools; however, it also depicts the possibility of attackers conducting similar adversarial attacks making the detector tools ineffective with increase in successful phishing attempts.

### **4.5 Large Language Models**

Large language models (LLMs) can generate convincing phishing emails using simple prompts. Crafting phishing emails has been easier than ever with the

use of LLMs. Experiments have been conducted to demonstrate the ability of LLMs to contribute toward social engineering and phishing attacks with simple prompts [40]. Attackers can misuse these wide variety of LLMs that have special features to heavily customise an initiative for advanced spear phishing. Although most of the publicly available LLMs have been designed with ethical guidelines and restrictions for creating adversarial contents, they can be manipulated using tricky prompts and fool them to generate malicious contents. For instance, if we provide a direct instruction to create a convincing email, it will deny for generating. However, if we modify the instruction to generate a sample of phishing emails for educational purposes, it will generate one. There are some LLMs which have been launched for cybercriminals such as WormGPT and FraudGPT which allows creation of phishing contents with very less effort [41].

#### **4.6 Deep Fakes**

Deepfake technology which builds realistic but counterfeit audiovisual contents using various AI algorithm, is an emerging threat as it has been misused for phishing and vishing. In the paper [42], the use of deep fakes for various deceptive purposes such as false advertising and malicious agendas have been discussed along with the problems in regulating them ethically and lawfully. It has created a serious problem in terms of security compromise in the digital space [43]. This technology has been used specifically for ‘spear phishing’ by deceiving an individual into revealing sensitive information. This form of phishing is generally carried out by fooling the victim through audio or video messages containing impersonation of a trusted individual. One such example is Fraudsters using deepfake technology to trick a finance worker into transferring \$25 million by impersonating company’s chief financial officer through a video conferencing call. With rapid advancement in deep fake technology, number of deep fake phishing attacks are expected to increase causing serious issue if not regulated [44].

## **5 Threats of AI-Powered Phishing**

### **5.1 Sophistication and Deception**

The sophistication of phishing attacks has increased tremendously with the use of AI [4]. Contextual understanding can lead to convincing conversation between attacker and victim which enhances the credibility of phishing attempts [45]. It is difficult for attackers to achieve contextual understanding through traditional phishing techniques but with different AI techniques it has become quite a simple task. AI generated phishing is grammatically correct with less to no typos such that its detection is difficult using existing phishing detection methods [46]. These attacks can also be orchestrated across multiple channels simultaneously and in a coordinated manner elevating the deception chances and successful attacks.

## 5.2 Adaptive and Evolving Tactics

One of the major differences between traditional phishing attacks and AI-based phishing attacks is the adaptability of AI phishing due to its dynamic nature [47]. The currently available static phishing defense system is not efficient enough to keep pace with the rapidly changing AI-based phishing. For example, existing system trained to detect phishing email based on spelling error and poor grammar cannot detect the phishing email with no flaws generated using AI. Zero-day attacks are the most challenging issue for any cybersecurity solution [48] and AI helps in generating and executing these immune zero-day phishing attacks making the existing defense system less effective.

## 5.3 Scale and Automation

If we look at the history of previously conducted phishing attacks, large scale attacks were not common as they are time consuming and expensive. But AI has been found to be a great way for attackers to perform mass phishing with less resources and short time-period. AI can generate numerous phishing emails with automated operations which can have massive damage in a very short period [5]. Also, attackers can now set a specific time to send phishing email automatically using various ML algorithms. This has increased the seriousness of the attacks as the damage they make is vast with limited time to act for recovery.

## 5.4 Evasion of Detection Systems

The traditional or currently implemented phishing detection tools are designed against previously conducted phishing patterns which will not effectively detect AI-powered phishing with dynamic nature. The evolving AI-generated phishing is unique for each attack which can easily bypass the existing phishing detectors. Research conducted shows that AI can help attackers to evade spam filters along with the security plugins by mimicking communications which resemble legitimate ones [49]. These actions are time consuming to perform manually but with the help of AI, it is quite efficient.

## 5.5 Social Engineering and Exploitation

The amount of personal data that an individual shares on the internet these days is like threats to the attackers for conducting personalised attacks [50]. However, extracting and analysing these data manually can easily analyse the vast amount of data by collecting them from social media and other online activities. Other ways in which AI-based phishing can enhance social engineering is through psychological profiling and emotion analysis [51]. Data collected related to the target can be psychologically profiled for understanding vulnerabilities and preferences which accounts to creation of highly convincing attacks. Since emotions are the primary factor which influences the chances of successful social engineering, AI can accommodate in tailoring phishing attempts by analysing emotional responses for higher success rate in the attacks.

### 5.6 Data Privacy and Security Concerns

Data privacy and security concerns coupled with AI-based phishing are immaculate mainly regarding the misuse of AI models and unintended consequences [52]. Effective and widespread attacks have increased with AI allowing less-skilled attackers to deploy sophisticated schemes. AI tools have helped to escalate the data breaches on a very large scales where victims lose their sensitive details that can be sold on the dark web or used for various other fraudulent activities. There are many other security and privacy concerns which can be classified into user, ethical, technological, institutional, regulatory and law perspectives [53]. All of these can lead to significant unintended consequences along with the amplification of risks associated with data breaches.

## 6 Challenges and Risks of AI-Powered Phishing

The essence of attack is the same even with the use of AI, but the scope and scale of the attacks are contrasting. Analysing the threats imposed by AI-based phishing, the differences between them with traditional phishing can be highlighted as shown in Table 1.

Features	Traditional Phishing	AI-Driven Phishing
Customisation	Low	High
Automation	Limited	Extensive
Sophistication	Basic, Template-Based	High, Personalised, Adaptive
Success Rates	Moderate	High
Detection Difficulty	Easier to Detect	Harder to Detect
Evolution Speed	Slow	Rapid

**Table 1.** Differences between Traditional and AI-based phishing.

Traditional phishing provides less opportunity for customising the phishing emails according to the targeted individual/group and the customisation process is extensive and difficult [54]. However, it is not the same in case of AI-driven phishing, where the customisation process is easily conducted based on the log activities and interest of the selected target. Conventional phishing emails were drafted manually limiting the scale and frequency of the attacks. But in AI-based phishing, there is a wide availability of automation allowing campaign and large-scale phishing which are more destructive in nature to be carried out effortlessly [5]. Previously, phishing emails were crafted by using basic sentences and were mostly template-based [55] making them less sophisticated, while with the help of AI, phishing emails are becoming more sophisticated, adaptive, and personalised. These are some of the reasons why detection of traditional phishing attacks was less successful with easy detection and recent AI-based phishing attacks have high success rate with less chances of being detected. The evolution

of traditional phishing was slow as they were manually induced which provided ample amount of time for the development of new defences. However, AI-based phishing is evolving rapidly and continuously offering less time for developing the defence systems against them.

## 7 Defense Strategies Against AI-Driven Phishing

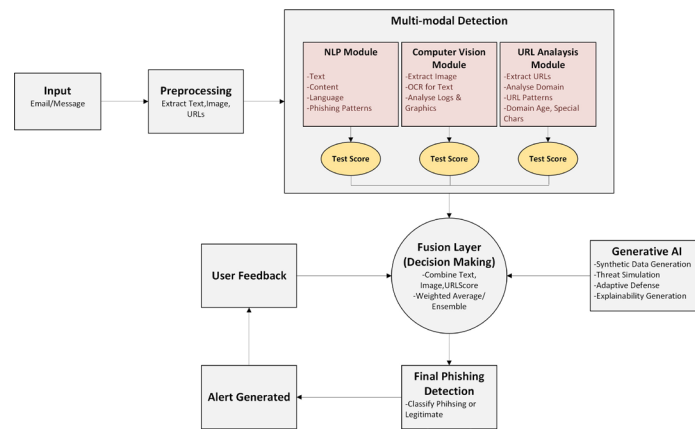
There are various ways to detect and prevent AI-based phishing attacks but with the pace at which AI-powered attacks are evolving, the only means by which these attacks can be tackled more efficiently is with the use of AI itself [14]. Modern phishing attacks are sophisticated and exploit multiple types of content such as text, images, and URLs [56].

### 7.1 Technical Measures

**Advanced Phishing Detector Tools** One of the efficient ways of detecting and preventing phishing attacks is to develop machine learning and AI-based smart detector tools which can adapt according to the changing nature of AI-phishing attacks. For example, various companies have integrated phishing detection tools such as Vectra Cognito with Amazon Web Services (AWS) to enhance the cybersecurity through real-time monitoring and response in the network [57]. Some of the other phishing detector tools such as Bacrracuda Essentials, Graphus, Hunto AI are popular among organisations as they claim to be more effective than the traditional tools. AI-driven solutions can drastically improve detection and prevention of advanced phishing attacks and this has been shown through case studies and real-world scenarios in the paper by H. N. Fakhouri et. al [58].

ML algorithms such as convolutions neural network (CNN), recurrent neural network (RNN), Random Forest have been explored by researchers for detecting phishing attacks. The experiment performed by Safi, A. et. al. and Sahingoz, O.K. et. al. depicts that CNN has the highest accuracy of 99.98% and 98.74% in respective experiments where multiple algorithms were tested [59] [60]. Algorithms with high detection accuracy and low false positive rates can be trained to detect AI-based phishing as well. Existing AI detectors for phishing can be incompetent when it comes to identifying zero-day attacks [61], making them less effective. It has become essential to develop and test hybrid AI-models which can identify zero-day attacks because hybrid models have been found to possess high accuracy compared to other models. Also, some of the AI-techniques such as reinforcement learning are adaptive, and this feature can be utilised against new tactics used in phishing. By selecting appropriate machine learning and AI techniques, their features can be combined to develop highly accurate phishing detector tools. Along with hybrid models, a holistic approach of combining AI-based tools with other traditional methods can improve the defence strategy against phishing [62].

**Multi-modal Approach with Generative AI** Multi-modal approaches for phishing detection have gained significant popularity in recent researches due to its improved accuracy, robustness, and scalability. In the research performed by Çolhák, F. et. al., multiText-LP model performance was superior to standalone model in detecting phishing emails through HTML content analysis with an accuracy of 96.80% [63]. To ensure all crucial cues embedded in different formats are properly analysed, multi-modal AI-approach has been proposed for phishing detection with motive to achieve highly accurate detection. The multi-modal detection framework presented as shown in Fig.2 combines the analysis of three different types of content: Text, Image and URL which can be classified into these components.



**Fig. 2.** Multi-modal approach to phishing detection with Generative AI.

*Input and Pre-processing:* It is the initial stage where incoming emails and messages are segregated into its core components (text, images, URLs). Extraction of relevant data and its preparation for analysis are carried out in this stage. For text, body of the email along with the subject line can be extracted using email parsing libraries or natural language processing. Similarly, for image, email parsing libraries in python and C\*, file handling libraries and Optical Character Recognition (OCR) can be utilised [64]. Likewise, for URL extraction, HTML parsing and URL handling libraries can be incorporated.

*Multi-modal Detection Modules:* Three different and independent modules are combined to form a multi-modal detection for identifying potential phishing indicators by focusing on different aspects of the email content.

- **NLP Module (Text Analysis):** The text extracted from stage 1 is analysed in this module. Text Cleaning, feature extraction and phishing pattern

detection are the processes that will be commenced in this module. Text Cleaning refers to removal of unnecessary elements for simplifying the analysis which can be achieved with tokenization, stop word removal, stemming and lemmatization [65] using libraries such as SpaCy and NLTK. Feature extraction is carried out to convert the text into suitable format which is suitable for processing in ML model. Machine learning classification models such as logistic regression, Support Vector Machine (SVM), Naïve Bayes and Deep learning models can be incorporated for detecting the phishing patterns [66]. This module looks for linguistic patterns, wording, and phrases which are commonly associated with phishing and gives a phishing likelihood score that represents the probability of the text being a part of a phishing attempt.

- **Computer Vision Module (Image Analysis):** This model focuses on extracting image from the emails/messages and analyse them to detect visual cues based on fake logos, brand inconsistencies, as well as embedded text inside images. Key processes involved in this module is Optical Character Recognition (OCR), Logo and Brand Detection, and visual pattern recognition. Text can be extracted from the image using Google Vision API such that, a common phishing tactic, embedded text in images is also processed. CNNs can be implemented for verifying the logo as authenticate or manipulated versions [67]. One of the important aspects of image analysis is detecting suspicious visual elements such as altered icons, low-quality images or fake warning signs [68] which can be done using various ML algorithms. The output from this module is represented in a likelihood score of 0 or 1.
- **URL Analysis Module:** This module evaluates the URLs (domain names, link structures, URL reputation) that are embedded trickily using various AI-techniques in the email to determine its association with phishing attempts. URLs are parsed in this step for suspicious features such as typo squatting, long URLs with excessive special characters. Services such as VirusTotal and Google Safe Browsing API can also be used to check for blacklisted domains [69]. Domain age and history is another indicator for identifying phishing URL as new domains with little history may be suspicious [70]. The outputs of this module will be similar to that of the previous 2 modules representing a likelihood score which are combined to obtain a weighted score.

*Generative AI Integration:* Generative AI can enhance the detection capability of multi-modal phishing detection modal by increasing the adaptiveness. Role of Gen AI in this approach can be utilised for 4 specific tasks: generating synthetic data, threat simulation, adaptive defence, and generation of explainability. For generating phishing samples, GANs can be deployed to augment training datasets allowing the model to learn from wide range of phishing scenarios and improve the adaptiveness [71]. Threat simulation is another aspect which can be initiated using Variational Autoencoders (VAEs) that can mimic the tactics used by cybercriminal and generate a simulation environment. Through continuous testing of detection systems using simulation will ensure that the model is

resilient to emerging threats. Adaptive defence can be leveraged by using Natural Language Generation (NLG) allowing the Gen AI to generate new phishing scenarios through dynamic learning from detected phishing attempts. This can upgrade the system's ability to detect zero-day attacks. It is essential to offer transparency in the decision-making process of AI model to improve user's trust. Using Gen AI to deliver explanation on the outputs, alerts or decision made by the system in human-readable format, will drastically improve the user's experience while using this model.

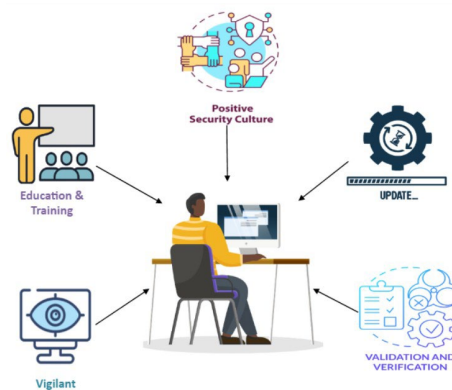
*Fusion Layer (Decision Making):* This layer is the centre element of the multi-modal phishing detection system that plays a pivotal role by acting as a convergence point and integrating different modalities. It creates a holistic view of the email to provide a better understanding of whether the email is likely to be phishing. It obtains outputs from multi-modals in the form of likelihood scores and outputs from synthetic data, threat simulation through generative AI. All these features are combined using feature-level fusion and allows the system to consider the context as a whole. For instance, the differences between the context of an email (text) and the visual elements associated with it might be detected by the system which could indicate phishing. This layer also uses adaptive weighting mechanisms which assigns equal importance to each modality along with cross-modal embeddings. If the phishing pattern is text-based, the score from text analysis might dominate other scores. In such cases, the dynamic approach ensures that the system adapts based on the characteristics of the specific email being analysed instead of relying heavily on a single modality. Various neural networks such as multimodal deep learning, Convolutional Neural Network (CNNs), Recurrent Neural Network (RNNs), attention-based fusion networks, can be utilised in this layer so that holistic view can be achieved with less false positives, reduced errors and biases [72]. A combination of techniques and contextual understanding offered by this layer with the integration of Gen AI, will help in improving accuracy, resilience and adaptability of the detection system.

*Final Phishing Detection:* This is the core decision-making component of the phishing detection system as it classifies whether an email is phishing or legitimate by applying various machine learning models. The combined feature vector from fusion layer is taken as an input for this input. The machine learning models for this layer is chosen based on the complexity of the email. For instance, logistic regression (LR) is a simple interpretable classification model [73] which can provide insights on how each feature of an email contributes to the final decision. This can be used in the model as it is quick to train but is not an ideal one for complex data and cannot be used on its own. A combination of ML such as SVM, CNNs, RNNs or ensemble learning can be used. The output is in the form of phishing likelihood score which is usually in the range of 0 to 1. The threshold can be predefined, for example, if score is above 0.7, the email can be classified as phishing, whereas if it is below 0.3, it can be classified as legitimate.

*Alert and Feedback loop:* User feedback have a great role in improving the model and increase the adaptation. This feature allows the system to evolve according to the real-world performance and user’s feedback to ensure that it is effective against new phishing techniques. Here, user can provide feedback related to false positive and false negatives. This provides the fusion layer information on the scale of attention or weighting to be given for each modality. This helps on continuous progress and adaptation against the new AI-driven tactics used by attackers Some of the key benefits of this feedback loop are dynamic learning, error correction, increased accuracy, and building user trust.

## 7.2 Human-Centered Preventive Measures

Humans are the major aspect of any social engineering as they are deceived and manipulated to disclose confidential information, which is why human-focused strategies play a vital role in mitigating phishing attacks. Some of the user-oriented security measures as shown in Fig.3 have been discussed below.



**Fig. 3.** Human-centered prevention methods.

**Education and Training** Awareness and proper understanding of the potential risk from AI-based cyberattacks are some of the effective preventive strategies which can be achieved through education and training [74]. Regular security awareness training programs can aid in educating individuals regarding the evolving nature of phishing attacks and provide them with the ability to identify potential red flags. Stimulated exercises like generating a controlled synthetic environment and inducing AI-based phishing attacks along with implementing effective procedures to tackle them can make an individual prepared to deal with similar attacks in the future.

**Regular updates** Keeping a regular update in software, systems and latest trends is one of the important security measures even for dynamic phishing attacks. Attackers often tend to exploit the vulnerabilities of software to execute zero-day attacks [75]. But the impact of this issue can be minimized with regular software updates which will ensure security patches against exploitation like any other attackers, cybersecurity personnel and researchers are also elevating defence by upgrading anti-phishing tools and spam filters which when implemented can block phishing attempts. It is also essential to get updated with recent PA patterns to be aware of such attacks and find a solution to defend them.

**Vigilance** Being observant of different activities while being on the internet or any online platform can elevate the early detection of notorious activities such as phishing attacks [76]. If an individual is vigilant, the chances of noticing unusual and suspicious emails and messages are extremely high. Carefully inspecting links and attachments in emails, assessing the domain names in the email addresses are some examples of vigilant actions which can help in identifying phishing. Not only this, but a vigilant individual can identify the legitimacy of highly convincing AI generated PA by making full use of available security tools.

**Promoting a Proactive Security Culture** As AI-based PA are highly dynamic, it is essential to understand the recent tactics used by attackers. For this, information sharing is essential among individuals, employees, employers and companies. Information shared and proper flow of communication among designated level within an organisation regarding any sort of unusual online activities, data breach, and system compromise can elevate the performance of incident response team to handle the situation with appropriate guidelines and minimize the negative impact of the situation. This can also help in preventing similar attacks (AI/non-AI based) in the future. For improved security, it is also important to encourage the usage of security measures against phishing.

**Verification Practices** Verification is an act of establishing a truth with a proof which is essential when providing access to data and information [77]. Some of the common verification practices are Double-checking requests and cross-referencing information [78]. Phishing attacks are generally accomplished by requesting access so verifying the requests can help in identifying deceptive form of tactics used by attackers and protect the data. Another simple yet effective method when it comes to tackling AI-based phishing is to use an alternative trusted channels (phone calls or in-person confrontation) for contacting the sender from unusual request. This measure is not only effective for traditional PA but also for AI-generated attacks. Some of the details such as email address, domain names, logos, language and signatures are great indicators of phishing attempts and hence they need to be verified for protection against phishing attacks.

### 7.3 Ethical Considerations

Controlling the development and use of any kind of technology and products arises the need for proper rules and regulation to flourish it in many ways [79]. AI is fast evolving field of technology that have spread across numerous sectors and have been incorporated heavily in human lives to carry out daily tasks. But with increasing misuse as well such as phishing attacks, regulating AI ethically is essential for protecting against AI-driven attacks. Accountability for errors, misconduct, and regulatory violations are also essential to be defined for minimising the misuse of AI [80]. This can be achieved in many ways through the involvement of stakeholders as well as individuals.

**Regulatory Frameworks and Compliance** Government Laws are an integral aspect in controlling the development and use of AI ethically so that their adverse use is minimised [81]. Increasing number of AI products and users have urged government bodies to regulate it. There are some of the Laws and Acts such as European Union AI Act, UK's National AI Strategy, AI Guidelines and Executive Orders in United States, and many more which plays significant role in regulating and preventing AI-based phishing attacks [82]. Cyber Resilience Act is another example of government at that aims to safeguard consumers and businesses using software and digital products [83]. Similarly, other AI regulations have been introduced in many countries such as China and Japan, further depicting the potential usefulness of AI products. They establish regulatory frameworks which promotes responsible and fair development, deployment and use of AI. Although the current acts and laws addresses most of the areas of AI, there are some limitations. For instance, regulations for misinformation created by Gen AI is inadequate in the EU legislation. Hence, these regulations should be more refined in the future for preventing any sorts of cybercrimes induced with the use of AI.

Data protection regulations such as GDPR (General Data Protection Regulation) and HIPAA (Health Insurance Portability and Accountability Act) provides a framework for protection of user's personal data in an organisation [84]. They provide guidelines on mandatory consent and specific or minimum personal data to be collected and processed. Adopting these regulations strictly will reduce the pool of data that can be exploited during AI-based phishing attacks. They also provide security measures for storing data which can prevent unauthorised access initiated through use of AI. Breach notification is another attribute which have been specified in these regulations that can help in recovery and prevention of further damage in case of data breach. Complying with the principles of data minimisation, purpose limitation and transparency are also some other aspects that these law aims regarding AI system for protecting the privacy rights of an individual. Therefore, every organisation must comply with data protection laws as a preventive measure against AI-based phishing attacks.

**Responsible AI Development and Deployment** When deploying any AI products, it is necessary that a set of principles are followed during not just

deployment but also in the design and development rather than blindly releasing them without overlooking the consequences it might induce [85]. To assure that AI incorporates the ethical implications, responsible AI which is a framework with principal guidelines for design and deployment needs to be implemented [86]. Principle for the responsible development of AI established by the Organisation for Economic Cooperation and Development (OECD) is an example of how organisations can regulate responsible AI [87]. This best practice can aid in minimising AI-based cyberattacks such as phishing attacks.

**Transparency and Explainability** Transparency and explainability, which are linked with each other, are the factors for achieving an understandable AI [88]. Deep understanding of AI fosters swift identification of AI's vulnerabilities and potential misuse as a result of which appropriate preventive measure can be identified. According to various studies [89] [90] [91], explainability results in higher transparency and have now been considered as a key requirement of AI systems. These elements have a positive impact on maintaining ethics around the use of AI and limits its misuse. They define the accountability retained during various instances and provides a clarity of the responsibility and actions to be implied by the designated individual or organisations.

**Cybersecurity Standards and Best Practices** Cybersecurity standards and guidelines provide a set of guidelines which if fully adhered to will secure a system by fulfilling the fundamental security requirements [92]. They help counter AI-based phishing along with other adverse events by providing a multi-layer defence strategy through regular security assessments, data encryption, continuous monitoring and incident response, and multi-factor authentication. DMARC (Domain-based Message Authentication, Reporting, and Conformance), SPF (Sender Policy Framework) and DKIM (DomainKeys Identified Mail) are some of the email filtering and authentication protocols which help in identifying and blocking phishing emails before they reach to user [93]. Adopting these practices reduce the chances of sensitive information to be exploited through phishing.

## 8 Future Directions and Recommendations

AI is advancing daily with the introduction of new techniques in a noticeably short time, and it is evident that attackers will be improving their strategies. The complexity of phishing will grow with ample use of AI tools and techniques in near future. To prevent high impact of phishing in various aspects, advancements must be made for defence using enhanced AI algorithms and Explainable AI (XAI). New solutions against phishing must be designed and developed with adaptive features. To achieve this, investment in AI research and development is recommended along with fostering a security-conscious culture. Empowering humans to adapt to innovative technology must also been done simultaneously to ensure the effectiveness of the defense strategies.

## 9 Conclusion

The use of AI by attackers to conduct phishing attacks is highly threatening. Various AI techniques have already been utilised for phishing attacks by generating convincing emails making them superior than traditional phishing in terms of sophistication and success rates. Considering the threat they pose; it is essential to have a proper defence against these AI-driven phishing attacks as the existing defence tools are not fully effective on these advanced attacks. Some solutions for fighting against these attacks is to develop phishing detector tools by incorporating various AI techniques and looking for holistic approach. A multi-modal approach with the integration of generative AI proposed can help in detecting the advance phishing attacks. Human-drive preventive measures and ethical/regulatory considerations should also be implemented simultaneously for better security against AI-driven phishing attacks.

AI is advancing daily with the introduction of new techniques, and it is evident that attackers will be improving their strategies. To prevent adverse impact of phishing in near future, advancements must be made for defence using enhanced AI algorithms and simultaneously empowering humans to adapt to new technologies must also be done to ensure the effectiveness of the defense strategies.

## References

1. Salahdine, F. & Kaabouch, N. Social engineering attacks: A survey. *Future Internet* **11**, 89 (2019).
2. Goenka, R., Chawla, M. & Tiwari, N. A comprehensive survey of phishing: mediums, intended targets, attack and defence techniques and a novel taxonomy. *International Journal of Information Security* **23**, 819–848 (2024).
3. Blauth, T. F., Gstrein, O. J. & Zwitter, A. Artificial intelligence crime: An overview of malicious use and abuse of ai. *Ieee Access* **10**, 77110–77122 (2022).
4. Kaloudi, N. & Li, J. The ai-based cyber threat landscape: A survey. *ACM Computing Surveys (CSUR)* **53**, 1–34 (2020).
5. Brundage, M. *et al.* The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv preprint arXiv:1802.07228* (2018).
6. Gupta, M., Akiri, C., Aryal, K., Parker, E. & Praharaj, L. From chatgpt to threatgpt: Impact of generative ai in cybersecurity and privacy. *IEEE Access* (2023).
7. Bécue, A., Praça, I. & Gama, J. Artificial intelligence, cyber-threats and industry 4.0: Challenges and opportunities. *Artificial Intelligence Review* **54**, 3849–3886 (2021).
8. Phishing activity trends report, 4th quarter 2023 (2023). URL <https://apwg.org/trendsreports/>.
9. Ooi, K.-B. *et al.* The potential of generative artificial intelligence across disciplines: Perspectives and future directions. *Journal of Computer Information Systems* 1–32 (2023).
10. Metta, S., Chang, I., Parker, J., Roman, M. P. & Ehuan, A. F. Generative ai in cybersecurity (2024).

11. Perlroth, N. *This is how they tell me the world ends: The cyberweapons arms race* (Bloomsbury Publishing USA, 2021).
12. Alkhalil, Z., Hewage, C., Nawaf, L. & Khan, I. Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science* **3** (2021).
13. Malik, J. K. & Choudhury, S. A brief review on cyber crime-growth and evolution. *Pramana Research Journal* **9**, 242 (2019).
14. Guembe, B. *et al.* The emerging threat of ai-driven cyber attacks: A review. *Applied Artificial Intelligence* **36**, 2037254 (2022).
15. Jaber, A. N. & Fritsch, L. *Covid-19 and global increases in cybersecurity attacks: review of possible adverse artificial intelligence attacks*, 434–442 (IEEE, 2021).
16. Yamin, M. M., Ullah, M., Ullah, H. & Katt, B. Weaponized ai for cyber attacks. *Journal of Information Security and Applications* **57**, 102722 (2021).
17. Salloum, S., Gaber, T., Vadera, S. & Shaalan, K. A systematic literature review on phishing email detection using natural language processing techniques. *IEEE Access* **10**, 65703–65727 (2022).
18. Arora, A. & Shantanu. A review on application of gans in cybersecurity domain. *IETE Technical Review* **39**, 433–441 (2022).
19. Grbic, D. V. & Dujlovic, I. *Social engineering with chatgpt*, 1–5 (IEEE, 2023).
20. Yu, J. *et al.* The shadow of fraud: The emerging danger of ai-powered social engineering and its possible cure. *arXiv preprint arXiv:2407.15912* (2024).
21. Wang, J., Herath, T., Chen, R., Vishwanath, A. & Rao, H. R. Research article phishing susceptibility: An investigation into the processing of a targeted spear phishing email. *IEEE transactions on professional communication* **55**, 345–362 (2012).
22. Alkhalil, Z., Hewage, C., Nawaf, L. & Khan, I. Phishing attacks: A recent comprehensive study and a new anatomy. *Frontiers in Computer Science* **3**, 563060 (2021).
23. Emanuela, A. R., Cristina, B. A. & Luminița, S. *Ai and prompt engineering: The new weapons of social engineering attacks*, 1–6 (IEEE, 2024).
24. Das, B. C., Amini, M. H. & Wu, Y. Security and privacy challenges of large language models: A survey. *arXiv preprint arXiv:2402.00888* (2024).
25. Community, O. Openai wrote better spear phishing emails. <https://community.openai.com/t/openai-wrote-better-spear-phishing-emails/7359> (2023). Accessed on [Insert access date here].
26. Otter, D. W., Medina, J. R. & Kalita, J. K. A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems* **32**, 604–624 (2021).
27. Khurana, D., Koli, A., Khatter, K. & Singh, S. Natural language processing: state of the art, current trends and challenges. *Multimedia tools and applications* **82**, 3713–3744 (2023).
28. Young, T., Hazarika, D., Poria, S. & Cambria, E. Recent trends in deep learning based natural language processing [review article]. *IEEE Computational Intelligence Magazine* **13**, 55–75 (2018).
29. Carvalho, I., Oliveira, H. G. & Silva, C. The importance of context for sentiment analysis in dialogues. *IEEE Access* **11**, 86088–86103 (2023).
30. Patton, D. U. *et al.* *Contextual analysis of social media*, 337–342 (ACM, 2020).
31. Mahiuddin, M., Khaliluzzaman, M., Chowdhury, M. S. A. & Arefin, M. N. Fake face generator: Generating fake human faces using gan. *International Journal of Advanced Computer Science and Applications* **13** (2022).
32. Zhang, Y., Gan, Z. & Carin, L. *Generating text via adversarial training*, Vol. 21, 21–32 (Academia. edu, 2016).

33. Liang, G. *et al.* A text gan framework for creative essay recommendation. *Knowledge-Based Systems* **232**, 107501 (2021).
34. Franceschi-Bicchierai, L. Hackers steal activism games and employee data (2023). URL <https://techcrunch.com/2023/02/21/hackers-allegedly-steal-activision-games-and-employee-data/>.
35. Kaushal, R., Ghose, V. & Kumaraguru, P. *Methods for user profiling across social networks*, 1572–1579 (IEEE, 2019).
36. Haleem, A., Javaid, M., Qadri, M. A., Singh, R. P. & Suman, R. Artificial intelligence (ai) applications for marketing: A literature-based study. *International Journal of Intelligent Networks* **3**, 119–132 (2022).
37. Karuna, P., Purohit, H., Jajodia, S., Ganesan, R. & Uzuner, O. Fake document generation for cyber deception by manipulating text comprehensibility. *IEEE Systems Journal* **15**, 835–845 (2021).
38. Apruzzese, G., Conti, M. & Yuan, Y. *Spacephish: The evasion-space of adversarial attacks against phishing website detectors using machine learning*, 171–185 (2022).
39. Song, F., Lei, Y., Chen, S., Fan, L. & Liu, Y. Advanced evasion attacks and mitigations on practical ml-based phishing website classifiers. *International Journal of Intelligent Systems* **36**, 5210–5240 (2021).
40. Emanuela, A. R., Cristina, B. A. & Luminița, S. *Ai and prompt engineering: The new weapons of social engineering attacks*, 1–6 (IEEE, 2024).
41. Firdhous, M. F. M., Elbreiki, W., Abdullahi, I., Sudantha, B. & Budiarto, R. *Wormgpt: A large language model chatbot for criminals*, 1–6 (IEEE, 2023).
42. Meskys, E., Liaudanskas, A., Kalpokiene, J. & Jurcys, P. Regulating deep fakes: legal and ethical considerations. *Journal of Intellectual Property Law Practice* **15**, 24–31 (2020).
43. May, R., Krüger, J. & Leich, T. *Sok: How artificial-intelligence incidents can jeopardize safety and security*, 1–12 (2024).
44. Mirsky, Y. & Lee, W. The creation and detection of deepfakes. *ACM Computing Surveys* **54**, 1–41 (2022).
45. Desolda, G., Ferro, L. S., Marrella, A., Catarci, T. & Costabile, M. F. Human factors in phishing attacks: A systematic literature review. *ACM Computing Surveys* **54**, 1–35 (2022).
46. Heiding, F., Schneier, B., Vishwanath, A., Bernstein, J. & Park, P. S. Devising and detecting phishing emails using large language models. *IEEE Access* **12**, 42131–42146 (2024).
47. ANDRIU, A.-V. Adaptive phishing detection: Harnessing the power of artificial intelligence for enhanced email security. *Romanian Cyber Security Journal* **5**, 3–9 (2023).
48. Guo, Y. A review of machine learning-based zero-day attack detection: Challenges and future directions. *Computer Communications* **198**, 175–185 (2023).
49. Basit, A. *et al.* A comprehensive survey of ai-enabled phishing attacks detection techniques. *Telecommunication Systems* **76**, 139–154 (2021).
50. Kim, W., Jeong, O.-R., Kim, C. & So, J. The dark side of the internet: Attacks, costs and responses. *Information systems* **36**, 675–705 (2011).
51. Alahmed, Y. & Abadla, R. Exploring the potential implications of ai-generated content in social engineering attacks. *International Journal of Computing and Digital Systems* **16**, 1–11 (2024).
52. Dilmaghani, S. *et al.* *Privacy and security of big data in ai systems: A research and standards perspective*, 5737–5743 (IEEE, 2019).
53. Golda, A. *et al.* Privacy and security concerns in generative ai: A comprehensive survey. *IEEE Access* **12**, 48126–48144 (2024).

54. Chiew, K. L., Yong, K. S. C. & Tan, C. L. A survey of phishing attacks: Their types, vectors and technical approaches. *Expert Systems with Applications* **106**, 1–20 (2018).
55. Potti, N., Wendt, J. B., Zhao, Q., Tata, S. & Najork, M. *Hidden in plain sight: Classifying emails using embedded image contents*, 1865–1874 (2018).
56. Alabdan, R. Phishing attacks survey: Types, vectors, and technical approaches. *Future Internet* **12**, 168 (2020).
57. Jain, J. Artificial intelligence in the cyber security environment. *Artificial Intelligence and Data Mining Approaches in Security Frameworks* 101–117 (2021).
58. Fakhouri, H. N. *et al.* *Ai-driven solutions for social engineering attacks: Detection, prevention, and response*, 1–8 (2024).
59. Safi, A. & Singh, S. A systematic literature review on phishing website detection techniques. *Journal of King Saud University-Computer and Information Sciences* **35**, 590–611 (2023).
60. Sahingoz, O. K., Buber, E. & Kugu, E. Dephides: Deep learning based phishing detection system. *IEEE Access* (2024).
61. Alsubaei, F. S., Almazroi, A. A. & Ayub, N. Enhancing phishing detection: A novel hybrid deep learning framework for cybercrime forensics. *IEEE Access* (2024).
62. van Geest, R., Cascavilla, G., Hulstijn, J. & Zannone, N. The applicability of a hybrid framework for automated phishing detection. *Computers & Security* **139**, 103736 (2024).
63. Çolhak, F., Ecevit, M. İ., Uçar, B. E., Creutzburg, R. & Dağ, H. Phishing website detection through multi-model analysis of html content. *arXiv preprint arXiv:2401.04820* (2024).
64. Patel, C., Patel, A. & Patel, D. Optical character recognition by open source ocr tool tesseract: A case study. *International journal of computer applications* **55**, 50–56 (2012).
65. Chai, C. P. Comparison of text preprocessing methods. *Natural Language Engineering* **29**, 509–553 (2023).
66. Karim, A., Shahroz, M., Mustofa, K., Belhaouari, S. B. & Joga, S. R. K. Phishing detection system through hybrid machine learning based on url. *IEEE Access* **11**, 36805–36822 (2023).
67. Battini, V. S., Kishan, S. R. & Valeti, V. D. *Fake logo detection using image processing*, 1–7 (IEEE, 2024).
68. Rocha, A., Scheirer, W., Boulton, T. & Goldenstein, S. Vision of the unseen. *ACM Computing Surveys* **43**, 1–42 (2011).
69. Bell, S. & Komisarczuk, P. *An analysis of phishing blacklists: Google safe browsing, openphish, and phishtank*, 1–11 (ACM, 2020).
70. Gowtham, R. & Krishnamurthi, I. A comprehensive and efficacious architecture for detecting phishing webpages. *Computers Security* **40**, 23–37 (2014).
71. Shirazi, H., Muramudalige, S. R., Ray, I., Jayasumana, A. P. & Wang, H. Adversarial autoencoder data synthesis for enhancing machine learning-based phishing detection algorithms. *IEEE Transactions on Services Computing* **16**, 2411–2422 (2023).
72. Temburne, J. V. & Diwan, T. Sentiment analysis in textual, visual and multi-modal inputs using recurrent neural networks. *Multimedia Tools and Applications* **80**, 6871–6910 (2021).
73. Lv, C. & Chen, D.-R. *Interpretable functional logistic regression*, 1–5 (ACM, 2018).
74. Ansari, M. F., Sharma, P. K. & Dash, B. Prevention of phishing attacks using ai-based cybersecurity awareness training. *Prevention* **3**, 61–72 (2022).

75. Bilge, L. & Dumitras, T. *Before we knew it: an empirical study of zero-day attacks in the real world*, 833–844 (2012).
76. Felt, A. P. & Wagner, D. Phishing on mobile devices (2011).
77. Al-Khawajah, N., Al-Billeh, T. & Manasra, M. Digital forensic challenges in jordanian cybercrime law. *Pakistan Journal of Criminology* **15** (2023).
78. Van Leuven, S., Kruikemeier, S., Lecheler, S. & Hermans, L. Online and news-worthy: Have online sources changed journalism? *Digital Journalism* **6**, 798–806 (2018).
79. de Almeida, P. G. R., dos Santos, C. D. & Farias, J. S. Artificial intelligence regulation: a framework for governance. *Ethics and Information Technology* **23**, 505–525 (2021).
80. Uzougbo, N. S., Ikegwu, C. G. & Adewusi, A. O. Legal accountability and ethical considerations of ai in financial services. *GSC Advanced Research and Reviews* **19**, 130–142 (2024).
81. Arcila, B. B. Ai liability in europe: How does it complement risk regulation and deal with the problem of human oversight? *Computer Law & Security Review* **54**, 106012 (2024).
82. Liebig, L., Güttel, L., Jobin, A. & Katzenbach, C. Subnational ai policy: shaping ai in a multi-level governance system. *AI SOCIETY* **39**, 1477–1490 (2024).
83. Novelli, C., Casolari, F., Hacker, P., Spedicato, G. & Floridi, L. Generative ai in eu law: liability, privacy, intellectual property, and cybersecurity. *arXiv preprint arXiv:2401.07348* (2024).
84. Kirwan, M. *et al.* What gdpr and the health research regulations (hrrs) mean for ireland: “explicit consent” – a legal analysis. *Irish Journal of Medical Science (1971-)* **190**, 515–521 (2021).
85. Kenthapadi, K., Lakkaraju, H. & Rajani, N. *Generative ai meets responsible ai: Practical challenges and opportunities*, 5805–5806 (ACM, 2023).
86. Peters, D., Vold, K., Robinson, D. & Calvo, R. A. Responsible ai – two frameworks for ethical design practice. *IEEE Transactions on Technology and Society* **1**, 34–47 (2020).
87. Akpuokwe, C. U., Adeniyi, A. O. & Bakare, S. S. Legal challenges of artificial intelligence and robotics: a comprehensive review. *Computer Science & IT Research Journal* **5**, 544–561 (2024).
88. Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkänen, K. & Kujala, S. Transparency and explainability of ai systems: From ethical guidelines to requirements. *Information and Software Technology* **159**, 107197 (2023).
89. Chazette, L., Karras, O. & Schneider, K. *Do end-users want explanations? analyzing the role of explainability as an emerging aspect of non-functional requirements*, 223–233 (IEEE, 2019).
90. Chazette, L. & Schneider, K. Explainability as a non-functional requirement: challenges and recommendations. *Requirements Engineering* **25**, 493–514 (2020).
91. Horkoff, J. *Non-functional requirements for machine learning: Challenges and new directions*, 386–391 (IEEE, 2019).
92. Srinivas, J., Das, A. K. & Kumar, N. Government regulations in cyber security: Framework, standards and recommendations. *Future generation computer systems* **92**, 178–188 (2019).
93. Hu, H., Peng, P. & Wang, G. *Towards understanding the adoption of anti-spoofing protocols in email systems*, 94–101 (IEEE, 2018).