AI

MDPI

Systematic Review

Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Review of Human Factors

Raja Jabir *, John Le and Chau Nguyen *

Institute of Cybersecurity and Cryptology, School of Computing and Information Technology, University of Wollongong, Wollongong, NSW 2522, Australia; johnle@uow.edu.au

* Correspondence: rj991@uowmail.edu.au (R.J.); chaun@uow.edu.au (C.N.)

Abstract

Despite the focus on improving cybersecurity awareness, the number of cyberattacks has increased significantly, leading to huge financial losses, with their risks spreading throughout the world. This is due to the techniques deployed in cyberattacks that mainly aim at exploiting humans, the weakest link in any defence system. The existing literature on human factors in phishing attacks is limited and does not live up to the witnessed advances in phishing attacks, which have become exponentially more dangerous with the introduction of generative artificial intelligence (GenAI). This paper studies the implications of AI advancement, specifically the exploitation of GenAI and human factors in phishing attacks. We conduct a systematic literature review to study different human factors exploited in phishing attacks, potential solutions and preventive measures, and the complexity introduced by GenAI-driven phishing attacks. This paper aims to address the gap in the research by providing a deeper understanding of the evolving landscape of phishing attacks with the application of GenAI and associated human implications, thereby contributing to the field of knowledge to defend against phishing attacks by creating secure digital interactions.

Keywords: cybersecurity; phishing; phishing countermeasures; human factors; artificial Intelligence (AI); generative artificial intelligence (GenAI); systematic literature review

check for updates

Academic Editor: Giovanni Diraco

Received: 15 June 2025
Revised: 21 July 2025
Accepted: 26 July 2025
Published: 31 July 2025

Citation: Jabir, R.; Le, J.; Nguyen, C. Phishing Attacks in the Age of Generative Artificial Intelligence: A Systematic Review of Human Factors. *AI* **2025**, *6*, 174. https://doi.org/10.3390/ai6080174

1. Introduction

In recent years, the term ‘cybersecurity’ has been increasingly used across various disciplines, moving away from previously used terms such as ‘computer security’ and ‘information security’ and hence increasing cybersecurity understanding [1]. However, it is critical to deepen our understanding of the associated risks and financial losses experienced on a global scale. This acknowledges that anyone using an electronic platform is prone to being a victim of cyberattacks [2]. Cyberattacks are advanced through social engineering, the deployment of harmful software, the sharing of confidential information, and the lack of efficient defence systems [3]. It is important to highlight that cybercrimes cause significant global risks, as attacks are not limited by geographic location and have the potential to scale rapidly, impacting critical infrastructures and leading to serious financial losses. According to the Australian Government’s Australian Signals Directorate (ASD) Cyber Threat Report 2023–2024, there were more than 87,400 cybercrime reports over the same financial year, with an estimate that one incident was reported every six minutes [4]. Furthermore, the Australian Government’s scam watcher [5] reported that the financial losses of phishing attacks in 2024 amounted to approximately AUD 20.5 million and

6.1 million in January 2025. Hence, more research and unconventional measures are required to protect our cybersecurity.

In the fight against cyberattacks, it is important to consider the role of humans. Cybersecurity risks are amplified and spread with the exploitation of human factors [3]; therefore, the main challenge of cybersecurity is to find the right balance between people's freedom and their security. We should also consider the risk of human error, which can lead to security breaches and cyberattacks [6]. This demonstrates the importance of paying attention to the aspects of human factors along with technological aspects to strengthen cybersecurity [7] and prevent cybersecurity breaches [8,9]. To prevent human errors, it is important to understand human behaviours toward cybersecurity [10].

Cybersecurity behaviours can be studied using different learning and behavioural theories [11], such as the Theory of Planned Behaviour (TPB), Protection Motivation Theory (PMT), Social Cognitive Theory (SCT), General Deterrence Theory (GDT), and the Technology Acceptance Model (TAM) [12]. These theories have allowed researchers to begin to understand the cybersecurity behaviours of individuals [13,14], providing a comprehensive assessment of human behaviour towards different systems and technologies. For example, Al-Qaysi et al. [15] used TAM to understand users' acceptance of new applications, application usage, attitudes, and intentions to measure technology adoption rates using human behaviour assessment. Efforts made to understand human behaviour are a step in the right direction, and more research is required to gain a deeper understanding and insight into the different factors intertwined in the hindrance of negative cybersecurity behaviour.

The rapid advancement of artificial intelligence (AI) has significantly transformed several aspects of our lives, changing the traditional way of work through its ability to automate tasks and generate human-like content [16,17] at the expense of our digital privacy and security [18]. This dependence on technology to connect with each other and perform the simplest day-to-day tasks has increased our vulnerability to cyberattacks. Furthermore, advances in generative artificial intelligence (GenAI) technology have led to the possibility of mimicking human activities, behaviours, and communication styles. AI technologies can be implemented to detect phishing attacks; unfortunately, these technologies can also be used maliciously by cybercriminals to advance cyberattacks. The World Economic Forum, in its Global Cybersecurity Outlook 2024, highlighted the risks of emerging technologies' capabilities of advancing social engineering attacks, such as phishing attacks [19]. These have increased the threat of digital deception through the deployment of GenAI technologies in phishing attacks. Therefore, there is an urgent need to assess the impact and challenges introduced by GenAI technology [20] and identify means to protect our digital interactions.

This paper aims to analyse various phishing attacks and their advancement using emerging AI technologies, understand how cybercriminals exploit human factors in phishing attacks, and examine how GenAI can be harnessed to conduct sophisticated phishing attacks. The research findings provide valuable information that supports researchers and cybersecurity professionals in understanding the advancement of AI-driven phishing attacks. This paper contributes to the research gap in phishing attacks in the following ways:

- Highlighting the rapid enhancement and wide accessibility of GenAI and the associated risk of misuse in advancing phishing attacks.
- Highlighting opportunities for using GenAI as a solution for phishing attacks.
- Providing a holistic approach to all human factors that have been exploited in phishing attacks and how they contribute to negative cybersecurity behaviours.
- Highlighting research directions to support researchers and practitioners on the topic.

The remainder of this paper consists of seven sections. Section 2 presents the research concepts and explores the current state of research in the fields of cybersecurity behaviour, human factors, and phishing attacks. Section 3 outlines the research method, research inclusion criteria, and research questions. Section 4 answers the identified research questions and scope by addressing the human factors exploited in phishing attacks, the proposed solutions, and the role of GenAI in advancing phishing attacks. Section 5 discusses the findings. Section 6 identifies future research directions. Section 7 concludes the research, providing an overview of the research, key findings, and future research directions.

2. Background and Related Work

This section provides the necessary background on the main concepts covered in this systematic literature review. This section establishes the foundation for the paper, in which we explore the state of the art of different types of phishing attacks and research on human factors.

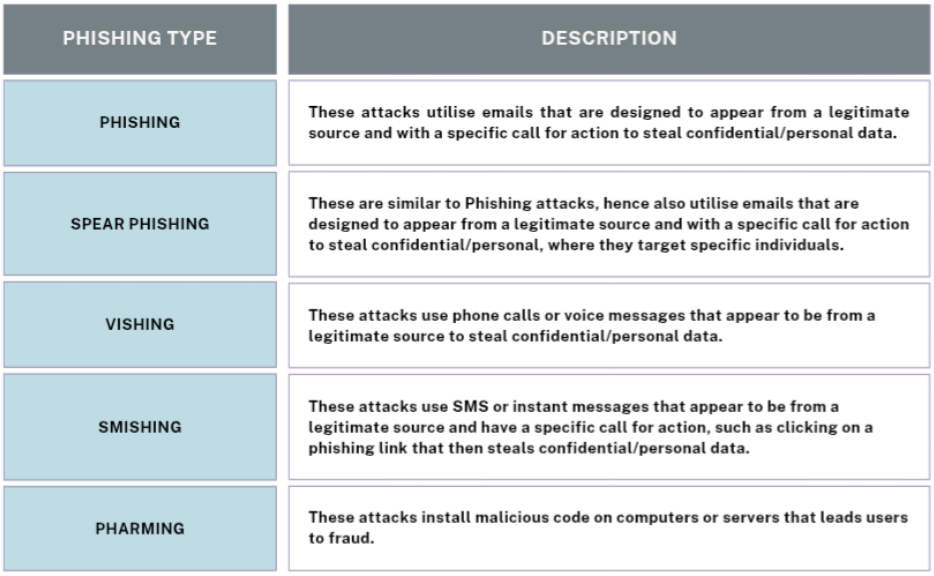
2.1. Phishing Attacks Description

Phishing attacks are considered the most common cybersecurity attacks, and they may cause significant damage to many organisations and individuals. Lastdrager defined phishing as "a scalable act of deception whereby impersonation is used to obtain information from a target" [21] (p. 8). Figure 1 provides a summary of the main types of phishing attacks, noting that all types have the objective of convincing people to share personal or confidential information (for example, bank details, account IDs, and passwords) [22]. The motivation behind the phishing attacks varies depending on the attacker's objective, and phishing is increasingly used as a front to advance different attacks such as identity theft, ransomware, espionage, and blackmail [23].

It is important to note that losses due to phishing attacks are not only financial. Phishing attacks exploit people's vulnerabilities, causing phishing victims to fall into depression or even die by suicide [24,25]. Most fall victim to phishing attacks due to carelessness or lack of knowledge; therefore, it is important to be educated about the traps and tricks used by attackers. Phishing emails are identified by spam filter classifications using the following features [26]:

- Email body: The email body content is scanned for attributes such as HTML, images or shapes, and specific sentences and phrases.
- Email Subject: The email's subject line is scanned for specific common terms, such as 'verify' or 'debit'.
- URLs: Emails are classified as suspicious if an IP address is used instead of the sender domain. Other attributes include, but are not limited to, the presence of external links and links with the '@' sign.
- Sender email: The sender's address is checked and compared with the reply-to reaction.
- Scripts or code: This refers to emails that include JavaScript, click-on activities, or any code present in the email's body or subject.

Phishing attacks have different forms, though emails are the most common communication medium used in phishing attacks because of their ease of use and low cost. Other mediums used include deceitful text messages or phone calls that impersonate legitimate companies or individuals. Therefore, the type of phishing attack, targeted victims, and medium used to advance these attacks should be considered to understand the associated risks and implement the appropriate protection mechanisms.



PHISHING TYPE	DESCRIPTION
PHISHING	These attacks utilise emails that are designed to appear from a legitimate source and with a specific call for action to steal confidential/personal data.
SPEAR PHISHING	These are similar to Phishing attacks, hence also utilise emails that are designed to appear from a legitimate source and with a specific call for action to steal confidential/personal, where they target specific individuals.
VISHING	These attacks use phone calls or voice messages that appear to be from a legitimate source to steal confidential/personal data.
SMISHING	These attacks use SMS or instant messages that appear to be from a legitimate source and have a specific call for action, such as clicking on a phishing link that then steals confidential/personal data.
PHARMING	These attacks install malicious code on computers or servers that leads users to fraud.

Figure 1. Main phishing types.

2.2. Generative AI in Phishing Attacks

AI has transformed several industries by improving performance and customer experience. GenAI, a branch of AI, represents the latest advance in this field. GenAI models are capable of producing high-quality novel content that mimics human creations by leveraging enhanced capabilities in machine learning (ML) models, such as deep learning (DL), which facilitate learning from complex training datasets. The landscape of GenAI is categorised into the following model architectures:

- Generative Adversarial Networks (GANs): GANs consist of two neural networks: the discriminator and the generator. The generator is designed to create sample data, and the discriminator verifies authenticity. The aim is to produce high-quality, realistic data by continuous refinement [27].
- Variational Autoencoders (VAEs): Input data are first encoded into a latent space and then decoded to rebuild the original/initial data. This enables the generation of new data that resemble the data in the original dataset [28].
- Transformer Models: These models advance natural language processing, leveraging long-range data dependencies and generating coherent and contextual content [29].

AI is now easily accessible to anyone, and many vendors compete to introduce technology to gain a competitive edge with an easy-to-use, friendly interface. However, this comes at the expense of digital privacy and security [18] by increasing vulnerability to cyberattacks, including phishing attacks. Attackers increasingly use GenAI tools to produce content that advances the credibility of phishing attacks. This is achieved by producing content such as text, voice, and video [30], replicating trusted entities' behaviour and communication style with the aim of deceiving targeted victims. The threat of digital deception through the deployment of GenAI technologies in phishing attacks is increasing significantly. Therefore, there is an urgent need to assess the impact and challenges introduced by GenAI technology and identify means to protect our digital interactions.

2.3. Human Factors in Phishing Attacks

Any system is only as secure as its weakest link, which is the individual using or operating it [31]. Therefore, to ensure that organisations are protected from cyberattacks, employees should follow cybersecurity policies and recommendations. Over the years, many theories have been developed to understand human behaviour, such as PMT and TPB, which study the factors that influence cybersecurity practices [12].

Several studies have evaluated multiple features of cybersecurity behaviours [10]; however, there is still a need for further analysis of existing theories and models to provide deeper insight into the factors that influence individuals' cybersecurity behaviours. Furthermore, understanding the relationship between all the elements that contribute to human behaviour, such as psychological, environmental, and social factors, will aid efforts to improve cybersecurity behaviours [32]. Chowdhury et al. [33] provide a theoretical framework that demonstrates the effect of time restrictions on cybersecurity behaviour based on psychological factors and context. Other reviews have focused on the educational aspect, such as a review by Hwang et al. [34], which explored the use of multimedia in cybersecurity education and awareness among individuals without professional experience by developing an educational cybersecurity game that studied the interactions of participants using the features of the game, the theory of learning, and the characteristics of users. Figure 2 summarises different elements of cybersecurity behavioural studies.

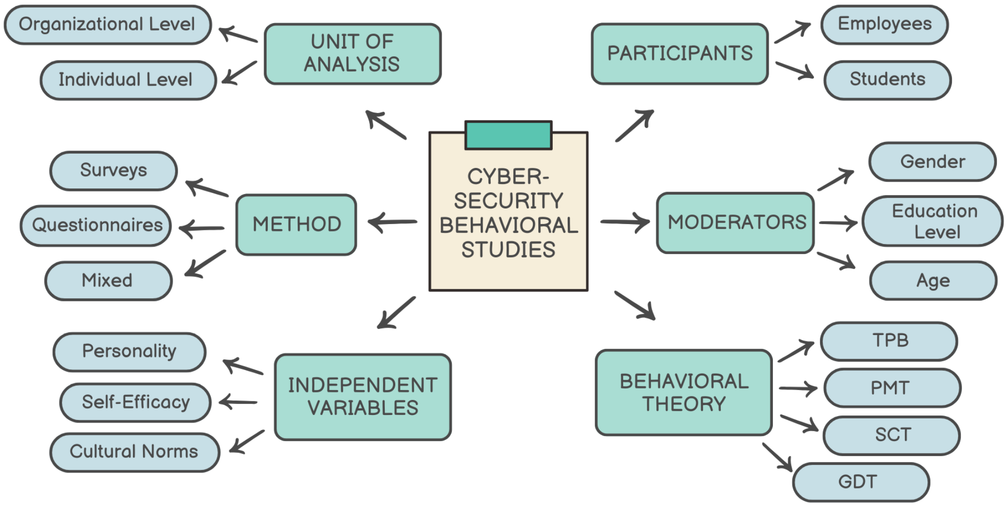


Figure 2. Cybersecurity behavioural studies.

Kalhoro et al. [35] studied the personal, technical, social, and environmental dimensions that contribute to the cybersecurity behaviours of software engineers and found that positive cybersecurity behaviours among software engineers are determined by their personality traits (specifically conscientiousness and openness personality types), level of technical knowledge, and experience. Related work focused on the importance of user awareness in minimising the effect of human errors on breach prevention [36] and found that there is a significant need for user-centred strategies that address human factors to effectively reduce security risks. Other studies have focused on understanding attacks, such as hacking, service interruption, unauthorised access, spam, and malware attacks [37].

Another area of research is human cognitive factors aimed at understanding the characteristics and thinking patterns that impact behaviour. Cognitive factors include attention, memory, and perception [38], which translate into how humans survive and make decisions. From this perspective, some studies have focused on the cognitive security dimension, which aims to understand how humans solve problems and make decisions considering the associated consequences [39]. The term cognitive security refers to the practices and efforts used to defend against cyberattacks; in addition, it can refer to the application of technologies that mimic human cognition against threats [38], such as deploying knowledge and patterns in technologies like AI for the detection and prevention of cyberattacks.

As part of our study, we identified a research gap, namely the limited studies on the advancement of phishing attacks and the exploitation of human factors using GenAI. Table 1 demonstrates the focus areas of some recent review articles on phishing attacks and provides a comparison that helps the reader to understand the specific contribution of this research paper.

As illustrated in Table 1, Desolda et al. [31] explored the role of human factors in phishing attacks and highlighted the lack of knowledge, awareness, and resources, as well as the role of norms in the success of these attacks. This paper also identified the use of games in security training as a countermeasure and emphasised the importance of a common framework and understanding the exploitation of human factors. Similarly, Arevalo et al. [38] studied human factors exploited in phishing attacks and identified several techniques to detect these attacks, such as educational games, leveraging user behaviour knowledge, and deploying cognitive security tools. This paper also introduced cognitive psychology as a means to study users' perceptions of phishing messages and understand how factors such as pressure, stress, and time filter into users' susceptibility to attacks.

On the other hand, Naqvi et al. [40] focused on the classification of phishing attack mitigation strategies by grouping them into anti-phishing systems, models or frameworks, and human-focused strategies. The study explored different tools and capabilities, such as ML. Similarly, Ayeni et al. (2024) [41] reviewed phishing attack detection techniques by classifying these attacks based on the method used and their communication channels. The detection techniques proposed were visual similarity, ML, and DL. Kyaw et al. [42] and Thakur et al. [43] studied DL as a mechanism to detect phishing emails, and both studies found significant advancements in phishing email detection, often outperforming traditional ML models. Thakur et al. [43] explored and assessed the accuracy of different techniques, highlighting the limitation of data set availability and the lack of focus on privacy preservation. Kyaw et al.'s [42] countermeasure focused on identifying email anomalies and patterns; this study also highlighted that data sets for training are limited, which leads to poor generalisations and model overfitting.

Schmitt and Flechais [44] studied the role of GenAI in social engineering, including phishing attacks, and how it increases the risk of attack. The study mentions the limitations of traditional training programmes when faced with AI phishing attacks.

Table 1. Comparison of recent phishing review papers.

Papers	Phishing	Human Factors	Generative AI	Phishing Countermeasures
Desolda et al. (2021) [31]	Y	Y	N	Y
Schmitt and Flechais (2023) [44]	Y	•	Y	Y
Arevalo et al. (2023) [38]	Y	Y	N	•
Naqvi et al. (2023) [40]	Y	N	N	Y
Thakur et al. (2023) [43]	Y	N	•	Y
Kyaw et al. (2024) [42]	Y	N	•	Y
Ayeni el al. (2024) [41]	Y	•	N	Y
Our work	Y	Y	Y	Y

Note: "Y" = Paper's Focus, "•" = Mentioned but not detailed, "N" Not Covered.

3. Methodology

The objective of a systematic review is to understand a topic of interest by studying relevant sources that contribute to the field of research [45]. To thoroughly study GenAI and human factors in relation to the transformation of phishing attacks, we performed a systematic review of the relevant literature on the topic using a scientific and reproducible method, namely the PRISMA guidelines [46] and Kitchenham methodology [47]. Kitchenham explained that review papers consist of three main phases: planning, conducting, and reporting. This section provides details on the steps followed in conducting the literature review, incorporating a structured approach for studying the research on phishing attacks. In this section, both the "planning" and "conducting" phases are explained; the "reporting" phase is detailed in the Results section.

The systematic review protocol was registered with the Open Science Framework (OSF) to ensure transparency. The registration provided details such as the search design, inclusion, and exclusion criteria and is available at https://osf.io/jczu3 (accessed on 15 July 2025).

3.1. Systematic Literature Review Planning Phase

The planning phase of the review paper consists of the following main activities: (a) preparing the right research questions, (b) selecting relevant search strings, (c) choosing the search databases, and (d) defining the inclusion and exclusion criteria.

3.1.1. Preparing the Right Research Questions

We studied the current literature to understand the topic of research and identify the research gap. The research questions were then carefully developed to address this gap. This study explores how humans are exploited in phishing attacks and the complexity added with the introduction of GenAI. Consequently, the following questions were formulated to address the research gap:

- Research Question 1 (RQ1): What factors make humans susceptible to phishing attacks?
- Research Question 2 (RQ2): How has GenAI increased the risks and sophistication of phishing attacks?
- Research Question 3 (RQ3): What are the most effective human-centred and technological solutions to mitigate phishing attacks?

The first research question (RQ1) supports an understanding of the state of the existing research in relation to how human factors in phishing attacks are exploited to make individuals more vulnerable and susceptible to phishing attacks. The second question (RQ2) aims to understand how the sophistication of phishing attacks has increased with the introduction of GenAI. The third question (RQ3) investigates existing solutions focusing on technology or human aspects that have been implemented to reduce or prevent phishing attacks.

3.1.2. Selecting Relevant Search Strings

To thoroughly search for publications covering the scope of our research, relevant keywords were identified based on the research questions in Section 3.1.1. The search strings were derived from our knowledge in the areas of cybersecurity and social engineering attacks, as well as by studying the most cited papers on the topic of phishing attacks. The main term used to retrieve the majority of the publications was "Phishing", and other terms such as "social engineering" were also selected to complement the search, as they are commonly used in relation to phishing attacks.

The variation in the spelling of words in British English and American English was also considered to avoid overlooking relevant research papers. An example of such a word includes, but is not limited to, "behaviour" versus "behavior". Strings such as "user behavio(u)r" were also used. In addition, to further enhance our search focus on the context of human factors, the keywords "human-computer interaction" and "human factors" were utilised. Considering the search criteria mentioned, an example of the final search strings used is: "human factors and phishing", "human factors and social engineering", "user behaviour and phishing", "artificial intelligence and phishing", "generative artificial intelligence and phishing", and "GenAI and phishing". We also considered unique syntax requirements or recommendations for the different digital libraries and research databases when performing the search.

3.1.3. Choosing the Search Databases

Selecting relevant research and publications from reputable data sources was the basis for our detailed literature review process. The selected papers were from previous conference proceedings, journals, and scientific digital libraries. The search engines used included Google Scholar, Scopus, and ScienceDirect, and the libraries searched were ACM, IEEE, and Springer. This provided us with multiple sources for our research scope.

3.1.4. Defining the Inclusion and Exclusion Criteria

The formulation of inclusion and exclusion criteria aims to eliminate biased publication selection. This process ensured that the selected publications were aligned with the objectives of our study and supported finding answers to identified research questions. References from other authors helped define these criteria [48,49]. The inclusion criteria were defined to gather the most relevant papers published between 2000 and 2025, specifically targeting phishing attacks' advancement using GenAI technology and human factor exploitation. The exclusion criteria were defined to eliminate all irrelevant papers. Hence, each publication considered for our research satisfied the inclusion and exclusion criteria, as presented in Table 2.

Table 2. Paper inclusion/exclusion criteria.

Inclusion Criteria	Exclusion Criteria
Publication is on phishing	Publication is not related to scope
Publication is on AI in phishing	Publication is not fully accessible
Publication is on human factors in phishing	Publication contains partial results
Publication reports complete results	
Publication language is English	

3.2. Conducting the Literature Review

The execution of the literature review process—the "conducting phase", as presented by Kitchenham [47]—was divided into two main parts: "literature review execution" and "data synthesis" (as explained below). This was accomplished using the PRISMA [46] guidelines and the CASP (Critical Appraisal Skills Programme) quality assessment tool during the article selection process to assess the relevance of the selected papers. Then, 116 papers were examined for this review, as illustrated in Figure 3.

- Literature Review Execution: At this stage, the search query was designed to be inclusive and was executed in several scientific digital libraries and databases, as mentioned in Section 3.1.3. The search objective was to study the relevant publications on phishing attacks, human factors, and AI from the years 2000 to 2025, thus including the latest publications on our research topic.
- Data Synthesis: At this stage, all duplicates were removed, and the articles were screened to ensure that they met the inclusion and exclusion criteria. The remain-

ing eligible papers were then included in this study after reading the full text and determining that they were relevant to the topic of the research.

As part of our study, further analysis was conducted on all eligible papers, gathering the publication title; a description of the paper's proposed phishing solution and technology; the phishing type and medium addressed, such as email, voice, message, etc.; the human factors addressed; and AI-driven phishing attacks. This provided us with a better understanding of the topic of research from different perspectives and the ability to answer the research questions. Section 4 presents the analysis and findings of the study.

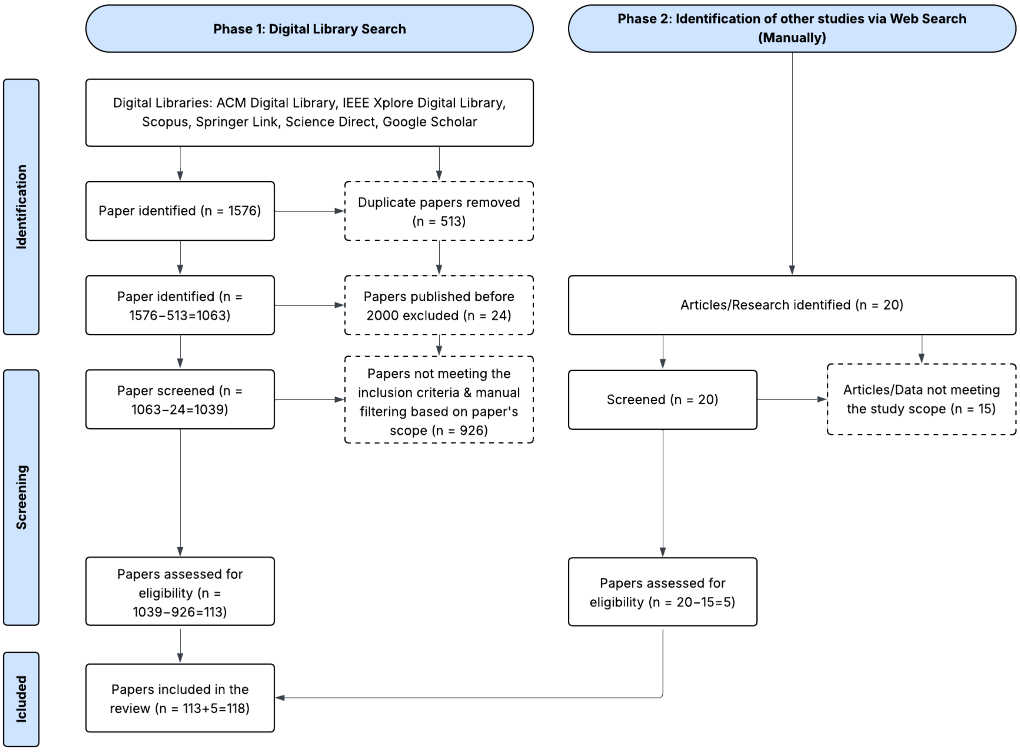

Phase 1: Digital Library Search
Phase 2: Identification of other studies via Web Search (Manually)
Identification
Digital Libraries: ACM Digital Library, IEEE Xplore Digital Library, Scopus, Springer Link, Science Direct, Google Scholar
Paper identified (n = 1576)
Duplicate papers removed (n = 513)
Paper identified (n = 1576−513=1063)
Papers published before 2000 excluded (n = 24)
Articles/Research identified (n = 20)
Screening
Paper screened (n = 1063−24=1039)
Papers not meeting the inclusion criteria & manual filtering based on paper's scope (n = 926)
Screened (n = 20)
Articles/Data not meeting the study scope (n = 15)
Papers assessed for eligibility (n = 1039−926=113)
Papers assessed for eligibility (n = 20−15=5)
Icluded
Papers included in the review (n = 113+5=118)

Figure 3. Selection procedure and results depicted in the PRISMA flow diagram.

4. Findings

This section presents our research results by examining the state of the literature on human factors in phishing attacks. The results are presented in logical order by answering the research questions mentioned in Section 3.1.

4.1. What Factors Make Humans Susceptible to Phishing Attacks? (RQ1)

After analysing existing research and publications on phishing attacks, we found limited explicit references to human factors in such attacks. Therefore, the included research papers were studied to identify the main areas related to human exploitation. The retrieved publications were studied, and the factors that can lead to human vulnerabilities were categorised into the following main reasons, which are further explored in this section.

4.1.1. Insufficient Training

On a daily basis, individuals receive several emails, and amid ongoing responsibilities and tasks, they may not pay attention to minor changes in email structure, the sender's email, or the linked website. Therefore, it is necessary to implement the right tools and proper education to inform them about suspicious or potential threats. It is also important to note that cybersecurity experts may, at times, be potential victims of cyberattacks [50]. Therefore, there is a persistent need to introduce up-to-date cybersecurity training that addresses the latest threats and attacks to all individuals and organisations. This would increase awareness of the latest cybersecurity attacks, including phishing attacks, and the best practices for preventing security attacks.

GenAI enables attackers to craft highly convincing phishing emails that eliminate commonly known signs, such as poor grammar or incorrect phrasing. This provides attackers with the capability of mimicking legitimate messages that are difficult to detect. In addition, deepfake audio and video that impersonate authoritative or well-known entities can be used to deceive targeted victims. Hence, it is critical to have advanced cybersecurity training that addresses GenAI phishing email detection to prevent attack escalation and financial losses.

4.1.2. Bias and Neglect

To further understand the human factors exploited in phishing attacks, we used the Dual Process Theory (DPT) psychological framework. This framework explains human decision making by dividing it into two systems of thinking: (a) 'System 1 Thinking', which operates without conscious reasoning and is therefore fast, intuitive, dependent on past experiences, heuristic, and automatic, and (b) 'System 2 Thinking', which operates on critical evaluation and is therefore slow, intentional, logical, and analytical [51].

Cognitive factors, which include cybersecurity and digital literacy, critical thinking, and impulsiveness, are important when studying human factors exploited in phishing attacks. Individuals who lack digital literacy or critical thinking are more susceptible to phishing attacks, as they are less likely to notice or detect suspicious sender details or links [52]. Overconfidence in cybersecurity knowledge and not having fallen victim to cybersecurity attacks in the past lead people to lower their guard during online transactions and email handling, leading them to fall victim to phishing attacks [31].

In the context of organisations, employees are influenced by general workplace guidelines and practices [32,53,54]. In cases of cyberattacks, if the organisation overestimates security practices and controls, it could leave the organisation vulnerable to cyberattacks, including phishing attacks. In addition, individuals tend to fall into 'authority bias', in which they trust messages that appear to come from their managers, executives, or known brands, preventing them from noticing signs of fraud [52].

Psychological traits such as trust, security attitudes, emotional reactivity, and risk perception play a critical role in understanding susceptibility to phishing attacks. Oner et al. [52] studied susceptibility to phishing attacks and found that individuals who are trusting are more likely to become victims of attackers impersonating legitimate entities. The study mentioned that emotional manipulation is a common tactic used by attackers when crafting messages that introduce a sense of urgency or trigger fear or excitement.

In phishing, attackers design their attacks to invoke System 1 thinking by creating a sense of urgency or triggering emotional reactions or fear. The aim is for potential victims to bypass rational thinking or careful analysis, leading them to respond to a phishing email or click on a suspicious link. GenAI has the potential to increase this risk by mimicking legitimate communications and personalised messaging, bypassing suspicion and triggering System 1 thinking.

4.1.3. External Influence

The impact on cybersecurity and susceptibility to phishing attacks involves not only internal factors but also external influences, such as our environment and workplace culture, which impact individual practices and development over time [55]. Security practices were measured by Alsharnouby et al. [56] using a questionnaire that included statements such as "Securing workstations (screen lock or logout) before leaving the work area to prevent unauthorised access", "Passwords should not be shared with anyone", etc. Sasse et al. [57] mentioned that good security practices, such as not sharing their passwords and locking their screen before leaving the office, are perceived as signs of mistrust in the social and workplace contexts. The impact of social and cultural influence, as discussed by Pham et al. [32], affects the behaviours of individuals.

4.2. How Has GenAI Increased the Risks and Sophistication of Phishing Attacks? (RQ2)

Cyberattacks are consistently trying to bypass all protective measures. This section explores how emerging technologies such as GenAI have increased the risk of cyberattack transformation. The offensive use of GenAI has contributed significantly to the sophistication of phishing attacks, which has added to the difficulty of deploying an effective defence plan. A deeper understanding of the various attack strategies effectively supports efforts to enhance our defence systems in protection against cyber attacks.

4.2.1. Defence System Evasion

GenAI can be used by cybercriminals to carry out attacks that bypass existing detection mechanisms [58]. This is achieved by utilising GenAI technology to create adversarial examples aimed at leading detection systems to false flags and predictions. Generative Adversarial Networks (GANs) can be used to generate malicious traffic to avoid detection by intrusion detection systems (IDS), spreading through the network as part of a phishing attack. Likewise, malicious websites and URLs can be created that resemble legitimate websites to avoid detection [59,60]. This misuse of GenAI in phishing attacks can advance such attacks by automating the generation of emails, pushing them to potential victims, and creating fake websites [61], making them more dangerous than traditional phishing attacks.

4.2.2. Phishing Attack Content

The risk of phishing attacks is significant, as according to a survey on cybersecurity breaches conducted by the United Kingdom (UK) government in 2024, phishing attacks affect 84 per cent of UK businesses [62]. Furthermore, the Australian Government's scam watcher [5] reported that financial losses due to phishing attacks in 2024 amounted to about AUD 20.5 million, and in January 2025 alone, these financial losses amounted to AUD 6.1 million. GenAI is rapidly evolving, mastering advanced capabilities in content creation, such as text, images, voice, and videos [30,63]. For instance, GenAI platforms can be used in website cloning as part of phishing attacks. With AI capabilities, legitimate websites can be easily cloned to deceive potential victims. Table 3 summarises the different types of content used in phishing attacks.

GenAI can be used to create realistic videos such as deepfakes, which are fake videos created to mimic the appearance and voice of a specific individual to deceive target victims to perform specific actions [64] in a phishing attack. Deepfakes are produced using individuals' visual and audio data, which can be retrieved by attackers using open-source intelligence (OSINT), closed-source intelligence (CSINT), or large public datasets of available information about individuals on various platforms such as social media. Therefore, concerns about deepfakes' use to spread misinformation are increasing [63,65].

GenAI provides the capability to generate convincing simulations of voice audio clips. This makes it difficult for automatic speaker verification (ASV) systems to distinguish between human voices and synthetic speech. Doan et al. [66] proposed a framework for audio deepfake detection called BTS-E, which correlates the speaker's breathing, talking, and pauses in an audio clip. This is effective, as natural human sounds and breathing are difficult to synthesise accurately using text-to-speech (TTS) systems. The study findings demonstrated that including the assessment of breathing sound characteristics significantly improved the performance of deepfake detection classifiers by 46 per cent.

Table 3. Phishing Content Usage.

Content	Description
Text	Crafting emails/text messages that are personalised to specific individuals.
Voice	Creating voice messages that can be used to impersonate trusted individuals.
Images	Creating images to add credibility to phishing attack content.
Videos	Creating realistic videos to add credibility to phishing attack content.

4.2.3. Language Models in Phishing Attacks

AI is rapidly evolving, and recently, large language models (LLMs) have been implemented through platforms such as ChatGPT and Gemini. These advances also introduce potential risks of misuse in conducting phishing attacks. For example, LLMs can be leveraged to produce convincing, personalised, targeted emails that increase the success rate of attacks using various communication channels [67]. LLMs are constantly developing and expanding by adding new capabilities. Malicious actors thrive on discovering and exploiting system vulnerabilities, and attackers can perform phishing attacks using LLM capabilities to access sensitive information and gain control [68,69]. The sophistication introduced by these models increases the challenges of phishing attack detection and prevention. For example, Qi et al. [70] introduced a framework called SpearBot, which uses an LLM to generate sophisticated phishing emails. This model was used to generate spear-phishing emails and continuously refine them, aiming to enhance their effectiveness in evading detection and deceiving target victims. The results of the study demonstrated a significant reduction in the cost and effort of conducting phishing attacks. It also highlighted that AI-generated emails were convincingly deceptive and often more effective compared to manually prepared phishing emails. These studies analysed GenAI's impact on phishing attacks by examining the use of LLMs in creating advanced and personalised phishing attacks due to their effectiveness in mimicking legitimate correspondence.

GenAI is increasingly being deployed to advance phishing attacks, allowing attackers to conduct well-crafted, deceptive attacks. Gupta et al. [71] discussed emerging threats enabled by GenAI, highlighting that advanced models such as ChatGPT can produce tailored communication that can be misused to advance phishing attacks. This enables criminals to use GenAI to understand human language and cognitive biases, which helps them rapidly automate the generation of deceptive, real-time, chat-based phishing messages, crafting convincing targeted interactions. The study also highlighted that ChatGPT, as a platform, is vulnerable to jailbreaks and prompt injection attacks, and hence, it can be compromised by attackers. Therefore, it is necessary to address how GenAI can benefit cybersecurity through secure coding, threat intelligence, and automation.

4.2.4. GenAI in Phishing Attack Personalisation

Most people use social media platforms daily, and some even have multiple profiles on different platforms. These profiles are used to share their experiences, cherished moments with family and friends, and even their thoughts and political opinions. With the large online presence nowadays, attackers have a wealth of information about the target individual's behaviour, work history, affiliations, and preferences, which can be analysed and leveraged in malicious phishing attacks. In targeted, personalised attacks, criminals invest in data collection and studying the potential victim's digital footprint to craft an effective phishing email that is hard to detect.

GenAI has the ability to produce realistic and tailored content that can be used to manipulate potential victims. This is achieved by creating content that is personalised and familiar to the target individual and does not raise any concerns about legitimacy.

Webb et al.'s [72] study demonstrated that GPT-3 and similar smart bots have the ability to mimic humans on various analogical reasoning tasks. This means that they have the ability to create text conversations and email messages that are not detectable, especially when crafted using previously collected personal information. GenAI tools can also be used to impersonate specific individuals' human factors by mimicking communication style, vocabulary, voice, and video to add to attackers' credibility.

Implementing proactive countermeasures against AI-driven cybersecurity threats starts with understanding AI systems and their underlying technologies and then understanding how they can be deployed in cyberattacks. Cybercriminals can leverage AI algorithms and big data to carry out sophisticated phishing attacks. AI platforms can generate targeted, personalised phishing content designed to avoid detection by traditional security systems [20,73]. GenAI capabilities using large language models [74] can scale social engineering attacks (such as phishing attacks) to an industrial level [58]. Furthermore, with the support of AI agents [75], phishing attacks can be automated. Open-source GenAI tools that are accessible to everyone represent the advancement of realistic AI-produced content such as text, images, voice, and video. This was evident with the introduction of ChatGPT (GPT-4) and Gemini, substantially raising risks of malicious use as part of social engineering attacks [58]. Therefore, to protect digital privacy and security, it is crucial to understand the threats imposed by GenAI tools on the current cybersecurity landscape.

4.3. What Are the Most Effective Human-Centred and Technological Solutions to Mitigate Phishing Attacks? (RQ3)

To answer this research question, we examined existing literature that has proposed solutions for detecting and preventing phishing attacks. Phishing attacks are constantly being studied to develop the best approaches to preventing these attacks, as shown in Figure 4, which summarises the available solutions presented in the literature. In addition, phishing solutions are presented in two different streams: technology-focused solutions and human-focused solutions.

4.3.1. Human-Focused Solutions

- Training and Awareness Programmes:

The focus is on the importance of educating users about the latest cybersecurity threats and best practices. This has been addressed in existing research by (a) introducing security training programmes [76], (b) increasing user knowledge about cybersecurity issues [77], and (c) informing users about experienced events [78]. These practices aim to reduce the likelihood of falling victim to cybersecurity attacks by increasing users' awareness of the actions required to protect themselves and the organisation as a whole, rather than relying on intuition to identify phishing emails [79].

An advancement in training methods is the design of an interactive security game that trains users on cybersecurity issues and attacks, which has been shown to be more effective than traditional training methods, as presented in different research publications, such as Dixon et al. [80], Ndibwile et al. [81], Wen et al. [77], Sheng et al. [82], and Kumaraguru et al. [83]. Interactive training material simulates possible cybersecurity attacks, allowing users to interact with these attacks through games and simulations on a real-time basis, receiving input and recommendations on users' responses.

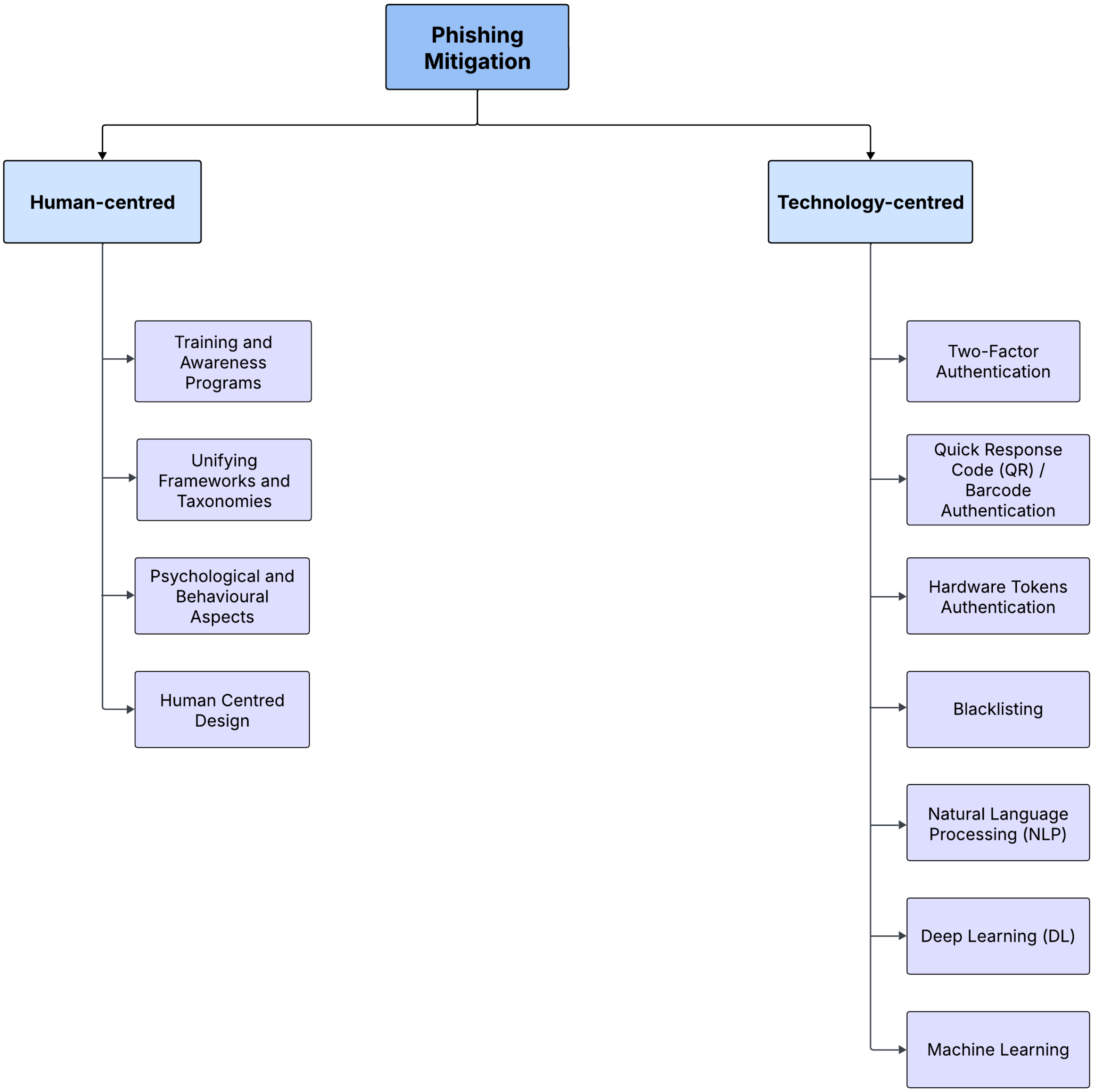

Phishing Mitigation
Human-centred
Technology-centred
Training and Awareness Programs
Unifying Frameworks and Taxonomies
Psychological and Behavioural Aspects
Human Centred Design
Two-Factor Authentication
Quick Response Code (QR) / Barcode Authentication
Hardware Tokens Authentication
Blacklisting
Natural Language Processing (NLP)
Deep Learning (DL)
Machine Learning

Figure 4. Human-centred versus technology-centred phishing mitigation classification.

Furthermore, Lim et al. [84] presented a training system that trains individuals on SMS and email phishing attacks. These training programmes simulate phishing attacks by incorporating the phishing site's URL into the received SMS and email and educating users on the dangers of the sites. The training led to a decrease in accessing these sites from 16 per cent to 12 per cent in the experiment, demonstrating the benefits of the training in reducing phishing attacks. Bowen et al. [8] focused on large corporations and government agencies in their study and used phishing attack simulations that resembled real phishing emails. They then tracked employees' responses and provided them with immediate feedback on their performance. This allowed for the identification of vulnerable groups that are more likely to fall victim to real phishing attacks. The study demonstrated that having targeted responses during phishing attack training strengthens the overall security posture.

- Unifying Frameworks and Taxonomies:

A unified understanding of frameworks and taxonomies is needed to address human factor vulnerabilities in relation to cyberattacks. Nurse [85] created an introductory taxonomy that highlighted different types of cyberattacks that exploit human vulnerabilities. This research reveals the need to explore approaches that prevent, detect, and deter attackers by studying their behaviour.

Psychological and behavioural aspects were considered when building security frameworks and taxonomies [86–89]. In addition, Metalidou et al. developed a preliminary framework [90] relating to human factors and a lack of cybersecurity awareness. The presented framework consists of several dimensions: (a) lack of motivation, referring to staff motivation to implement security practices and managers' identification of the source of staff motivation; (b) lack of awareness, referring to the lack of necessary knowledge of best security practices that prevent cyberattacks; (c) users' risky beliefs, relating to users' beliefs about security practice; (d) behaviour, referring to users' actions in relation to security practice; and (e) inadequate use of technology, meaning that human use of technology is as essential as the implementation of defence systems for the success of cyberattack prevention. Steves et al. [91] created a phishing email grading system called the "Phish Scale," which can be utilised in user security training. The scale was founded on two dimensions: (a) the number of cues in the email, which consist of indicators commonly present in phishing emails, such as spelling and grammatical mistakes, a sense of urgency, limited-time offers, and mistakes in visual presentation; and (b) target audience alignment, which indicates how the email is relevant to the target victims, such as personalised emails that include user information and preferences, etc.

Henshel et al. [9] proposed a predictive framework for cybersecurity risk assessment that focuses on the trust factor. They argued that it is important to have a holistic perspective on risk assessment that does not overlook the complex human roles of system users, defenders, or even attackers. In this study, the authors used 'trust' only for humans and 'confidence' when discussing non-human elements. The objective is to consider unique human characteristics such as integrity, attention, accuracy, and expertise when attempting to understand the human influence in conducting security risk assessments.

More investment is required to create more concise frameworks and taxonomies with a standardised vocabulary, especially in relation to studying human factors in phishing attacks, which establishes a foundation for future research.

- Psychological and Behavioural Aspects:

Several studies have recommended changes in the psychological, behavioural, and attitudinal aspects of users. Choong and Theofanos [92] evaluated the strengths and weaknesses of a company's security policies by examining employees' perspectives on these policies. The study found that security policies tend to neglect human factors and therefore fail to achieve the desired results. For example, regarding password maintenance and change policies, passwords require specific complexity and length, and it can be hard for individuals to remember them, leading them to write the password down and leave it in the office or on a sticky note that can be easily retrieved by attackers.

Another study by Levesque et al. [93] explored user actions in relation to best security practices, such as installing the latest antivirus software and malware detection. The results showed that users with high computer knowledge were more susceptible to phishing attacks. In addition, the authors assessed the age and gender of the users and found that they were not directly related to malware attacks, which differs from the finding presented by Nsiempba et al. [94].

Kavvadias and Kotsilieris' study [95] assessed both demographic factors, such as gender, education, age, technical skills, and psychological traits, including impulsivity, emotion, and trust. The study found that among the demographic factors, older age, younger age, females in certain contexts, and those with lower technical skills were more susceptible to phishing attacks. As for psychological traits, individuals who were more impulsive, more trusting, and prone to emotional responses to authority or urgency cues were also at greater risk. The study concluded that implementing tailored awareness training programmes that consider individual differences can help reduce individuals' vulnerability to phishing attacks.

- Human-Centred Design:

Human-focused designs propose adding improvements to the end-user interfaces of applications to increase user knowledge and awareness, leading to informed decision-making. Xiong et al. [78] presented a training interface related to phishing attacks that included embedded warning messages, and the authors derived the results of the experiment, which indicated that the appearance of these warning messages led to the successful identification of phishing websites.

William and Li [96] designed an experiment to assess the cognitive process of identifying whether web pages are legitimate or not by the presence of the https padlock icon on the website's link. During the experiment, the results indicated a lack of knowledge in understanding the purpose of the icon, which indicates that a website is secure. Furthermore, Zhao et al. [97] studied the design of phishing websites and how they are presented as legitimate websites. The authors implemented a toolkit to study the behaviour of participants when targeted with examples of phishing attacks.

4.3.2. Technology-Focused Solutions

- Two-Factor Authentication:

With the increase in scams targeting vulnerable individuals to deceive them and steal their money, many banks have implemented two-factor authentication techniques to increase the security of their bank transactions, as explored by Varshney et al. [23]. However, despite adding a second level of protection, two-factor authentication can be ineffective in complex phishing attacks in which an attacker mimics the bank's website to deceive the bank client into sharing their assigned login credentials along with the second authentication factor in real time.

- Quick-Response Code (QR) or Barcode Authentication:

Several researchers, such as Dodson et al. [98] and Jindal and Misra [99], have studied phishing prevention mechanisms that allow the exchange of credentials in different formats instead of the traditional plain text format. As such, barcodes can be used for authentication, as they are generated by a legitimate website and then scanned by the associated mobile application for user verification purposes. QR codes can also be used to send secret challenges to website users, who respond using their dedicated private keys. Successfully responding to these challenges is treated as a mandatory website login verification step. The strength of the QR and barcode authentication mechanism is that the secret resides within the user's mobile phone; therefore, to succeed in gaining unauthorised access, one must steal the login credentials and gain access to the user's mobile phone simultaneously.

- Hardware Token Authentication:

A vendor-specific hardware token is used to hold the secret keys required during the user authentication process. Lu et al. [100] studied the use of hardware tokens, such as RSA keys, which display, in real time, second-factor authentication information, such as an OTP or a PIN. The user then enters the OTP or PIN along with their login credentials for authentication.

- Blacklisting:

This anti-phishing technique is based on blacklisting suspicious sender emails and hyperlinks. The phishing email detection process involves extracting the sender's email address from the received email and comparing it with existing blacklists to verify its legitimacy and prevent phishing attempts [101]. The efficiency of this technique depends on the manual efforts of revising and maintaining the existing blacklists, and on individuals' effective reporting of phishing emails. Several public phishing databases, such as PhishTank [102], can be referenced when studying phishing attacks.

- Natural Language Processing (NLP):

NLP techniques can be used to extract specific features from emails in relation to syntax, context, content, and semantics, which can be used to detect phishing emails [22]. NLP is used in various tasks, such as text classification and information extraction, and to automate the phishing detection process. It can be used for semantic analysis in order to classify emails as legitimate or spam based on the words in the email and their meaning [22].

- Deep Learning (DL):

DL can be deployed for email feature extraction instead of manual extraction. The use of DL helps increase the accuracy and efficiency of feature extraction, as studied by Sallom et al. [22]. This research found that models such as convolutional neural networks (CNNs) can support phishing email detection, as well as email header, body, and phrase processing. It can also be used to extract features from URLs to detect attempts to increase reliability.

Alhuzali et al. [103] conducted a study comparing and evaluating 14 DL and ML models across 10 different datasets. They found that advanced DL models such as BERT and RoBERTa achieved 99 per cent accuracy in detecting phishing emails, outperforming traditional ML models. Among the DL approaches, the BERT and RoBERTa models effectively captured subtle patterns in phishing emails, whereas other DL models, including CNNs, Long Short-Term Memory (LSTM), and DistilBERT, did not reach the same level of accuracy. As for ML models, the Stochastic Gradient Descent (SGD) classifier achieved the highest average accuracy rate at 98 per cent, and models such as Extra Trees and Random Forest also delivered competitive results. However, DL models generally require substantial computational resources and significant training time to be applicable in real-world applications.

Mahmud et al. [104] presented a hybrid DL model for detecting SMS phishing, commonly known as 'smishing'. The model, called CNN-Bi-GRU, first preprocesses the SMS text by cleaning and converting it to a numerical format. It then identifies text patterns and captures contextual word sequences in both the forward and backward directions. This hybrid DL approach achieved an accuracy of 99.82 per cent in detecting smishing messages.

- Machine Learning (ML):

The application of ML algorithms in phishing detection has been explored in several studies. The main aim of the ML anti-phishing technique is to use ML algorithms for email classification by training them using legitimate and phishing emails [22]. Alam et al. [105] conducted an experiment with two ML models, namely Random Forest and Decision Tree, and then measured their accuracy. The results of the experiment showed that the Random Forest algorithm achieved an accuracy of 97 per cent, and Decision Tree achieved an accuracy of 93.84 per cent. Similarly, Rawal et al. [106] explored the use of ML algorithms to classify whether emails are phishing emails or ham by extracting relevant features and applying ML classifiers. The authors used natural language processing techniques to extract key features from an email. Then, using ML classifiers on the phishing email dataset, performance was evaluated using metrics such as accuracy, precision, and recall. The results suggested that ML is promising for detecting phishing emails, as Support Vector Machines (SVM) and Random Forest classifiers achieved the best performance, with an accuracy of 99.87 per cent in email classification. Using ML in phishing attack detection is a promising area of research, and more efforts should be directed towards it.

Nagy et al. [107] investigated phishing URL detection using various ML models. The experiment compared sequential processing, in which models are trained step by step, with parallel processing. Multiple components are trained simultaneously using threads or multi-processing to improve the phishing URL detection speed without sacrificing accuracy. The models used were Random Forest (RF), Naive Bayes (NB), LSTM, and CNN. Among them, NB achieved the highest accuracy of 96.01 per cent, while RF, LSTM, and CNN each achieved a 100 per cent recall rate in detecting phishing URLs.

Brissett and Wall [108] proposed an approach that combines ML and watermarking techniques to detect AI-generated phishing emails. The approach converts emails into numerical representations and uses the Logistic Regression algorithm to classify them as either legitimate or phishing emails. As part of the research, watermark tokens were suggested and tested as a detection mechanism by adding watermark tokens during AI content creation in order to facilitate the detection of AI-generated phishing emails.

Eze and Shamir [109] studied GenAI-generated phishing emails using ML methods such as topic modelling and deep neural networks. Their approach identified these emails with high accuracy by examining writing styles, sentiments, and vocabulary usage. The study also explored key challenges, including false-positive email classification and the ongoing need to continuously re-train the phishing detection systems with new AI-generated phishing email samples.

This section explored the different solutions to phishing attacks by highlighting the improvements in detection rates due to advances in ML and DL studies. However, it is important to consider the risks and additional challenges introduced by AI-driven phishing attacks, which traditional anti-phishing solutions will struggle to address due to the sophisticated and human-like GenAI phishing content that exploits targeted victims' vulnerabilities.

4.3.3. GenAI in Attack Prevention

Cyber threats are evolving, and with the challenges presented by GenAI-enabled phishing attacks, it is important to develop adaptive cybersecurity defence systems against evolving attack vectors. In RQ2, we answered the question "How are phishing attacks evolving with the introduction of GenAI? (RQ2)," providing an opportunity to examine how GenAI can advance phishing attacks. However, GenAI technology can also be used to enhance security measures. We explore this concept further below.

- Advanced Training Programmes: GenAI can be used in cybersecurity training and awareness programmes by developing sophisticated phishing attacks, simulating potential threats such as phishing emails [110], and identifying the best approach to handling them. These simulations allow employees to be trained on the nuances of phishing attacks, thus reducing the probability that breaches will be successful.
- Security Testing: GenAI can improve security testing by automating the generation of test cases. Hilario et al. [111] introduced the potential of using GenAI in software testing to ensure that applications are secured against a wide range of attacks.
- Defence Mechanisms: GenAI can be used to simulate various attack scenarios. This supports security professionals in developing adaptive defensive strategies [58,112] against sophisticated threats. Sai et al. [113] provided a review examining different security products that can improve security measures by leveraging GenAI tools such

as Google Cloud Security AI Workbench and Microsoft Security Copilot. Sai et al. [113] also identified that GenAI can be used in areas such as threat intelligence, vulnerability scanning, and secure code development. Although these studies have not explicitly examined their application to phishing attacks, we foresee great potential in applying them to the prevention and detection of phishing attacks. In addition, AI-powered techniques such as PhiShield, a spam filter browser extension, are used for real-time protection against phishing emails. The system combines signature-based checks with an LSTM AI model to alert users to phishing emails, achieving a detection accuracy of 98 per cent [114].

5. Discussion

Phishing attacks are considered the most common cybersecurity attacks, and they cause significant damage to a large number of organisations and individuals by exploiting people's vulnerabilities. However, phishing risks have increased significantly by capitalising on technology such as GenAI to, for example, produce personalised phishing messages. These targeted phishing attacks require thorough research and preparation to gather information about and insights into the target victims' details (for example, their company, family members, or posts), online presence, and behaviours. In addition, when targeting organisations, attackers may gather information about the implemented systems and infrastructure, which can be used in attacks to customise phishing messages that, for example, require employees to deploy urgent software patches or log in to a company website link after an upgrade or maintenance, which is then used in subsequent attacks. Phishing attacks have different forms, though emails are the most common communication medium used in phishing attacks because of their ease of use and low cost. Other mediums include deceitful text messages or phone calls that impersonate legitimate companies or individuals. Attackers with malicious intentions scam victims into sharing their data without raising doubts.

To further study the impact of phishing attacks on Australian businesses and individuals, we examined all reported phishing attacks. Our study referenced officially published phishing attack data from the Australian Government. ASD is an Australian Government agency that collects and studies intelligence, including cyberwarfare. The ASD Cyber Threat Report 2023–2024 reported that many Australian-based individuals and businesses faced cyberattacks, as demonstrated in Figure 5, which provides the number of phishing attacks reported in Australia from January 2020 to January 2025.

Phishing attacks will continue to evolve, with attackers taking advantage of technological advances to increase their financial gains. The Australian Government's scam watcher [5] reported significant financial losses from phishing attacks over a period of five years (Figure 6), with the amount of losses being AUD 1.6 million in 2020 and AUD 4.3 million in 2021. The amount of financial losses due to phishing attacks exponentially increased by approximately 472 per cent, from AUD 4.3 million in 2021 to AUD 24.6 million in 2022. In the consecutive years, the loss rate only increased by 4.88 per cent, making it AUD 25.8 million in 2023, before dropping to AUD 20.5 million in 2024. For the year 2025, in January 2025, the financial losses were AUD 6.1 million.

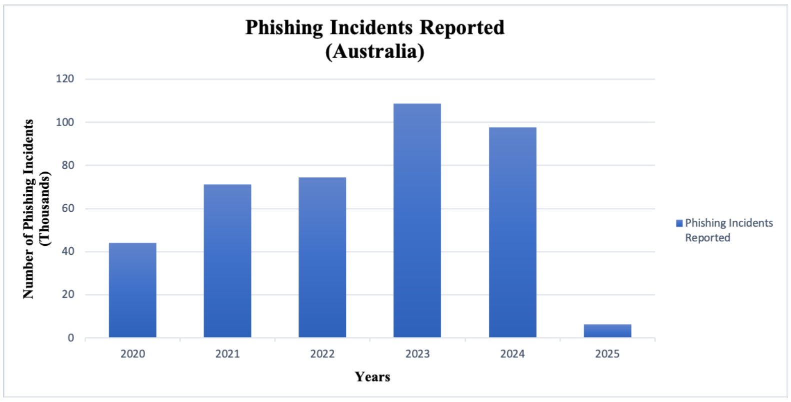


Figure 5. Number of phishing incidents reported in Australia. The data presented in this chart were obtained from the Australian Government's scam watcher site [5].

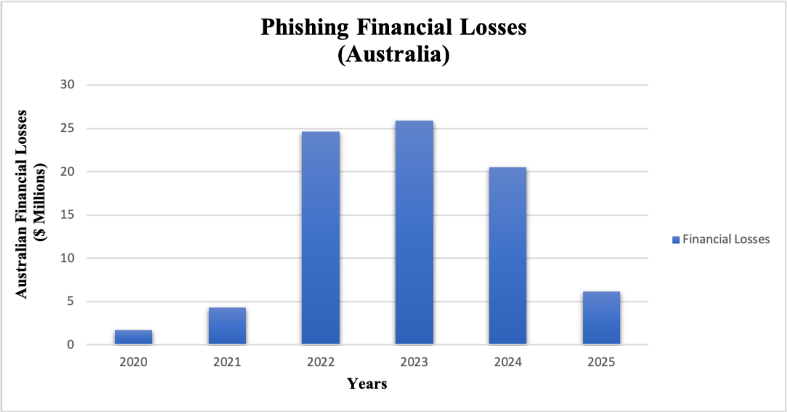


Figure 6. Phishing financial losses in Australia. The data presented in this chart were obtained from the Australian Government's scam watcher site [5].

A deeper analysis was conducted to understand the characteristics of the phishing victims, such as their age group and gender. This was achieved by referring to the Australian Government's scam watcher [5] to study the trend over the past five years, from 2020 to 2024, including January 2025 (Figures 7 and 8). In reference to the number of phishing attacks reported by age group, for individuals under 18 years of age, there were 1256 reports, and there were 10,155 reports for the age group 18–24. For the age group 25–34, the number of reports was 31,574, and there were 40,913 for the age group 35–44. Phishing attack reports continued to increase as age increased, with 45,093 reports for the age group 45–54 and 54,247 reports for the age group 55–64. The number of reported attacks increased by 83 per cent for the age group 65 years and above, with a total of 99,488 reported phishing incidents. It is also important to note that the number of reports by individuals who did not specify their age was 119,948. If we exclude the unspecified age group, we notice that the number of victims increases as the age group increases, making older adults more prone to deception and phishing attacks. In terms of gender analysis, the number of males who reported a phishing attack was 200,139, while the number of females was 183,743—49.7 per cent compared to 45.6 per cent—as presented in Figure 8.

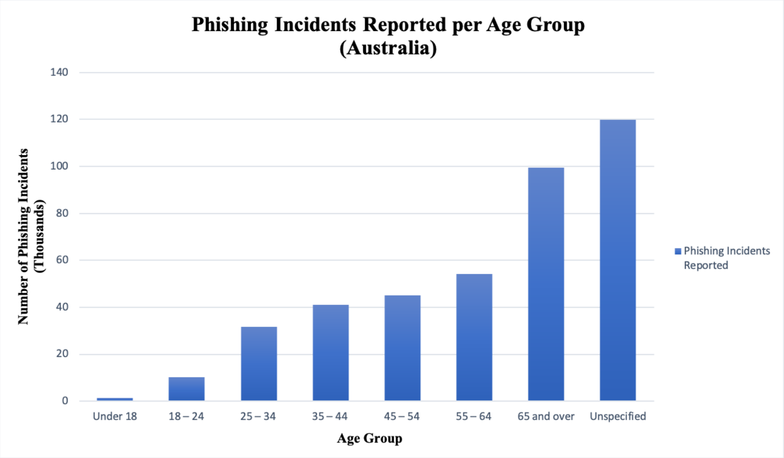


Figure 7. Phishing incidents per age group in Australia. The data presented in this chart were obtained from the Australian Government's scam watcher site [5].

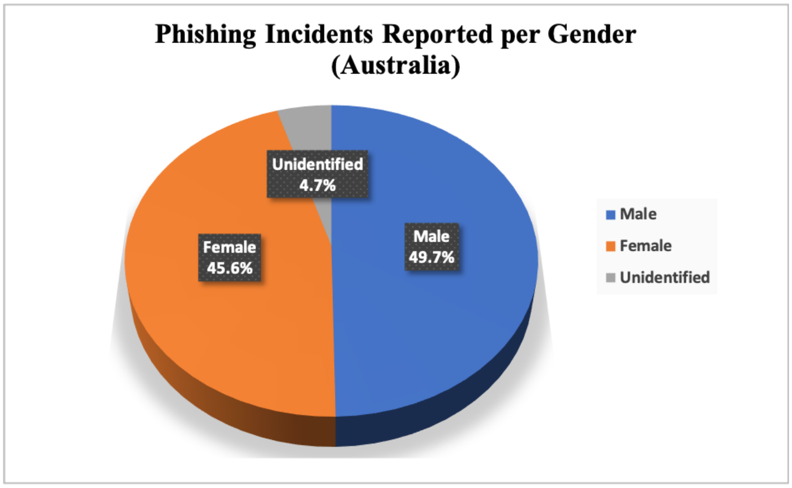


Figure 8. Phishing incidents by gender in Australia. The data presented in this chart were obtained from the Australian Government's scam watcher site [5].

By reviewing the current literature on the topic of study, we evaluated the different human factors addressed in phishing attacks and concluded that the focus was narrowed in most cases to psychological, cognitive, and emotional aspects. However, other studies have focused on human characteristics, such as demographics, gender, and educational level. In Table 4, we provide a summary and comparison of the relevant literature on human factors in phishing attacks, mentioning the human factors studied, the research goal, key findings, and any limitations.

When designing human-focused solutions to prevent phishing attacks, it is important to take a holistic approach that considers all the factors that contribute to individuals' behaviours. Human behaviour is affected by both internal and external factors. Internal factors include different characteristics, personalities, experiences, and biological natures. External factors include organisational culture, system design, social norms, training, and unforeseen incidents. As a contribution to the field of study, we developed a holistic view of all possible human factors and attributes that can be exploited in phishing attacks, as illustrated in Figure 9, providing a more detailed description of our proposed classification for both internal and external human factors.

Table 4. Human factors in phishing attacks: research comparison analysis.

Paper	Human Factors Covered	Research Goal	Key Findings	Limitations
[115]	Psychological factors	Study factors influencing phishing reporting intentions	Identified factors affecting reporting intentions	Online questionnaire (Number of participants = 284)
[116]	Emotional factors in phishing victimisation	Explore emotional factors significant in phishing victimisation	1. Attackers manipulate victims' psychology and emotions; 2. Comparison of phishing types; 3. Review of training approaches.	Relied on literature review without original empirical data
[38]	Cognitive factors in phishing detection	Identify human and cognitive factors in phishing attacks	1. Cybercriminals exploit human vulnerabilities; 2. Difficulty in identifying phishing emails; 3. Importance of cognitive and psychological factors.	Limited to existing literature only
[117]	Demographic, behavioural, and psychological factors	Examine factors influencing phishing susceptibility through a simulated campaign	1. The limited effects of gender and age; 2. Previous phishing victims are more susceptible; 3. Impulsivity is correlated with phishing susceptibility; 4. Better security habits are linked to lower susceptibility.	Focused on a single university population
[118]	Emotional and psychological factors	Explore human/emotional factors in phishing victimisation	1. Emotional manipulation in attacks; 2. Phishing types; 3. Training approaches.	Limited to existing literature only

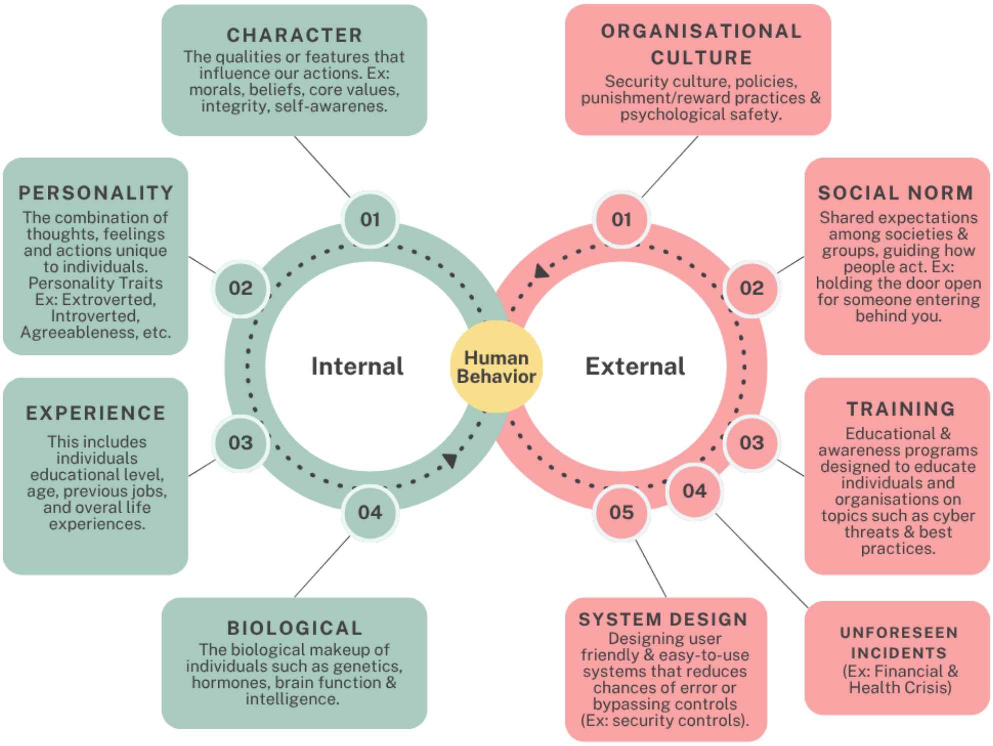

CHARACTER
The qualities or features that influence our actions. Ex: morals, beliefs, core values, integrity, self-awarenes.
ORGANISATIONAL CULTURE
Security culture, policies, punishment/reward practices & psychological safety.
PERSONALITY
The combination of thoughts, feelings and actions unique to individuals. Personality Traits Ex: Extroverted, Introverted, Agreeableness, etc.
SOCIAL NORM
Shared expectations among societies & groups, guiding how people act. Ex: holding the door open for someone entering behind you.
01
02
03
04
Internal
Human Behavior
External
01
02
03
04
05
EXPERIENCE
This includes individuals educational level, age, previous jobs, and overal life experiences.
TRAINING
Educational & awareness programs designed to educate individuals and organisations on topics such as cyber threats & best practices.
BIOLOGICAL
The biological makeup of individuals such as genetics, hormones, brain function & intelligence.
SYSTEM DESIGN
Designing user friendly & easy-to-use systems that reduces chances of error or bypassing controls (Ex: security controls).
UNFORESEEN INCIDENTS
(Ex: Financial & Health Crisis)

Figure 9. Human behaviour factors.

AI-driven phishing attacks significantly improve the sophistication of cybercrime by leveraging GenAI for personalisation, scalability, and multichannel deception. Unlike traditional phishing, which relies on generic templates and broad targeting, AI-powered methods exploit advanced data analysis to craft highly tailored and realistic campaigns, thereby making them more effective and difficult to detect. Table 5 provides a comparison between traditional phishing attacks and AI-driven phishing attacks, assessing message quality, personalisation, attack scale, audience-targeting mechanisms, attack vectors, and foreseen detection challenges.

Table 5. Traditional versus AI-driven phishing attacks.

Aspect	Traditional Phishing	AI-Driven Phishing
Message Quality	Message is generic, with grammatical errors and typing mistakes	Mimics real-life communication styles, with no spelling or grammatical mistakes
Personalisation	Broad targeting without personalisation	Carefully crafted, personalised messages
Scale	Manual messaging with limited scalability	High-volume generation and automation
Targeting Approach	Indiscriminately large audience targeting	Strategic targeting based on AI-driven analysis
Attack Vectors	Multi-channel; primary attack channel is email	Multi-channel, deploying AI technology
Detection Challenges	Easier to detect due to grammatical/typing errors	Harder to detect, with AI-driven attacks overcoming traditional controls

6. Future Research

More research is required to combat evolving phishing attacks, as researchers and security professionals must be up to date with the latest trends and techniques of phishing attacks. Based on our research, we recommend the following areas for future research:

- Awareness and Education Programmes: With the complexity introduced by AI usage in phishing attacks, traditional cybersecurity awareness programmes are ineffective. The focus should be on increasing users' knowledge of AI-driven attacks and advanced techniques that cybercriminals are deploying to deceive individuals and organisations. These programmes should also include interactive, situation-based training modules that highlight how human factors are exploited in phishing attacks.
- Implementing AI and ML in Defence Systems: This focuses on leveraging the latest techniques and algorithms to prevent GenAI-driven attacks. Future research in this area should focus on the introduction of robust AI/ML models and strategies to detect and defend against sophisticated phishing attacks.
- Explainable AI: More research should focus on AI model transparency; this would support security professionals and researchers in studying how AI models in defence systems identify anomalies and make decisions.

7. Conclusions

The threat of phishing attacks is increasing exponentially with advances in technology. Attackers continuously exploit human vulnerabilities by compromising digital interactions and breaking the foundations of trust and security. This study examined the importance of designing human-focused solutions to prevent phishing attacks and the need for holistic solutions that consider all factors that contribute to people's behaviours. Human behaviour is affected by both internal and external factors. Internal factors include different characteristics, personalities, experiences, and biological natures. External factors include organisational culture, social norms, system design, training, and unforeseen incidents, such as financial and health crises. This study also examined the capabilities introduced by GenAI in the context of phishing attacks and how this technology has increased the effectiveness of these attacks. GenAI can analyse data and produce realistic content, such as voice, video, images, and text. This is then used in targeted, personalised attacks to increase their success rates. GenAI indicates a paradigm shift in the manner in which phishing attacks are executed, making them undetectable and scalable. More research is required to introduce unconventional measures to protect our cybersecurity against phishing attacks, such as implementing AI/ML defence systems, investing in upskilling security professionals and organisations' knowledge using Explainable AI, and enhancing cybersecurity training programmes to include AI-driven attacks and human factor components. As AI continues to evolve and become accessible to everyone, it is crucial to proactively combat emerging AI threats by introducing countermeasures and increasing awareness.

Author Contributions: Conceptualisation, R.J.; methodology, R.J.; validation, R.J., J.L. and C.N.; formal analysis, R.J.; investigation, R.J.; writing—original draft preparation, R.J.; writing—review and editing, R.J., J.L. and C.N.; visualisation, R.J.; supervision, J.L. and C.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
GenAI	Generative Artificial Intelligence
ML	Machine Learning
NLP	Natural Language Processing
DL	Deep Learning

References

1. Schatz, D.; Bashroush, R.; Wall, J. Towards a more representative definition of cyber security. *J. Digit. Forensics Secur. Law* **2017**, *12*, 8. [CrossRef]
2. Sule, M.J.; Zennaro, M.; Thomas, G. Cybersecurity through the lens of digital identity and data protection: Issues and trends. *Technol. Soc.* **2021**, *67*, 101734. [CrossRef]
3. King, Z.M.; Henshel, D.S.; Flora, L.; Cains, M.G.; Hoffman, B.; Sample, C. Characterizing and measuring maliciousness for cybersecurity risk assessment. *Front. Psychol.* **2018**, *9*, 39. [CrossRef]
4. Australian Signals Directorate (ASD) Cyber Threat Report 2023–2024. Available online: https://www.cyber.gov.au/sites/default/files/2024-11/asd-cyber-threat-report-2024.pdf (accessed on 1 March 2025).
5. Australian Scamwatch. Available online: https://www.scamwatch.gov.au (accessed on 1 March 2025).
6. Mohammad, T.; Hussin, N.A.M.; Husin, M.H. Online safety awareness and human factors: An application of the theory of human ecology. *Technol. Soc.* **2022**, *68*, 101823. [CrossRef]
7. Hong, Y.; Furnell, S. Understanding cybersecurity behavioral habits: Insights from situational support. *J. Inf. Secur. Appl.* **2021**, *57*, 102710. [CrossRef]
8. Bowen, B.M.; Devarajan, R.; Stolfo, S. Measuring the human factor of cyber security. In Proceedings of the 2011 IEEE International Conference on Technologies for Homeland Security (HST), Waltham, MA, USA, 15–17 November 2011; IEEE: Piscataway, NJ, USA, 2011; pp. 230–235.

9. Henshel, D.; Cains, M.; Hoffman, B.; Kelley, T. Trust as a human factor in holistic cyber security risk assessment. *Procedia Manuf.* **2015**, *3*, 1117–1124. [CrossRef]
10. Alanazi, M.; Freeman, M.; Tootell, H. Exploring the factors that influence the cybersecurity behaviors of young adults. *Comput. Hum. Behav.* **2022**, *136*, 107376. [CrossRef]
11. Shah, P.R.; Agarwal, A. Cybersecurity behaviour of smartphone users through the lens of fogg behaviour model. In Proceedings of the 2020 3rd International Conference on Communication System, Computing and IT Applications (CSCITA), Mumbai, India, 3–4 April 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 79–82.
12. Lebek, B.; Uffen, J.; Breitner, M.H.; Neumann, M.; Hohler, B. Employees' information security awareness and behavior: A literature review. In Proceedings of the 2013 46th Hawaii International Conference on System Sciences, Wailea, HI, USA, 7–10 January 2013; IEEE: Piscataway, NJ, USA, 2013; pp. 2978–2987.
13. Li, L.; He, W.; Xu, L.; Ash, I.; Anwar, M.; Yuan, X. Investigating the impact of cybersecurity policy awareness on employees' cybersecurity behavior. *Int. J. Inf. Manag.* **2019**, *45*, 13–24. [CrossRef]
14. Simonet, J.; Teufel, S. The influence of organizational, social and personal factors on cybersecurity awareness and behavior of home computer users. In Proceedings of the ICT Systems Security and Privacy Protection: 34th IFIP TC 11 International Conference, SEC 2019, Lisbon, Portugal, 25–27 June 2019; Proceedings 34; Springer: Berlin/Heidelberg, Germany, 2019; pp. 194–208.
15. Al-Qaysi, N.; Granić, A.; Al-Emran, M.; Ramayah, T.; Garces, E.; Daim, T.U. Social media adoption in education: A systematic review of disciplines, applications, and influential factors. *Technol. Soc.* **2023**, *73*, 102249. [CrossRef]
16. Chui, M.; Hazan, E.; Roberts, R.; Singla, A.; Smaje, K. *The Economic Potential of Generative AI*; McKinsey & Company: New York, NY, USA, 2023.
17. Hatzius, J.; Briggs, J.; Kodnani, D.; Pierdomenico, G. The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani). *Goldman Sachs* **2023**, *1*. Available online: https://static.poder360.com.br/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf (accessed on 1 April 2024).
18. Kaur, R.; Gabrijelčič, D.; Klobučar, T. Artificial intelligence for cybersecurity: Literature review and future research directions. *Inf. Fusion* **2023**, *97*, 101804. [CrossRef]
19. Bueermann, G.; Rohrs, M. World Economic Forum (2024) Global Cybersecurity Outlook 2024. Available online: https://www3.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2024.pdf (accessed on 1 April 2024).
20. Renaud, K.; Warkentin, M.; Westerman, G. *From ChatGPT to HackGPT: Meeting the Cybersecurity Threat of Generative AI*; MIT Sloan Management Review: Cambridge, MA, USA, 2023.
21. Lastdrager, E.E. Achieving a consensual definition of phishing based on a systematic review of the literature. *Crime Sci.* **2014**, *3*, 9. [CrossRef]
22. Salloum, S.; Gaber, T.; Vadera, S.; Shaalan, K. Phishing email detection using natural language processing techniques: A literature survey. *Procedia Comput. Sci.* **2021**, *189*, 19–28. [CrossRef]
23. Varshney, G.; Kumawat, R.; Varadharajan, V.; Tupakula, U.; Gupta, C. Anti-phishing: A comprehensive perspective. *Expert Syst. Appl.* **2024**, *238*, 122199. [CrossRef]
24. Hakim, Z.M.; Ebner, N.C.; Oliveira, D.S.; Getz, S.J.; Levin, B.E.; Lin, T.; Lloyd, K.; Lai, V.T.; Grilli, M.D.; Wilson, R.C. The Phishing Email Suspicion Test (PEST) a lab-based task for evaluating the cognitive mechanisms of phishing detection. *Behav. Res. Methods* **2021**, *53*, 1342–1352. [CrossRef]
25. Wash, R.; Cooper, M.M. Who provides phishing training? facts, stories, and people like me. In Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–12.
26. Dhamija, R.; Tygar, J.D.; Hearst, M. Why phishing works. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 22–27 April 2006; pp. 581–590.
27. Goodfellow, I.J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial nets. In Proceedings of the NIPS'14: Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 2, Montreal, QC, Canada, 8–13 December 2014; pp. 2672–2680. https://dl.acm.org/doi/10.5555/2969033.2969125 (accessed on 1 April 2025).
28. Bengesi, S.; El-Sayed, H.; Sarker, M.K.; Houkpati, Y.; Irungu, J.; Oladunni, T. Advancements in generative AI: A comprehensive review of GANs, GPT, autoencoders, diffusion model, and transformers. *IEEE Access* **2024**, *12*, 69812–69837. [CrossRef]
29. Ronge, R.; Maier, M.; Rathgeber, B. Towards a definition of Generative artificial intelligence. *Philos. Technol.* **2025**, *38*, 31. [CrossRef]
30. Archana, R.; Jeevaraj, P.E. Deep learning models for digital image processing: A review. *Artif. Intell. Rev.* **2024**, *57*, 11. [CrossRef]
31. Desolda, G.; Ferro, L.S.; Marrella, A.; Catarci, T.; Costabile, M.F. Human factors in phishing attacks: A systematic literature review. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [CrossRef]
32. Pham, H.C.; Pham, D.D.; Brennan, L.; Richardson, J. Information security and people: A conundrum for compliance. *Australas. J. Inf. Syst.* **2017**, *21*. [CrossRef]

33. Chowdhury, N.H.; Adam, M.T.; Skinner, G. The impact of time pressure on human cybersecurity behavior: An integrative framework. In Proceedings of the 2018 26th International Conference on Systems Engineering (ICSEng), Sydney, NSW, Australia, 18–20 December 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–10.
34. Hwang, M.I.; Helser, S. Cybersecurity educational games: A theoretical framework. *Inf. Comput. Secur.* **2022**, *30*, 225–242. [CrossRef]
35. Kalhoro, S.; Rehman, M.; Ponnusamy, V.; Shaikh, F.B. Extracting key factors of cyber hygiene behaviour among software engineers: A systematic literature review. *IEEE Access* **2021**, *9*, 99339–99363. [CrossRef]
36. Herath, T.B.; Khanna, P.; Ahmed, M. Cybersecurity practices for social media users: A systematic literature review. *J. Cybersecur. Priv.* **2022**, *2*, 1–18. [CrossRef]
37. Zhang, Z.; Gupta, B.B. Social media security and trustworthiness: Overview and new direction. *Future Gener. Comput. Syst.* **2018**, *86*, 914–925. [CrossRef]
38. Arévalo, D.; Valarezo, D.; Fuertes, W.; Cazares, M.F.; Andrade, R.O.; Macas, M. Human and Cognitive Factors involved in Phishing Detection. A Literature Review. In Proceedings of the 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE), Las Vegas, NV, USA, 24–27 July 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 608–614.
39. Andrade, R.O.; Yoo, S.G. Cognitive security: A comprehensive study of cognitive science in cybersecurity. *J. Inf. Secur. Appl.* **2019**, *48*, 102352. [CrossRef]
40. Naqvi, B.; Perova, K.; Farooq, A.; Makhdoom, I.; Oyedeji, S.; Porras, J. Mitigation strategies against the phishing attacks: A systematic literature review. *Comput. Secur.* **2023**, *132*, 103387. [CrossRef]
41. Ayeni, R.K.; Adebiyi, A.A.; Okesola, J.O.; Igbekele, E. Phishing attacks and detection techniques: A systematic review. In Proceedings of the 2024 International Conference on Science, Engineering and Business for Driving Sustainable Development Goals (SEB4SDG), Omu-Aran, Nigeria, 2–4 April 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 1–17.
42. Kyaw, P.H.; Gutierrez, J.; Ghobakhlou, A. A Systematic Review of Deep Learning Techniques for Phishing Email Detection. *Electronics* **2024**, *13*, 3823. [CrossRef]
43. Thakur, K.; Ali, M.L.; Obaidat, M.A.; Kamruzzaman, A. A systematic review on deep-learning-based phishing email detection. *Electronics* **2023**, *12*, 4545. [CrossRef]
44. Schmitt, M.; Flechais, I. Digital deception: Generative artificial intelligence in social engineering and phishing. *Artif. Intell. Rev.* **2024**, *57*, 324. [CrossRef]
45. Baker, J. The technology–organization–environment framework. In *Information Systems Theory: Explaining and Predicting Our Digital Society, Vol. 1*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 231–245.
46. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E.; et al. The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71. [CrossRef]
47. Kitchenham, B. Procedures for performing systematic reviews. *Keele UK Keele Univ.* **2004**, *33*, 1–26.
48. Fink, A. *Conducting Research Literature Reviews: From the Internet to Paper*; Sage Publications: Thousand Oaks, CA, USA, 2019.
49. Torraco, R.J. Writing integrative literature reviews: Guidelines and examples. *Hum. Resour. Dev. Rev.* **2005**, *4*, 356–367. [CrossRef]
50. Stanton, J.M.; Stam, K.R.; Mastrangelo, P.; Jolton, J. Analysis of end user security behaviors. *Comput. Secur.* **2005**, *24*, 124–133. [CrossRef]
51. Zhang, R.; Bello, A.; Foster, J.L. BYOD security: Using dual process theory to adapt effective security habits in BYOD. In Proceedings of the Future Technologies Conference, Vancouver, BC, Canada, 20–21 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 372–386.
52. Oner, U.; Cetin, O.; Savas, E. Human factors in phishing: Understanding susceptibility and resilience. *Comput. Stand. Interfaces* **2025**, *94*, 104014. [CrossRef]
53. Mayer, P.; Kunz, A.; Volkamer, M. Reliable behavioural factors in the information security context. In Proceedings of the 12th International Conference on Availability, Reliability and Security, Reggio Calabria, Italy, 29 August–1 September 2017; pp. 1–10.
54. Safa, N.S.; Sookhak, M.; Von Solms, R.; Furnell, S.; Ghani, N.A.; Herawan, T. Information security conscious care behaviour formation in organizations. *Comput. Secur.* **2015**, *53*, 65–78. [CrossRef]
55. Corradini, I.; Nardelli, E. Building organizational risk culture in cyber security: The role of human factors. In Proceedings of the Advances in Human Factors in Cybersecurity: Proceedings of the AHFE 2018 International Conference on Human Factors in Cybersecurity, Loews Sapphire Falls Resort at Universal Studios, Orlando, FL, USA, 21–25 July 2018; Springer: Berlin/Heidelberg, Germany, 2019; pp. 193–202.
56. Alsharnouby, M.; Alaca, F.; Chiasson, S. Why phishing still works: User strategies for combating phishing attacks. *Int. J. Hum.-Comput. Stud.* **2015**, *82*, 69–82. [CrossRef]
57. Sasse, M.A.; Brostoff, S.; Weirich, D. Transforming the 'weakest link'—A human/computer interaction approach to usable and effective security. *BT Technol. J.* **2001**, *19*, 122–131. [CrossRef]

58. Neupane, S.; Fernandez, I.A.; Mittal, S.; Rahimi, S. Impacts and risk of generative AI technology on cyber defense. *arXiv* **2023**, arXiv:2306.13033. [CrossRef]
59. AlEroud, A.; Karabatis, G. Bypassing detection of URL-based phishing attacks using generative adversarial deep neural networks. In Proceedings of the Sixth International Workshop on Security and Privacy Analytics, New Orleans, LA, USA, 18 March 2020; pp. 53–60.
60. Apruzzese, G.; Conti, M.; Yuan, Y. SpacePhish: The evasion-space of adversarial attacks against phishing website detectors using machine learning. In Proceedings of the 38th Annual Computer Security Applications Conference, Austin, TX, USA, 5–9 December 2022; pp. 171–185.
61. Yigit, Y.; Buchanan, W.J.; Tehrani, M.G.; Maglaras, L. Review of generative ai methods in cybersecurity. *arXiv* **2024**, arXiv:2403.08701. [CrossRef]
62. 2024 UK Cyber Security Breaches Survey. Available online: https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2024/cyber-security-breaches-survey-2024 (accessed on 1 April 2025).
63. Gambin, A.F.; Yazidi, A.; Vasilakos, A.; Haugerud, H.; Djenouri, Y. Deepfakes: Current and future trends. *Artif. Intell. Rev.* **2024**, *57*, 64. [CrossRef]
64. Bray, S.D.; Johnson, S.D.; Kleinberg, B. Testing human ability to detect 'deepfake'images of human faces. *J. Cybersecur.* **2023**, *9*, tyad011. [CrossRef]
65. Kaur, A.; Noori Hoshyar, A.; Saikrishna, V.; Firmin, S.; Xia, F. Deepfake video detection: Challenges and opportunities. *Artif. Intell. Rev.* **2024**, *57*, 1–47. [CrossRef]
66. Doan, T.P.; Nguyen-Vu, L.; Jung, S.; Hong, K. Bts-e: Audio deepfake detection using breathing-talking-silence encoder. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; IEEE: Piscataway, NJ, USA, 2023; pp. 1–5.
67. Seymour, J.; Tully, P. Generative models for spear phishing posts on social media. *arXiv* **2018**, arXiv:1802.05196. [CrossRef]
68. Greshake, K.; Abdelnabi, S.; Mishra, S.; Endres, C.; Holz, T.; Fritz, M. Not what you've signed up for: Compromising real-world llm-integrated applications with indirect prompt injection. In Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security, Copenhagen, Denmark, 30 November 2023; pp. 79–90.
69. Zou, A.; Wang, Z.; Carlini, N.; Nasr, M.; Kolter, J.Z.; Fredrikson, M. Universal and transferable adversarial attacks on aligned language models. *arXiv* **2023**, arXiv:2307.15043. [CrossRef]
70. Qi, Q.; Luo, Y.; Xu, Y.; Guo, W.; Fang, Y. SpearBot: Leveraging large language models in a generative-critique framework for spear-phishing email generation. *Inf. Fusion* **2025**, *122*, 103176. [CrossRef]
71. 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**, *11*, 80218–80245. [CrossRef]
72. Webb, T.; Holyoak, K.J.; Lu, H. Emergent analogical reasoning in large language models. *Nat. Hum. Behav.* **2023**, *7*, 1526–1541. [CrossRef]
73. Taddeo, M.; McCutcheon, T.; Floridi, L. Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nat. Mach. Intell.* **2019**, *1*, 557–560. [CrossRef]
74. Anil, R.; Dai, A.M.; Firat, O.; Johnson, M.; Lepikhin, D.; Passos, A.; Shakeri, S.; Taropa, E.; Bailey, P.; Chen, Z.; et al. Palm 2 technical report. *arXiv* **2023**, arXiv:2305.10403. [CrossRef]
75. Reed, S.; Zolna, K.; Parisotto, E.; Colmenarejo, S.G.; Novikov, A.; Barth-Maron, G.; Gimenez, M.; Sulsky, Y.; Kay, J.; Springenberg, J.T.; et al. A generalist agent. *arXiv* **2022**, arXiv:2205.06175. [CrossRef]
76. Asfoor, A.; Rahim, F.A.; Yussof, S. Factors influencing information security awareness of phishing attacks from bank customers' perspective: A preliminary investigation. In Proceedings of the Recent Trends in Data Science and Soft Computing: Proceedings of the 3rd International Conference of Reliable Information and Communication Technology (IRICT 2018), Kuala Lumpur, Malaysia, 23–24 July 2018; Springer: Berlin/Heidelberg, Germany, 2019; pp. 641–654.
77. Wen, Z.A.; Lin, Z.; Chen, R.; Andersen, E. What. hack: Engaging anti-phishing training through a role-playing phishing simulation game. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK, 4–9 May 2019; pp. 1–12.
78. Xiong, A.; Proctor, R.W.; Yang, W.; Li, N. Embedding training within warnings improves skills of identifying phishing webpages. *Hum. Factors* **2019**, *61*, 577–595. [CrossRef] [PubMed]
79. Sturman, D.; Auton, J.C.; Morrison, B.W. Security awareness, decision style, knowledge, and phishing email detection: Moderated mediation analyses. *Comput. Secur.* **2025**, *148*, 104129. [CrossRef]
80. Dixon, M.; Gamagedara Arachchilage, N.A.; Nicholson, J. Engaging users with educational games: The case of phishing. In Proceedings of the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, Glasgow, Scotland, UK, 4–9 May 2019; pp. 1–6.

81. Ndibwile, J.D.; Kadobayashi, Y.; Fall, D. UnPhishMe: Phishing attack detection by deceptive login simulation through an Android mobile app. In Proceedings of the 2017 12th Asia Joint Conference on Information Security (AsiaJCIS), Seoul, Republic of Korea, 10–11 August 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 38–47.
82. Sheng, S.; Magnien, B.; Kumaraguru, P.; Acquisti, A.; Cranor, L.F.; Hong, J.; Nunge, E. Anti-phishing phil: The design and evaluation of a game that teaches people not to fall for phish. In Proceedings of the 3rd Symposium on Usable Privacy and Security, Pittsburgh, PA, USA, 18–20 July 2007; pp. 88–99.
83. Kumaraguru, P.; Sheng, S.; Acquisti, A.; Cranor, L.F.; Hong, J. Teaching Johnny not to fall for phish. *ACM Trans. Internet Technol. (TOIT)* **2010**, *10*, 1–31. [CrossRef]
84. Lim, I.k.; Park, Y.G.; Lee, J.K. Design of security training system for individual users. *Wirel. Pers. Commun.* **2016**, *90*, 1105–1120. [CrossRef]
85. Nurse, J.R. Cybercrime and you: How criminals attack and the human factors that they seek to exploit. *arXiv* **2018**, arXiv:1811.06624. [CrossRef]
86. Avery, J.; Almeshekah, M.; Spafford, E. Offensive deception in computing. In Proceedings of the International Conference on Cyber Warfare and Security, Dayton, OH, USA, 2–3 March 2017; Academic Conferences International Limited: Reading, UK, 2017; p. 23.
87. Gangire, Y.; Da Veiga, A.; Herselman, M. A conceptual model of information security compliant behaviour based on the self-determination theory. In Proceedings of the 2019 Conference on Information Communications Technology and Society (ICTAS), Durban, South Africa, 6–8 March 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
88. McElwee, S.; Murphy, G.; Shelton, P. Influencing outcomes and behaviors in simulated phishing exercises. In Proceedings of the SoutheastCon 2018, St. Petersburg, FL, USA, 19–22 April 2018; IEEE: Piscataway, NJ, USA, 2018; pp. 1–6.
89. Noureddine, M.A.; Marturano, A.; Keefe, K.; Bashir, M.; Sanders, W.H. Accounting for the human user in predictive security models. In Proceedings of the 2017 IEEE 22nd Pacific Rim International Symposium on Dependable Computing (PRDC), Christchurch, New Zealand, 22–25 January 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 329–338.
90. Metalidou, E.; Marinagi, C.; Trivellas, P.; Eberhagen, N.; Skourlas, C.; Giannakopoulos, G. The human factor of information security: Unintentional damage perspective. *Procedia-Soc. Behav. Sci.* **2014**, *147*, 424–428. [CrossRef]
91. Steves, M.P.; Greene, K.K.; Theofanos, M.F. A phish scale: Rating human phishing message detection difficulty. In Proceedings of the Workshop on Usable Security (USEC) 2019, San Diego, CA, USA, 24 February 2019.
92. Choong, Y.Y.; Theofanos, M. What 4500+ people can tell you–employees' attitudes toward organizational password policy do matter. In Proceedings of the Human Aspects of Information Security, Privacy, and Trust: Third International Conference, HAS 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, 2–7 August 2015; Proceedings 3; Springer: Berlin/Heidelberg, Germany, 2015; pp. 299–310.
93. Lévesque, F.L.; Chiasson, S.; Somayaji, A.; Fernandez, J.M. Technological and human factors of malware attacks: A computer security clinical trial approach. *ACM Trans. Priv. Secur. (TOPS)* **2018**, *21*, 1–30. [CrossRef]
94. Nsiempba, J.J.; Lévesque, F.L.; de Marcellis-Warin, N.; Fernandez, J.M. An empirical analysis of risk aversion in malware infections. In Proceedings of the Risks and Security of Internet and Systems: 12th International Conference, CRiSIS 2017, Dinard, France, 19–21 September 2017; Revised Selected Papers 12; Springer: Berlin/Heidelberg, Germany, 2018; pp. 260–267.
95. Kavvadias, A.; Kotsilieris, T. Understanding the role of demographic and psychological factors in users' susceptibility to phishing emails: A review. *Appl. Sci.* **2025**, *15*, 2236. [CrossRef]
96. Williams, N.; Li, S. Simulating human detection of phishing websites: An investigation into the applicability of the ACT-R cognitive behaviour architecture model. In Proceedings of the 2017 3rd IEEE International Conference on Cybernetics (CYBCONF), Exeter, UK, 21–23 June 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–8.
97. Zhao, R.; John, S.; Karas, S.; Bussell, C.; Roberts, J.; Six, D.; Gavett, B.; Yue, C. The highly insidious extreme phishing attacks. In Proceedings of the 2016 25th International Conference on Computer Communication and Networks (ICCCN), Waikoloa, HI, USA, 1–4 August 2016; IEEE: Piscataway, NJ, USA, 2016; pp. 1–10.
98. Dodson, B.; Sengupta, D.; Boneh, D.; Lam, M.S. Secure, consumer-friendly web authentication and payments with a phone. In Proceedings of the Mobile Computing, Applications, and Services: Second International ICST Conference, MobiCASE 2010, Santa Clara, CA, USA, 25–28 October 2010; Revised Selected Papers 2; Springer: Berlin/Heidelberg, Germany, 2012; pp. 17–38.
99. Jindal, S.; Misra, M. Multi-factor authentication scheme using mobile app and camera. In Proceedings of the International Conference on Advanced Communication and Computational Technology (ICACCT), Kurukshetra, India, 6–7 December 2019; Springer: Berlin/Heidelberg, Germany, 2019; pp. 787–813.
100. Lu, Y.; Li, L.; Peng, H.; Yang, Y. A novel smart card based user authentication and key agreement scheme for heterogeneous wireless sensor networks. *Wirel. Pers. Commun.* **2017**, *96*, 813–832. [CrossRef]
101. Sheng, S.; Wardman, B.; Warner, G.; Cranor, L.; Hong, J.; Zhang, C. An empirical analysis of phishing blacklists. In Proceedings of the Sixth Conference on Email and Anti-Spam (CEAS), Mountain View, CA, USA, 16–17 July 2009; Carnegie Mellon University: Pittsburgh, PA, USA, 2009.

102. Phishtank Webiste. Available online: https://www.phishtank.com/ (accessed on 1 May 2025).
103. Alhuzali, A.; Alloqmani, A.; Aljabri, M.; Alharbi, F. In-Depth Analysis of Phishing Email Detection: Evaluating the Performance of Machine Learning and Deep Learning Models Across Multiple Datasets. *Appl. Sci.* **2025**, *15*, 3396. [CrossRef]
104. Mahmud, T.; Prince, M.A.H.; Ali, M.H.; Hossain, M.S.; Andersson, K. Enhancing cybersecurity: Hybrid deep learning approaches to smishing attack detection. *Systems* **2024**, *12*, 490. [CrossRef]
105. Alam, M.N.; Sarma, D.; Lima, F.F.; Saha, I.; Ulfath, R.E.; Hossain, S. Phishing attacks detection using machine learning approach. In Proceedings of the 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 20–22 August 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1173–1179.
106. Rawal, S.; Rawal, B.; Shaheen, A.; Malik, S. Phishing detection in e-mails using machine learning. *Int. J. Appl. Inf. Syst.* **2017**, *12*, 21–24. [CrossRef]
107. Nagy, N.; Aljabri, M.; Shaahid, A.; Ahmed, A.A.; Alnasser, F.; Almakramy, L.; Alhadab, M.; Alfaddagh, S. Phishing urls detection using sequential and parallel ml techniques: Comparative analysis. *Sensors* **2023**, *23*, 3467. [CrossRef]
108. Brissett, A.; Wall, J. Machine learning and watermarking for accurate detection of AI generated phishing emails. *Electronics* **2025**, *14*, 2611. [CrossRef]
109. Eze, C.S.; Shamir, L. Analysis and prevention of AI-based phishing email attacks. *Electronics* **2024**, *13*, 1839. [CrossRef]
110. Bethany, M.; Galiopoulos, A.; Bethany, E.; Karkevandi, M.B.; Vishwamitra, N.; Najafirad, P. Large language model lateral spear phishing: A comparative study in large-scale organizational settings. *arXiv* **2024**, arXiv:2401.09727. [CrossRef]
111. Hilario, E.; Azam, S.; Sundaram, J.; Imran Mohammed, K.; Shanmugam, B. Generative AI for pentesting: The good, the bad, the ugly. *Int. J. Inf. Secur.* **2024**, *23*, 2075–2097. [CrossRef]
112. Kucharavy, A.; Schillaci, Z.; Maréchal, L.; Würsch, M.; Dolamic, L.; Sabonnadiere, R.; David, D.P.; Mermoud, A.; Lenders, V. Fundamentals of generative large language models and perspectives in cyber-defense. *arXiv* **2023**, arXiv:2303.12132. [CrossRef]
113. Sai, S.; Yashvardhan, U.; Chamola, V.; Sikdar, B. Generative ai for cyber security: Analyzing the potential of chatgpt, dall-e and other models for enhancing the security space. *IEEE Access* **2024**, *12*, 53497–53516. [CrossRef]
114. Mun, H.; Park, J.; Kim, Y.; Kim, B.; Kim, J. PhiShield: An AI-Based Personalized Anti-Spam Solution with Third-Party Integration. *Electronics* **2025**, *14*, 1581. [CrossRef]
115. Marin, I.A.; Burda, P.; Zannone, N.; Allodi, L. The influence of human factors on the intention to report phishing emails. In Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany, 23–28 April 2023; pp. 1–18.
116. Jari, M. An overview of phishing victimization: Human factors, training and the role of emotions. *arXiv* **2022**, arXiv:2209.11197. [CrossRef]
117. Greitzer, F.L.; Li, W.; Laskey, K.B.; Lee, J.; Purl, J. Experimental investigation of technical and human factors related to phishing susceptibility. *ACM Trans. Soc. Comput.* **2021**, *4*, 1–48. [CrossRef]
118. Jari, M. A comprehensive survey of phishing attacks and defences: Human factors, training and the role of emotions. *Int. J. Netw. Secur. Its Appl.* **2022**, *14*, 5. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.