Advanced Research on Information Systems Security

ISSN: 2795-4609 | ISSN: 2795-4560
Print & Online

AI-Driven Threats in Social Learning Environments: A Multivocal Literature Review

Pedro de Almeida Perdigão[a,b]*, Nuno Mateus-Coelho [a,c], José Brás [a,b]

[a]*Universidade Lusófona, Campo Grande, 376, 1749-024 Lisbon, Portugal*
[b]*CGI Innovation Hub Lisbon, 999022 Lisboa, Portugal*
[c]*CTS – Centro de Tecnologia e Sistemas – UNINOVA, Lisbon, Portugal*

[a,b]*Email: palmeidaperdigao@gmail.com*
[a,c]*Email: nuno.coelho@ulusofona.pt*

Abstract

In recent years, artificial intelligence (AI) has become important in improving educational processes by facilitating personalized learning and enhancing collaborative platforms. However, the same technologies that offer these advantages can also enable sophisticated cyber threats. This multivocal literature review (MLR) explores four major areas of concern in social learning environments: (1) phishing and social engineering, (2) AI-generated misinformation, (3) deepfake media, and (4) AI-driven detection systems. Gathering insights from recent academic articles, industry reports, and news/blog analyses, the study demonstrates AI's dual function as both a channel for educational innovation and a tool for malicious exploitation. Findings indicate that AI-powered attacks not only erode trust and academic integrity but also target the inherent vulnerability of collaborative platforms, including Massive Open Online Courses (MOOCs). Additionally, while academic literature focuses on theoretical solutions such as explainable AI (XAI) and advanced machine learning detection, gray literature highlights practical challenges like regulatory gaps, limited funding, and insufficient user training. Blockchain-based audit trails and robust user-awareness campaigns also emerge as critical strategies for enhancing security. This review highlights the importance of interdisciplinary collaboration among policymakers, researchers, educators, and technology developers to ensure that AI's benefits are not dominated by its misuse. By adopting adaptive security

policies, fostering digital literacy, and integrating transparent detection tools, stakeholders can strengthen the resilience of social learning environments against rapidly evolving AI-driven threats.

Keywords: *artificial intelligence (AI); social engineering; phishing; deepfake; misinformation; cybersecurity, education, AI-driven detection systems.*

Citation: P. de Almeida Perdigão, N. Mateus Coelho, and J. Cascais Brás, "AI-Driven Threats in Social Learning Environments: A Multivocal Literature Review", ARIS2-Journal, vol. 5, no. 1, pp. 4–37, May 2025.

DOI: https://doi.org/10.56394/aris2.v5i1.60

---

* Corresponding author. Email address: palmeidaperdigao@gmail.com

1. Introduction

Artificial intelligence (AI) has rapidly become a transformative force in modern society, reshaping various fields such as education, healthcare, and business. Within educational contexts, AI integration has facilitated innovative teaching approaches, improved personalization, and fostered collaborative learning experiences. In particular, personalization enables AI systems to analyze user data to tailor instructional content, enhancing both engagement and effectiveness [1,2].

However, these same technological advances have also introduced unprecedented security challenges. The dual nature of AI is a catalyst for progress and a potential vehicle for harm, demands a more meticulous examination of its role in social learning environments. Malicious actors now leverage AI to orchestrate sophisticated cyber threats, including phishing attacks, misinformation campaigns, and deepfake media, all of which compromise the trust, security, and integrity of educational ecosystems [3,4,34].

Massive Open Online Courses (MOOCs) and other social learning platforms are especially vulnerable to such attacks. Phishing and social engineering tactics, for instance, exploit the natural trust among participants by utilizing AI to generate highly convincing fraudulent messages. AI-driven misinformation proliferating through collaborative environments damages shared knowledge and erodes educational credibility [5,6,33].

As AI-powered tools become increasingly accessible, their potential misuse in these contexts poses significant concerns. Advanced language models can produce deceptive reviews or malicious feedback on discussion boards, while deepfakes can manipulate video lectures or collaborative discussions, fostering confusion and eroding trust. Addressing these issues requires not only the adoption of AI-enabled defenses but also robust digital literacy initiatives to empower users to recognize and combat such threats [7,8,78].

This study employs a multivocal literature review (MLR) methodology to investigate the complex nature of AI-driven threats within social learning ecosystems. By examining a range of academic articles, gray literature, and related sources, this research aims to illustrate the complex relationship between technological innovation and

the evolving landscape of cybersecurity challenges in education [9,69,70].

2. Background

The digital transformation of education, driven by the growing adoption of artificial intelligence (AI), has become both a great opportunity and a significant challenge [85]. While AI empowers personalized learning and promotes collaborative engagement, it also magnifies the potential for cyber threats [62]. The education sector has emerged as a prime target for malicious entities seeking to exploit vulnerabilities within social learning environments [10,11,12,13,19,83].

In 2024, reported cyberattacks on educational institutions in the United States rose by 37% compared with the previous year, reflecting systematic gaps in cybersecurity protocols [14]. Across the globe, the education/research sector has become the most frequently targeted industry, averaging 3,828 weekly attacks per organization. The government/military and healthcare sectors followed closely, with 2,553 and 2,434 weekly attacks, respectively, as demonstrated in Figure 1 [15]. These statistics highlight the urgency of implementing robust measures to protect sensitive information and maintain the credibility of educational services.

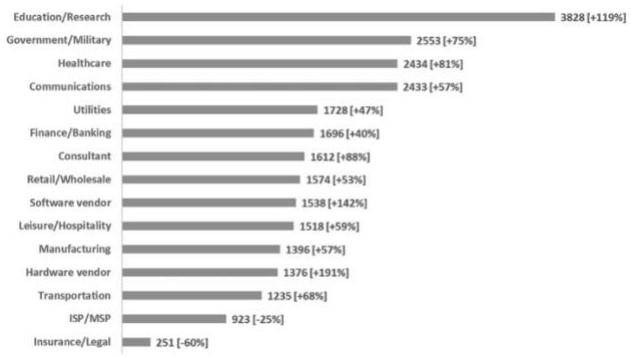

Figure 1: Global avg. weekly cyber-attacks per industry (Q3 2024 compared to Q3 2023) [15]

Figure 2 illustrates recent statistics highlighting the alarming scale of these threats, showing that institutions worldwide faced an average of 1,876 attacks per week in Q3 2024: an alarming 75% increase from the same period in 2023 [15].

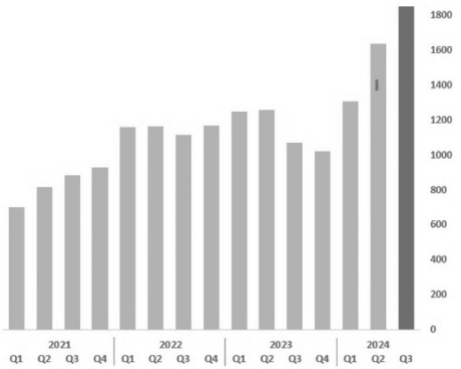

Figure 2: Avg. weekly cyber attacks per organization (Global 2021-2024) [15]

The financial impact of such breaches is equally concerning. The average cost of a data breach reached 4.88$ million in 2024, representing the highest average on record [16,17]. Human error remains a predominant factor, contributing to 88% of security breaches [16]. Once compromised, organizations typically require 194 days to detect a breach, followed by an extended lifecycle of approximately 292 days from identification to full containment of the threat [16,17]. Additionally, in 2024, nearly two-thirds of educational institutions experienced ransomware incidents, with average ransom payments reaching into the millions [18].

The education/research sector due to the current scenario has escalated from "moderate" to "high" the cyber risk rating over the past two years, indicating a troubling trend that demands urgent attention from schools, universities, and nonprofit organizations [10,12,19]. Educational institutions are particularly vulnerable to these threats due to their reliance on collaborative platforms, such as Massive Open Online Courses (MOOCs) and other social learning networks. These platforms often lack robust security infrastructure [62] to protect and prevent against sophisticated AI-driven attacks [5]. Phishing schemes, powered by generative AI, exploit trust among platform users to steal credentials or distribute malware [72,83], while deepfake technology, which is capable of producing highly convincing but fraudulent media, is increasingly being used for harassment and misinformation campaigns, weakening the integrity of educational system [4,6].

Such vulnerabilities emphasize the need for proactive measures to protect educational environments. Advanced AI-driven detection systems, combined with increased user awareness and digital literacy, can help mitigate risks associated with cyber threats [7,8,20]. Addressing these multifaceted challenges demands an integrated approach that merges technological innovation, stronger cybersecurity policies, and ongoing adaptations to an evolving threat landscape [9,21,83].

2.1. Phishing and social engineering attacks

Phishing and social engineering attacks remain among the most persistent cybersecurity challenges for both industry and academia. These techniques exploit human vulnerabilities such as trust, curiosity, or authority bias to persuade victims into exposing personal data or installing malicious software [22,84,87]. Although phishing already existed before recent advancements in artificial intelligence (AI), the growth in AI-driven generative models and data analytics has substantially increased attackers' ability to personalize and automate malicious content [23,83].

Recent studies show that AI can automate the creation of contextually relevant phishing emails or text messages, often by extracting information from social media or institutional websites [6]. Phishing attacks in Massive Open Online Courses (MOOCs) and Learning Management Systems (LMS) are particularly concerning, as attackers may impersonate instructors or administrative staff to acquire login credentials or request unauthorized payments. Time-sensitive announcements, such as assignment deadlines or exam schedules, provide ideal hooks, compelling hurried students or staff to click unfamiliar links. A recent article highlights the frequency of impersonation attacks in educational institutions, emphasizing the need for reinforced awareness [24].

A growing volume of the literature recommends machine learning (ML) and ensemble detection frameworks, which analyze sender reputation, text semantics, and user behavior. For instance, a study on phishing website detection using deep learning models explores advanced detection mechanisms to identify phishing websites effectively [25]. Alongside technical measures, security awareness training remains crucial [22,26,27,84,87]. Phishing simulations demonstrate that well-structured education programs can significantly reduce click-through rates on fraudulent links. A case study on safeguarding higher education institutions from phishing highlights the effectiveness of staff-awareness initiatives in mitigating such risk [26}.

Implementing a multi-layered approach that combines technical controls, user education, and institutional policies is essential to mitigate the persistent threat of phishing and social engineering in educational environments. By combining these strategies, institutions can reinforce the resilience of their online learning platforms and maintain the integrity of the educational experience.

- **Technical controls:** Robust technical measures form the first line of defense. Advanced email filtering, intrusion detection systems, and strong authentication protocols are fundamental anti-phishing strategies. Regular security audits and systematic software updates further bolster these defenses [26,27,87].
- **User education:** Educating staff and students about recognizing and responding to phishing attempts is crucial. Regular training initiatives such as including phishing simulations and awareness programs, can empower individuals to identify suspicious activities and reduce the likelihood of successful attacks [26,27,83,84,87].
- **Institutional policies:** Clear cybersecurity policies and procedures ensure a coordinated response to threats. Such policies often define acceptable use, establish incident-response protocols, and outline procedures for reporting suspicious activities. A well-articulated policy framework fosters a culture of security awareness and accountability [21,28,83].

By integrating these strategies, educational institutions can create a strong defense against phishing and social engineering attacks, defending their digital learning environments and maintaining the integrity of their educational platforms.

2.2. AI-Generated misinformation and learning effects

The rise of transformer-based language models has dramatically simplified the production of misinformation at scale [29]. These models can generate text that appears both coherent and contextually relevant, blurring the line between legitimate information and deceptive content [30,79]. Researchers highlight that misinformation can now be mass-produced, customized to specific topics or demographics, and automatically disseminated across online platforms [6,31,79].

In educational contexts, misinformation disrupts peer-to-peer learning and undermines academic integrity. Students frequently rely on discussion forums, social media groups, and wikis to share resources and clarify doubts and thoughts [32]. When AI-generated misinformation is introduced, it can spread quickly before lecturers or peers identify inaccuracies [33,34]. Prolonged exposure to misleading material inhibits knowledge acquisition and can erode trust in digital resources [35].

Current detection methods typically rely on natural language processing (NLP) and machine learning to flag doubtful claims or verify factual statements. For instance, stylometric analysis can detect anomalies in writing style, while graph-based fact-checking compares statements against reliable knowledge databases [36,37,38]. However, these systems face challenges: adversaries frequently adapt content to evade existing detection patterns [39,76], therefore experts highlight the need for digital literacy initiatives teaching students and educators to cross-check sources and remain skeptical of viral content [40,41].

From a learning perspective, misinformation disrupts knowledge construction by introducing errors that may be internalized or repeatedly circulated among peers. Once false information is integrated into learners understanding, correcting these inaccuracies often demands substantial cognitive effort and targeted instructional intervention. Even brief exposures to misinformation can create "illusions of knowledge," leading to overconfidence in incorrect information [42]. Prolonged misperceptions not only jeopardize academic performance and critical thinking development but also erode trust in legitimate educational materials [43].

Research has shown that once encoded, misinformation can continue to influence reasoning, even when a correction is remembered. This phenomenon, known as the continued influence effect, highlights the challenges in correcting false information in memory [44]. Additionally, studies have found that student's confidence in their knowledge does not always correlate with accuracy. Metacognitive experiences and subjective feelings can lead to overconfidence in incorrect information, which can hinder learning and critical thinking development [42].

These findings underscore the importance of developing effective strategies to mitigate the impact of misinformation on learning and to maintain trust in educational resources.

2.3. Deepfake media

The rise of deepfake media, driven by generative AI and large models, has introduced significant challenges to information integrity within society. Deepfakes, which blend "deep learning" and "fake" media, are highly realistic synthetic creations that can portray people saying or doing things that never happened. As AI advances, the realism of these creations has increased, making it increasingly difficult to differentiate falsified content from genuine information. This capability to convincingly replicate voices, faces, and behaviors raises pressing concerns about cyberbullying, harassment, and privacy violations in educational environments [6,46].

Deepfake-based impersonations can disrupt virtual lectures, circulate false announcements in institutional communications, or fabricate incriminating media featuring students or staff. Such incidents often lead to harassment, weaken user trust, and damage organizational reputations. The emotional impact on victims, especially those targeted in harassment campaigns, can be profound, heightening anxiety, causing reputational harm, or prompting social isolation [47,78,83].

Much of the research in media forensics emphasizes detection, seeking to identify pixel-level irregularities, mismatched lighting or facial movements, and audio-to-video ("lip-sync") inconsistencies [50,51,52,53]. Convolutional neural networks (CNNs) and other deep learning methods are frequently employed to spot these indicators in controlled environments [51]. However, real-world detection remains problematic as deepfake

generation methods evolve, they continuously refine the authenticity of synthetic media, complicating even state-of-the-art detection tools. Audio-to-video synchronization in particular has raised the quality of deepfakes, requiring more advanced strategies for identifying forgeries [54,61,95].

In response, educational institutions are advised to adopt clear media authentication guidelines and offer training on recognizing deepfake threats. One study, "Perception vs. Reality: Understanding and Evaluating the Impact of Synthetic Image Deepfakes over College Students" [56], highlights the critical role of awareness initiatives, including the strategic use of synthetic media for training, to help individuals more effectively identify such content [56, 57, 58]. Additionally, schools must provide comprehensive support to individuals portrayed in deepfake-related incidents. While disciplinary measures often focus on punishing offenders, some reports indicate a lack of adequate victim support structures. Ensuring counseling, mental health services, and other resources is crucial for mitigating the psychological harm associated with deepfake harassment [47,59,60,61].

By implementing a holistic strategy that unites policy development, awareness programs, and robust victim support, educational institutions can significantly reduce the risks posed by deepfake technology and safeguard the integrity of their learning communities [86].

2.4. AI-Driven detection systems

AI-driven detection systems are fundamental in modern cybersecurity, leveraging machine learning and deep learning to identify anomalies, detect phishing attempts, and flag manipulated media. These systems analyze user behaviors, such as login frequency and content creation rates, to identify suspicious patterns indicative of automated attacks or fraudulent activity [8,93].

In the educational sector, implementing AI-driven detection systems presents unique challenges. Learning platforms vary widely in design, data availability, and resource constraints, making it difficult to deploy a "one-size-fits-all solution". A detection engine effective in one Learning Management System (LMS) may struggle in another lacking robust APIs or real-time data feeds. Additionally, ethical concerns arise regarding data privacy, user profiling, and the potential over-blocking of legitimate student content. Experts advocate for a balanced strategy: integrating detection tools that offer transparency, regularly updating detection models to counter evolving attacks, and training educators and learners to recognize threats [21,62].

Recent advancements suggest that Explainable AI (XAI) could enhance trust in automated detection by clarifying how and why certain content is flagged [21,64]. XAI techniques, such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), provide insights into AI decision-making processes, making them more transparent and trustworthy [63]. Additionally, blockchain-based audit trails offer secure logs that can facilitate rapid forensic analysis when breaches occur. Continuous collaboration among AI researchers, educators, and policymakers is crucial to ensure that detection systems keep pace with adversarial evolution in social learning environments [64]. Additionally, the integration of AI in cybersecurity education is essential. A systematic literature review highlights the importance of teaching AI and cybersecurity together, emphasizing the need for comprehensive education to prepare future professionals for the evolving digital

landscape [5].

Concluding, while AI-driven detection systems offers potential in enhancing cybersecurity within educational environments, addressing the associated challenges and ethical considerations is critical [92]. Implementing transparent, adaptable, and collaborative approaches will be key to effectively leveraging AI in this domain.

3. Research methodology

3.1. Methods

The review protocol specifies the research question being addressed and the methods that are used to perform the review. To find the maximum number of studies related to the research question, a search strategy was used to detect as much of the relevant literature as possible using multiple keywords and datasets.

The research was carried out by using search strings to search for information on the main topic, "Generative AI", associating it with other interrelated keywords like social learning, education, cybersecurity and cyber threats (e.g. phishing, deepfakes and misinformation). Regarding academic data sources, the publications domain was identified by searching several electronic bibliographic databases, listed below, to build the datasets. The papers were collected based on their title, keywords, abstract, submission for review and publication in academic journals. Google Search (www.google.com (accessed 02 September 2024)) was also chosen to search for gray literature.

3.2. Data source and searches

The PRISMA (preferred reporting items for systematic reviews and meta-analyses) guidelines were followed in the conduct and reporting of this systematic review.

The articles were collected between September 2024 and January 2025; and restrictions were applied regarding language (only English) and dates between 1998 and 2025 with a strong preference of articles no older than 6 years old. The following keywords were applied to the search:

- "Generative AI" AND "misinformation" AND "learning".
- "Generative AI" AND "phishing" AND "education".
- "Generative AI" AND "deepfakes".
- "Generative AI" AND "social learning".
- "Generative AI" AND "education" AND "security"

Bibliographies from relevant publications were checked to identify relevant articles. The following databases for eligible studies:

- IEEE Xplore Digital Library (https://ieeexplore.ieee.org/Xplore/home.jsp (accessed on 02 October 2024)).
- ACM (https://dl.acm.org (accessed on 02 October 2024)).
- SpringerLink (https://link.springer.com (accessed on 02 October 2024)).
- SpringerNature (https://www.springernature.com/gp (accessed on 02 October 2024)).

- MDPI (https://www.mdpi.com (accessed on 02 October 2024).
- Google Search (https://www.google.com (accessed on 02 October 2024)).

Google Search was considered a limitation in terms of the replicability of the searches performed at a given time but, according to some authors [65], website search methods may differ, and it is more important to have a considered rationale for the process, taking the goals and objectives of each review into account, rather than specifying a single method. The planning and execution of the research, as well as the screening of results and the structure of its management, must be properly organized for this type of approach [65]. They recommend performing a gray literature search using at least one traditional search engine like Google with the first 12 pages (instead of the first 5 pages) and an accurate search of academic databases that are more closely aligned with the topic under analysis, to ensure that all the relevant literature is considered and that the conclusions are more comprehensive [66,67].

3.3. Eligibility criteria

For the qualitative analysis, it was included articles related to main keywords (AI, detection systems, social learning, education and cyber threats), present in the title, abstract, key contents or subject relevance. They were found in journals, conference papers, blogs or gray literature (limited to the first 12 pages of Google Search).

3.4. Study selection

In the initial search stage (first filtration, shown in Figure 1), the filtering criteria - inclusion and exclusion criteria filters (all fields; all documents and full text, abstract, reviewed publications in journals, academic journals and gray literature) - were used together with the search string. This step is illustrated in Table 1, as part of the full MLR protocol to find the final sample for the elaboration of the article, which produces a list of the articles found, together with the filters used. All publications that met the inclusion criteria were selected and analyzed.

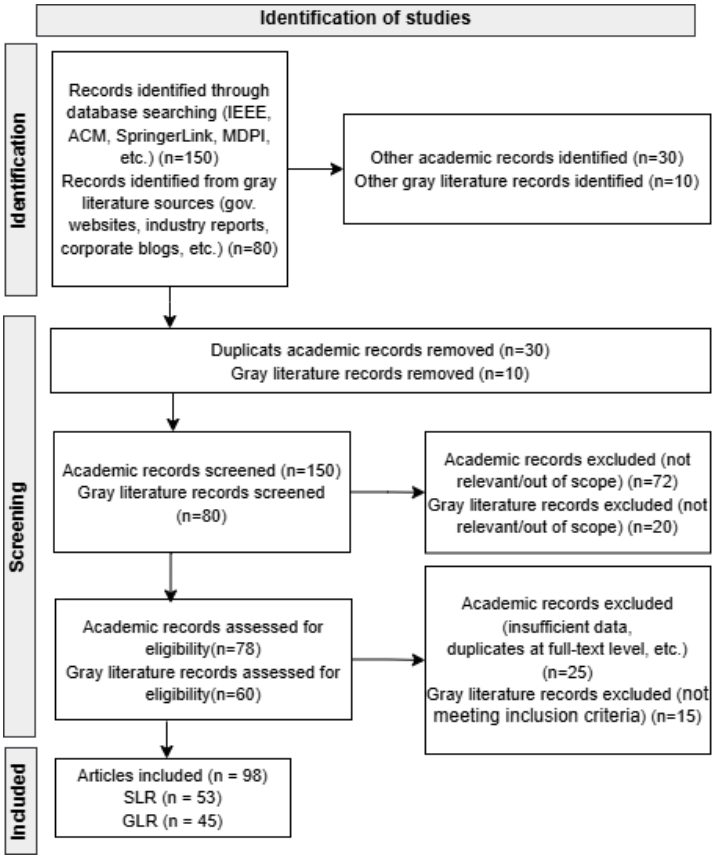

Figure 1: PRISMA 2020 flow diagram (adapted from [68])

Table 1: Inclusion and exclusion criteria

Inclusion criteria	Exclusion criteria
Related to the main keywords	Paper not in English
Documents in English	Documents with publication date earlier than 2018
Key topics AI, social learning, education and cyber threats	Not related to the key topics stated in this article.
Article, journals, conference papers, blogs or gray literature	Papers by unidentified authors
Limit results to first 12 pages of Google search	No publication date
Title, abstract, key contents or subject relevance	

In the case of the Google search engine, we consider it to be a valid source of gray literature, governmental and institutional reports. Although Google Search has its limitations and should not be used as the only source for systematic reviews, it was used here as it can be suitable for the purposes of qualitative systematic reviews. For the initial results, only the first twelve pages of the results were counted, which were then used for review and

selection [67].

The study has the following research question: What are the primary AI-driven threats in social learning environments, and how effective are current detection systems in mitigating these risks?

The overview of the review process can be found in Figure 1, which provides a visual representation of the study selection process that was applied. This diagram represents the different selection steps used in the systematization of the selection process.

An inclusion and exclusion criteria were adopted to identify the relevant literature for this study. The screening criteria for including or excluding articles for this research are summarized and illustrated in Table 1.

A software package (Zotero) was used to facilitate the task of searching and collecting the literature. This ensures that unique results are obtained, as the software detects and eliminates duplicate entries, therefore solving the problem of consistency in the returned and collected results and organizing it into different sets according to query strings and the academic or gray literature categories. Finally, it facilitates the work of retrieving the results of the distinct ID sets (academic and gray literature) that are easily merged in the study process.

4. Multivocal Literature Review

The multivocal literature review (MLR) [69] is like the systematic literature review (SLR) [70] and aims to incorporate the so-called "gray literature" to supplement the published (formal) literature. MLRs are SLRs which include both scholarly writing (also known as academic writing or formal writing) and the (informal) gray literature (GL) which is not considered in the SLR. GL is a multisource of information, which may exist in the form of blogs, videos, webpages and white papers that are produced outside academic forums and are not subject to any quality control mechanism (e.g., the peer review process) prior to publication.

By including information that normally would not be considered due to its "gray" nature [69], MLRs are important for the completeness of the research. An MLR in each subject field is essentially a combination of the sources that would be studied in an SLR and a GLR in the same field. Thus, an MLR is, in principle, expected to provide a more complete picture of the evidence in each field. Figure 2 represents the relationship between SLR, GLR and MLR.

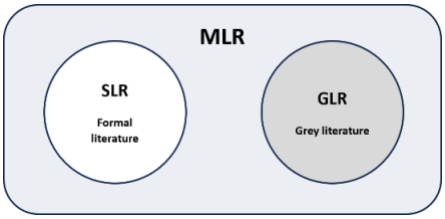

Figure 2: Relationship between SLR, GLR and MLR [69]

The aim of this research is to analyze the results of the MLR to provide a comprehensive overview of the

current state of misuse of Generative AI within four key areas of interest. The findings intent to contribute to understanding the broader implications of AI-driven threats in social learning environments. This research will focus on:

- How generative AI facilitates the creation of realistic phishing emails and social engineering schemes, which can affect the overall trust within social learning networks.
- The impact of AI-generated misinformation learning outcomes, focusing on how the spread of false knowledge damages the integrity of educational content and can erode the trust within collaborative learning platforms.
- Potential misuse of AI-generated deepfake media for malicious purposes, including for e.g. (cyberbullying, defamation and harassment).
- Efficiency of AI detection systems in identifying and mitigating malicious content, such as for e.g. (phishing, misinformation, deepfakes, etc.) within learning networks.
- Ethical and policy frameworks that guide responsible adoption of AI.

Table 2 distinguishes between "white literature" and "gray literature", listing the appropriate choice of publications in each case. "Black" or other types of literature subject to exclusion are also classified, to clarify the choices made during the assessment.

Table 2: Spectrum of "white", "gray" and excluded literature (adapted from [69]).

"White" literature	"Gray" literature	"Black" literature
Published journal papers	Preprints	Ideas
Conference proceedings	e-Prints	Concepts
Books	Technical reports	Thoughts
	Lectures	
	Data sets	
	Blogs	
	Technical reports	
	White papers	
	Government documents	
	Audio-video media	

The MLR workflow is summarized in Figure 3 and has three phases. The initial phase of the research

("planning the MLR") comprises two steps:

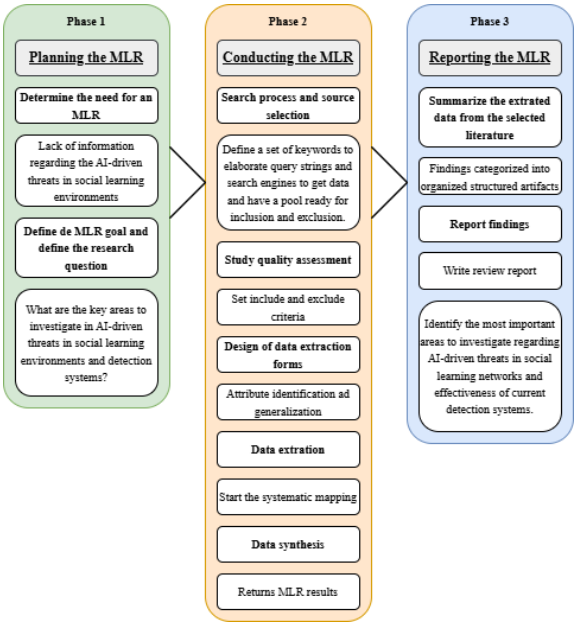

Figure 3: Multivocal literature review (MLR) phases and steps adopted in this research (adapted from [69])

- Determining the need for an MLR for the given topic.
- Defining the MLR goal and setting up the research questions.

Once the MLR is planned, we proceed to the next phase of the research, namely "conducting the MLR". This phase is divided into five stages:

- Search process and selection: identification of primary studies to address the research question, application of standard comprehensive search techniques by means of defined search strings, and definition of the selection criteria for performing the selection process.
- Study quality: assessment of sources to determine the extent to which a source is valid and free from bias.
- Design of data extraction: creation of forms to gather all the information needed to address the review question and the study quality criteria.
- Data extraction: extraction of the data items needed to answer the research questions.
- Data synthesis: synthesis of data in such a way that the question(s) can be answered.

"Reporting the MLR" is the final phase and is very similar to the SLR guidelines provided by Kitchenham and Charters [71] for planning the MLR, specifying dissemination mechanisms, formatting the main report and evaluating the report.

5. Reporting the MLR

5.1. Motivation

As a professor, my passion lies in fostering environments where students can learn, collaborate, and grow without

barriers. Teaching is more than sharing knowledge, it is about creating spaces where curiosity thrives, trust is cultivated, and students feel empowered to explore ideas. However, the integration of artificial intelligence (AI) into education, while transformative, introduces challenges that threaten the very foundation of these spaces.

The rise of AI-driven threats such as phishing attacks, misinformation, and deepfake media, poses unique risks to both students and educational institutions. These threats not only disrupt learning but also undermine trust and integrity within social learning environments. As a professor, I find it deeply concerning that the same technologies enabling personalized and adaptive learning can also be weaponized to exploit vulnerabilities in our educational systems.

This review seeks to answer a critical question: What are the primary AI-driven threats in social learning environments, and how effective are current detection systems in mitigating these risks? While existing research highlights the potential of AI to revolutionize education, there remains a pressing need to address its darker implications. Understanding the mechanisms by which these threats operate is essential for developing effective countermeasures. It is equally important to understand whether current detection systems are capable of addressing these threats without compromising the educational experience or invading user privacy.

By consolidating insights from academic literature, industry reports, and gray literature, my motivation stems from a commitment to ensuring that the integration of AI into education enhances rather than threatens the learning experience. By identifying the challenges and highlighting recommendations, this work seeks to contribute to the development of secure, trustworthy, and resilient social learning environments where students and educators can thrive.

6. Conducting the MLR

This section describes how the review was conducted, which is the second phase of the process. In this stage, the research is carried out by searching for information in selected databases using pre-defined queries and analyzing the extracted data.

7. Reporting the MLR

This section organizes the research findings to identify critical areas for future investigation regarding AI-driven threats and detection systems in social learning environments.

The objective is to explore what the scientific and gray literature reveals about the primary threats posed by AI and the gaps in current detection systems. The findings were derived from analyzing the outcomes and proposed future directions highlighted in the reviewed literature. This analysis led to three primary clusters: emerging threat vectors, AI detection strategies, and ethical and policy frameworks. These clusters emphasize the need for a more comprehensive and structured research approach to mitigate AI-driven threats and enhance the effectiveness of detection systems in social learning environments.

These findings highlight the importance of continued research to strengthen the security and integrity of

educational environments, ensuring that technological solutions remain adaptable to evolving threats while fostering trust and collaboration among educators, students, and institutions.

7.1. Emerging threat vectors

The reviewed literature reveals several AI-driven threats impacting social learning environments, highlighting the growing sophistication of attacks. These threats exploit AI to enhance deception, increase automation, and manipulate trust-based interactions in digital education settings. The following subsections examine the primary AI-driven threats, their impact, key findings, points of convergence and divergence, and recommended mitigation strategies.

7.1.1. Phishing and social engineering

AI-powered phishing and social engineering attacks have become increasingly sophisticated, leveraging machine learning to personalize attacks and bypass traditional security mechanisms.

1. **Key findings - AI-enhanced sophistication**
    - **AI-enhanced personalization:** AI enables cybercriminals to craft context-aware, highly customized phishing messages, mimicking institutional communications with authentic branding and tone [7,25,27,89].
        - **Example**: Phishing emails simulating exam schedule updates or login authentication requests.
        - Gray literature insight: Reports like those from Forbes and Wired emphasize that generative AI enables attackers to create highly convincing emails targeting students and educators by exploiting their trust in institutional systems [74,78].
    - **Real-time adaptation:** Advanced AI models dynamically adjust phishing strategies based on user behavior, increasing the effectiveness of attacks [26,83].
2. **Key findings - Targets and methods**
    - **Exploitation of urgency:** Attackers frequently use time-sensitive scenarios (e.g., assignment deadlines, system maintenance notices) to manipulate users into acting without verification [15,27,84].
    - **Educational sector vulnerabilities:** Higher education institutions, with diverse systems and varied levels of digital literacy, are identified as prime targets. These environments often lack robust security protocols, such as multi-factor authentication, aggravating vulnerabilities. Additionally, limited cybersecurity budgets and reliance on legacy systems increase the attack perimeter, making it easier for cybercriminals to exploit institutional weaknesses [10,11,26,62].
3. **Points of convergence**
    - **AI as an asset and a liability:** Both academic and gray literature emphasize that AI enhances phishing threats but also improves detection capabilities [7,23,25,92].
    - **Need for awareness and training:** User education is a widely recommended countermeasure, with training programs significantly reducing phishing success rates [26,27,87].
    - **Institutional weaknesses:** The absence of uniform security policies and inconsistencies in

email verification protocols are common vulnerabilities [15,83].

4. **Points of divergence**
    - **Detection strategies:**
        - Academic literature focuses on automated phishing detection using deep learning models, analyzing linguistic patterns and sender metadata [25,39].
        - Gray literature prioritizes practical, immediate solutions, such as phishing simulations, user training, and institutional awareness campaigns [27,72].
    - **User impact:**
        - Academic literature primarily explores technical approaches to phishing detection, focusing on algorithmic enhancements such as machine learning-based email classification, NLP-driven text analysis, and sender reputation scoring to improve security measures [25,39,63].
        - Gray literature, however, highlights the human impact of phishing, including trust erosion, psychological distress, and institutional consequences when successful attacks compromise educators' and students' personal data [83,89].
5. **Recommendations from the literature**
    - **Academic recommendations:**
        - Deploy machine learning-based phishing detection using behavioral analytics and anomaly detection [25,39].
        - Integrate Explainable AI (XAI) to enhance transparency and increase trust in automated detection systems [63].
    - **Gray literature recommendations:**
        - Conduct regular phishing awareness training and simulated attacks for students and staff [27,72].
        - Enforce multi-factor authentication (MFA) and institution-wide cybersecurity policies to mitigate risks [83].
        - Encourage partnerships between educational institutions and cybersecurity companies to stay ahead of evolving phishing approaches [10,15].

7.1.2. Misinformation amplified by AI

Misinformation amplified by AI represents a pressing challenge for social learning environments. The deployment of generative AI and large language models has drastically boosted the scale, speed, and sophistication of misinformation campaigns, complicating efforts to maintain informational integrity in educational contexts. Below, the key findings, points of convergence and divergence, and authors' recommendations are summarized.

1. **Key findings - Proliferation and sophistication of AI-generated misinformation**
    - **Ease of misinformation creation:** Generative AI tools can produce realistic and contextually accurate text, making it increasingly difficult for users to differentiate between legitimate and false information [73]. Academic literature highlights that this capability has accelerated the spread of misinformation in education, particularly on platforms like forums, wikis, and

discussion boards [6,29,31,34].

- **Example:** Tools like GPT models can generate fake articles or "expert" opinions that appear credible, amplifying their impact on uninformed audiences.

- **Rapid dissemination:** Gray literature reveals that the rapid propagation of AI-generated misinformation on social media and collaborative platforms magnifies its reach. Articles highlight how algorithms prioritize engagement over accuracy, aggravating the problem [12,19,76].

2. **Key findings - Educational impacts**
   - **Erosion of trust:** Misinformation undermines trust in educational materials and platforms. Both academic and gray literature agree that repeated exposure to false content can lead students to question the reliability of all digital resources, creating "illusions of knowledge" [29,35,76].
   - **Cognitive and emotional strain:** Exposure to conflicting information forces learners to expend cognitive effort determining credible content, leading to frustration and disengagement. Gray sources emphasize the psychological aspect, highlighting that misinformation's emotional charge can aggravate stress for students and educators equally [12,79].
3. **Points of convergence**
   - **Role of AI:** Both academic and gray literature agree that AI plays a dual role, serving as both the catalyst for misinformation and a potential solution through advanced detection systems [6,34,79,83].
   - **Necessity of digital literacy training:** Sources converge on the need for digital literacy initiatives to empower students and educators to critically evaluate online content. Practical training programs and awareness campaigns are frequently recommended [12,19,76].
4. **Points of divergence**
   - **Detection approaches:**
     - Academic sources tend to prioritize the development of technical solutions, such as machine learning models and natural language processing (NLP) algorithms, to identify and flag misleading content [6,29,30,35].
     - Gray literature leans toward manual verification, platform-based moderation, and policy interventions to mitigate misinformation risks [12,19,79].
   - **Interpretation of AI's impact:**
     - Academic sources frame AI as a neutral tool misused for misinformation [29,34].
     - Gray literature highlights corporate and governmental responsibilities, questioning platform accountability in misinformation propagation [76,79].
5. **Recommendations from the literature**
   - **Academic recommendations:**
     - **Advanced detection mechanisms:** Develop NLP-based systems that analyze textual patterns, evaluate content legitimacy against established knowledge bases, and flag anomalies in real-time [29,33,35].
     - **Explainable AI (XAI):** Employ XAI to enhance transparency in misinformation detection, fostering trust in automated systems and enabling users to understand why

content is flagged as misleading [6,34].

- **Gray literature recommendations:**
  - **Digital literacy programs:** Launch educational initiatives focused on teaching students and educators to critically evaluate sources and recognize common misinformation patterns [12,76].
  - **Policy and platform accountability:** Advocate for stricter policies on content moderation and algorithmic accountability within collaborative educational platforms [19,79]
  - **Collaborative efforts:** Encourage partnerships between educational institutions, AI researchers, and platform providers to address the systemic challenges of misinformation propagation [19,76].

7.1.3. Deepfake media

Deepfake media has emerged as a significant threat in social learning environments. Leveraging generative AI, deepfakes produce hyper-realistic synthetic content, enabling malicious actors to fabricate videos, images, and audio. These manipulations create risks to trust, privacy, and the integrity of digital educational spaces. Below is an analysis of the findings from academic and gray literature, including points of convergence and divergence, as well as recommendations from the authors.

1. **Key findings - Emergence and capabilities of deepfake media**
   - **Technological advancements:** Deepfakes, powered by sophisticated AI models, can convincingly mimic voices, faces, and behaviors, creating content that is nearly indistinguishable from reality. Academic sources point out how these advancements make it increasingly difficult to detect and prevent misuse [4,46,55].
     - **Real-world examples:**
       - Fabricated videos of lecturers or students making inappropriate statements can disrupt classes and harm reputations [59,60].
       - The use of AI-generated voice clones is also rising in higher education, with both legitimate applications (e.g., automated announcements) and risks of impersonation through phishing and fraud [49,78,99].
   - **Impact on trust:** Gray literature highlights that deepfake media weakens trust in digital communications and content authenticity. This erosion of trust is particularly concerning in educational environments, where collaboration and communication rely on integrity and transparency [59,75,78].
2. **Key findings - Educational impacts**
   - **Targeted harassment:** Deepfake technologies are used to target students and educators through fabricated media, leading to reputational damage and psychological harm. Both academic and gray sources underscore the emotional charge on victims [4,57,60,77,100,101].
   - **Institutional disruptions:** Instances of deepfakes being used to impersonate administrators or disseminate false announcements have been documented. This leads to confusion and damages

the credibility of institutional communications [46,49,57,62].

3. **Points of convergence**
    - **Dual-use technology:** Both academic and gray literature agree that while deepfake technology has legitimate applications (e.g., virtual learning environments), its misuse for malicious purposes outweighs these benefits in current contexts [4,55,78].
    - **Challenges in detection**: Sources consistently highlight the limitations of current detection methods. While AI-based tools can flag deepfake artifacts, evolving techniques often beat detection capabilities, making manual verification necessary in critical cases [23,46,55,60]. In addition, further emphasizes that adversarial AI techniques are being used to avoid detection, making it imperative for detection models to continuously evolve. Without iterative improvements, AI-powered forensic tools risk becoming obsolete against increasingly sophisticated manipulations.
4. **Points of divergence**
    - **Proposed countermeasures:**
        - Academic literature primarily explores technical approaches to deepfake detection, such as multimodal falsification detectors that analyze inconsistencies in audio-visual synchronization and blockchain-based content authentication [50,52,55].
        - Gray literature emphasizes institutional policies, victim support frameworks, and awareness campaigns [57,59].
    - **Scope of impact:**
        - Academic sources discuss the broader societal implications of deepfakes, such as their role in misinformation campaigns [4,46].
        - Gray literature narrows the focus to specific educational settings, addressing how these technologies disrupt student learning and faculty operations [57,59].
5. **Recommendations from the literature**
    - **Academic recommendations:**
        - **Develop advanced detection tools:** Invest in AI-powered detection systems, including multimodal approaches that analyze audio, video, and metadata for anomalies [50,55].
        - **Blockchain for media authentication:** Implement blockchain-based audit trails to verify the authenticity of media and detect manipulation [46,55].
    - **Gray literature recommendations:**
        - **Support victims:** Establish institutional protocols for responding to deepfake incidents, including counseling for victims and reputational repair strategies [57,60].
        - **Awareness and training:** Educate students and staff on recognizing deepfakes and understanding their implications. Workshops and simulations can prepare users to identify manipulative content [60,78].
        - **Policy development:** Advocate for strong regulations governing the creation and

distribution of synthetic media, emphasizing accountability for misuse [57,59].

7.1.4. Compounded vulnerabilities

Compounded vulnerabilities rise from the combination of multiple AI-driven threats, such as phishing, misinformation, and deepfake media, within social learning environments. These interconnected risks amplify the potential for disruption, creating challenges that demand a multi-faceted response. Below is the MLR analysis of the sub-topic, highlighting key findings, points of convergence and divergence, and recommendations from the literature.

1. **Key findings - The Interconnected nature of threats.**
    - **Phishing and misinformation:** The combination of phishing and misinformation aggravates vulnerabilities, as attackers leverage AI to create tailored, misleading messages that target user trust. Academic sources highlight how AI-enhanced phishing campaigns can spread false narratives, intensifying their impact [7,25,39].
        - **Example:** Phishing emails that incorporate fabricated "official" institutional policies or fake news about administrative changes.
    - **Deepfakes as catalysts:** Deepfake technologies often serve as amplifiers in compounded vulnerabilities. Fabricated videos or audio clips can reinforce phishing or misinformation efforts, making them more convincing and difficult to detect [4,46,50].
    - **Adaptive attacks:** The adaptability of AI-enhanced threats makes detection challenging. Academic sources emphasize that attackers often modify tactics to bypass existing defenses, leveraging multi-modal approaches such as combining video manipulation with phishing [23,46,50].
2. **Key findings - Educational impacts**
    - **Systemic weaknesses:** The compounded effects of these threats exploit systemic weaknesses in learning management systems (LMS) and institutional cybersecurity frameworks. Both academic and gray literature emphasizes that fragmented security protocols and under-resourced IT teams make educational institutions attractive targets [10,15,19,62,83,99,100,101].
    - **Limited preparedness, loss of trust, and operational disruption:** Repeated breaches and the propagation of false information erode trust in educational institutions. Gray literature reports that these attacks disrupt daily operations, causing delays, financial losses, and reputational damage. Additionally, gray literature also highlights the unpreparedness of many institutions to address the interconnected nature of these threats. Resource constraints, inadequate training, and fragmented policies contribute to the challenges [15,19,27,60,72].
3. **Points of convergence**
    - **The urgency of a holistic approach:** Both academic and gray sources highlight the need for integrated strategies that address the intersection of threats rather than isolated issues. Combining technical defenses, user education, and robust policies is consistently recommended [7,8,15,39,63,83,92].
    - **Importance of institutional resilience:** There is agreement across sources on building

resilience within institutions by enhancing their ability to detect, respond, and recover from multi-vector attacks. This includes adopting proactive measures such as incident response simulations and cross-functional collaboration [9.19,21,62,72,80].

- **Collaboration across stakeholders**: The literature agrees on the need for collaboration between AI researchers, cybersecurity experts, and educational institutions to tackle these multifaceted challenges effectively [7,21,83].

4. **Points of divergence**
   - **Technical vs. Policy Solutions:**
     - Academic literature often emphasizes technical innovations, such as machine learning-based anomaly detection and blockchain for tamper-proof audit trails [39,63,64].
     - Gray literature highlights the importance of policy changes, user awareness campaigns, and partnerships with cybersecurity firms to audit and address immediate vulnerabilities [10,15,19,80,83].
   - **Scope of analysis:**
     - Academic sources analyze the theoretical implications of combined vulnerabilities, often exploring abstract scenarios to identify potential risks and impact on global cybersecurity standards and AI ethics [8,21,25,55].
     - Gray sources provide practical examples and case studies, such as real-world breaches in schools or universities, to illustrate the tangible impact of these threats, such as the psychological impact on victims and disruption of learning processes [19,60,72,78,83].
5. **Recommendations from the literature**
   - **Academic recommendations:**
     - **Enhance detection capabilities:** Develop integrated detection systems that can identify and mitigate combined threats, such as phishing attempts incorporating deepfake elements or misinformation campaigns [8,39,46,55,63].
     - **Explainable AI and Blockchain:** Employ XAI to enhance transparency in detection processes and blockchain for secure tracking of digital assets, reducing vulnerabilities from deepfake and phishing combinations [7,55,63,64].
   - **Gray literature recommendations:**
     - **Adopt multi-layered defense strategies:** Implement a combination of technical defenses (e.g., advanced firewalls, real-time monitoring), institutional policies, regular risk assessments, and assign resources for cybersecurity training to address systemic vulnerabilities [27,80,83].
     - **Focus on user training and awareness:** Conduct regular training programs and real-world simulations for students and staff to recognize multi-vector attacks and respond effectively [19,27,72,83,93].
     - **Collaborate with industry experts:** Partner with cybersecurity firms and leverage external expertise to stay ahead of evolving threats and implement best practices

[15,72,83].

7.2. AI detection strategies

The evolving landscape of AI-driven threats necessitates sophisticated detection strategies to combat phishing, deepfake media, and misinformation in social learning environments. The reviewed literature highlights multiple AI-based approaches, including machine learning (ML), deep learning (DL), behavioral analysis, and explainable AI (XAI) to enhance detection capabilities. While academic research focuses on the development and theoretical advancement of detection models, gray literature emphasizes practical implementation, real-world applications, and policy-driven responses.

1. **Points of convergence**
    - **The necessity of AI-powered detection systems:** Both academic and gray literature agree that traditional cybersecurity tools are insufficient in detecting AI-generated threats, necessitating the use of machine learning, natural language processing (NLP), and deep learning techniques [8,20,21,92].
    - **Explainability and trust in detection models:** While academic literature explores technical aspects of explainability through XAI [63,64], gray literature highlights the market demand for transparent AI systems that can justify why certain content is flagged [88,92].
    - **Behavioral-based anomaly detection:** Both sources highlight that AI-based threat detection should extend beyond content analysis to include user behavior analytics, such as login frequency, browsing habits, and sudden changes in engagement [8,25,63,89,93,97]. AI-enhanced behavioral analytics can help detect anomalies indicative of phishing attacks, account takeovers, or deepfake-driven identity fraud. Research indicates that combining behavioral tracking with AI-driven anomaly detection significantly enhances early threat detection and mitigates security breaches in educational institutions [48].
    - **Integration of Blockchain technology:** Both sources emphasize that Blockchain-based solutions offer tamper-proof audit trails, enhancing media authentication and providing robust forensic capabilities to counteract deepfake media [55,94,95].
2. **Points of divergence**
    - **Theoretical advancements vs. practical deployments:** Academic research primarily focuses on advancing detection models, proposing techniques such as adversarial retraining, contrastive learning, and multimodal fusion for detecting manipulated media [4,46,55]. On the other hand, gray literature highlights scalability and real-time efficiency, discussing industry-driven frameworks and the integration of AI models into existing cybersecurity solutions [90,92,98].
    - **Ethical and regulatory considerations:** Academic literature often discusses AI bias, fairness in detection systems, and ethical concerns surrounding deepfake detection and misinformation control [6,34,62]. In contrast, gray literature leans more toward pragmatic implementation strategies, such as government-backed initiatives and industry-driven frameworks for AI detection [83,86,92].
    - **Effectiveness of watermarking and AI-forensics:** Gray literature reports optimism regarding

watermarking techniques (e.g., SynthID), digital fingerprinting, and AI-generated content tracking [94,96], while academic literature remains skeptical about their robustness against adversarial attacks and bypassing methods [53,55].

3. **Recommendations from the literature**
    - **Academic literature recommendations:**
        - **Developing inclusive training datasets**
            - Academic studies highlight the importance of diverse and representative datasets for training AI detection systems. These datasets must account for cultural, linguistic, and contextual differences to enhance robustness and adaptability [62,63].
            - Researchers suggest refining adversarial training, federated learning, and zero-shot detection techniques to improve AI-driven detection mechanisms [4,20,53].
        - **Advancing explainable AI (XAI) mechanisms**
            - Academic literature highlights the necessity of integrating Explainable AI (XAI) to provide transparent decision-making processes, fostering end-user trust and reducing false positives and negatives [22,63,64].
            - XAI tools also facilitate better understanding and evaluation by stakeholders, such as educators and administrators [64].
        - **Blockchain integration for media authentication:** Blockchain is recommended in academic studies as an effective mechanism for creating tamper-proof logs and verifiable audit trails, which are critical for forensic investigations and confirming media [52,55].
        - **Fostering interdisciplinary collaboration:** Academic research highlights the need for collaboration across academia, industry, and policymakers to share resources, datasets, and best practices to develop robust and scalable AI detection systems [63].
    - **Gray literature recommendations:**
        - **Prioritizing real-time scalability and efficiency:** Gray literature highlights the challenges of real-time detection, supporting lightweight AI models optimized for scalability and rapid response in dynamic environments [94,95,97].
        - **Enhancing awareness and education campaigns:**
            - Reports and articles reveal the importance of empowering users such as students, educators, and administrators, through training programs that highlight AI threats and mitigation strategies. This is especially critical in education environments where technological literacy may vary widely [72,83,96].
            - Campaigns could include simulated phishing attacks and hands-on training to improve threat recognition [84,94].
        - **Focusing on ethical AI development:** Ethical concerns raised in gray literature include privacy issues in data collection and potential misuse of AI systems.

Organizations are recommended to adopt transparent policies and utilize privacy-preserving techniques, such as differential privacy, to ensure user trust [24,92,94].

- **Collaboration across sectors:** Industry and governmental organizations emphasize the need for global partnerships to address the cross-border nature of AI-driven threats. For instance, joint initiatives can streamline resource-sharing and the creation of standardized frameworks [83,92,94,96].
- **Leveraging Blockchain and advanced detection tools:** Articles highlight incorporating blockchain for ensuring data authenticity and deploying cutting-edge tools for deepfake detection. Security companies and tech companies advocate for the widespread adoption of AI-generated content tracking, such as SynthID, and real-time AI forensics like watermarking techniques and real-time forgery detection models [94,95].
- **Integration with commercial cybersecurity frameworks:** Experts recommend embedding AI-driven detection models within existing cybersecurity tools, using multi-layered authentication, behavioral analytics, and blockchain verification [90,93,98].

7.3. Ethical and policy frameworks

The rapid integration of AI in education has required the development of ethical and policy frameworks to guide responsible AI adoption. The reviewed literature, both academic and gray, emphasizes key principles such as transparency, fairness, accountability, and data privacy. However, while academic literature primarily focuses on theoretical foundations and long-term governance models, gray literature tends to highlight practical challenges and institutional compliance strategies.

1. **Key findings - Ethical considerations in AI-Driven education.**
    - **Fairness and bias mitigation:** Academic sources explore algorithmic fairness, emphasizing the need to mitigate bias in AI-driven assessments [81], while gray literature raises concerns about real-world discrimination cases in educational AI deployment [85].
    - **Data privacy and security:** The ethical responsibility of institutions to protect student data is a consistent topic, particularly considering regulatory challenges such as GDPR and FERPA [62,85,92].
2. **Key findings - Policy gaps and governance models**
    - **Institutional and national-level policies:** Academic research discusses policy frameworks at both institutional and governmental levels, proposing AI-specific regulatory structures [21,81]
    - **Practical policy implementation:** Gray literature highlights challenges in AI governance, such as unclear accountability in AI-driven grading systems and content moderation [85,92].
    - **International regulations and compliance:** Reports from organizations like the World Economic Forum discuss the need for international AI governance standards in education [85].
3. **Points of convergence**
    - **Need for ethical AI development:** Both academic and gray literature emphasize that AI must

align with ethical principles to prevent harm and maintain trust in educational institutions [81,92].

- **Transparency as a core principle:** There is agreement that AI-driven decisions, particularly in automated grading and student evaluations, must be explainable and interpretable [45,63,83,85,92].
- **Collaboration between stakeholders:** Sources converge on the need for collaboration among policymakers, educators, AI developers, and students to ensure AI is implemented responsibly in education [21,81,83,85,92].

4. **Points of divergence**
   - **Technical vs. Practical Perspectives:**
     - Academic literature focuses on theoretical discussions about algorithmic fairness, AI governance, and privacy-by-design methodologies [45,62].
     - Gray literature presents real-world cases of AI-related debates, such as biased grading systems and faulty student monitoring tools [85, 92].
   - **Regulatory vs. self-governance approaches:**
     - Academic studies support strict AI regulations and governmental oversight [21,81].
     - Gray literature emphasizes institutional self-regulation and industry-led compliance frameworks [85,92].
   - **Regulatory vs. self-governance approaches:**
     - Academic research proposes explainability frameworks to hold AI decision-making accountable [63,64].
     - Gray literature raises concerns about practical challenges in enforcing AI accountability at the institutional level [92].
5. **Recommendations from the literature**
   - **Academic recommendations:**
     - **Develop comprehensive AI governance models**: Researchers emphasize the need for standardized policies across institutions, incorporating ethical AI principles at national and international levels [21,81].
     - **Integrate explainable AI (XAI) in educational systems**: To improve trust and accountability, institutions should adopt AI models that allow educators and students to understand AI-driven recommendations and decisions [45,63].
     - **Enforce data protection regulations:** Institutions must align AI-driven learning platforms with regulatory requirements such as GDPR, FERPA, and AI-specific legal frameworks [62].
   - **Gray literature recommendations:**
     - **Enhance institutional AI ethics training:** Universities and schools should offer mandatory AI ethics training for teachers, students, and administrators to ensure responsible AI usage [85,92].
     - **Develop AI policy guidelines at the institutional level:** Institutions should draft internal AI policies addressing transparency, fairness, and data privacy to mitigate risks

associated with biased AI models and unethical automation [85].

- **Promote multi-stakeholder collaboration**: AI governance should involve policymakers, teachers, industry leaders, and students to create adaptive and mandatory policies [81,85].

8. Conclusion

In this multivocal literature review (MLR), the duality of artificial intelligence (AI) in education becomes clear: on one hand, AI supports innovative teaching methods, personalization, and collaboration; on the other, malicious actors leverage AI to orchestrate damaging phishing attacks, misinformation campaigns, and deepfake media. Through a detailed examination of academic sources and gray literature, we identified how these threats weaken trust, disrupt learning activities, and create ethical dilemmas in educational contexts. Although significant progress has been achieved in developing AI-driven detection mechanisms, such as advanced machine learning models and blockchain-based audit trails challenges remain in scalability, user awareness, and ethical governance.

The research shows that technical solutions alone are insufficient. Instead, a balanced approach that merges technological defenses with user-focused strategies and clear policy guidelines is required. Institutions should prioritize digital literacy and ethics training to raise awareness about AI-driven threats, while explainable AI (XAI) frameworks can foster transparency, enhancing trust in automated detection. Collaborative policymaking among educators, researchers, industry experts, and governments is also essential to address cross-border issues and evolving attack methods.

Moving forward, future investigations might explore how to integrate these technical and human-centered solutions in ways that respect data privacy, mitigate algorithmic bias, and maintain the flexibility needed for educational innovation. By adopting an adaptive, ethical, and collaborative posture, the education sector can better safeguard social learning ecosystems against the escalating landscape of AI-powered threats and ensure that the transformative promise of AI remains an asset, rather than a vulnerability, to educational progress.

9. Author contributions

N.M.C. and J.B. methodology.

10. Funding

This research received no external funding.

11. Data availability statement

No new data was created.

12. Conflicts of interest

The author declares no conflict of interest.

13. References

[1] Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. Intelligence Unleashed: An Argument for AI in Education. *Pearson*. 2016. [Online]. Available: https://discovery.ucl.ac.uk/id/eprint/1475756/

[2] Kinshuk, Chen, NS., Cheng, IL. et al. Evolution Is not enough: Revolutionizing Current Learning Environments to Smart Learning Environments. *Int J Artif Intell Educ 26, 561–581.* 2016. [Online]. Available: https://doi.org/10.1007/s40593-016-0108-x

[3] Fritsch, L., Jaber, A., Yazidi, A. An Overview of Artificial Intelligence Used in Malware. In: Zouganeli, E., Yazidi, A., Mello, G., Lind, P. (eds) Nordic Artificial Intelligence Research and Development. NAIS 2022. Communications in Computer and Information Science, vol 1650. Springer, Cham. 2022. [Online]. Available: https://doi.org/10.1007/978-3-031-17030-0_4

[4] M. R. Shoaib, Z. Wang, M. T. Ahvanooey and J. Zhao, Deepfakes, Misinformation, and Disinformation in the Era of Frontier AI, Generative AI, and Large AI Models. *2023 International Conference on Computer and Applications (ICCA)*, Cairo, Egypt, pp. 1-7. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10401723

[5] S. Laato, A. Farooq, H. Tenhunen, T. Pitkamaki, A. Hakkala and A. Airola. AI in Cybersecurity Education-A Systematic Literature Review of Studies on Cybersecurity MOOCs. 2020 IEEE 20th International Conference on Advanced Learning Technologies (ICALT), Tartu, Estonia, pp. 6-10. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9156050

[6] Ferrara, E. GenAI against humanity: nefarious applications of generative artificial intelligence and large language models. J Comput Soc Sc 7, 549–569. 2024. [Online]. Available: https://doi.org/10.1007/s42001-024-00250-1

[7] S. M.Nour and S. A.Said- Harnessing the Power of AI for Effective Cybersecurity Defense, 2024 6th International Conference on Computing and Informatics (ICCI), New Cairo - Cairo, Egypt, pp. 98-102. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10485059

[8] X. Zhang, P. Wang, H. Jia, Z. Huang and R. Zhao. AI-Powered Cybersecurity: Enhancing Threat Detection and Defense in the Digital Age, 2024 IEEE 7th International Conference on Electronic Information and Communication Technology (ICEICT), Xi'an, China, pp. 1026-1031. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10670798

[9] J. Wei-Kocsis, M. Sabounchi, B. Yang and T. Zhang. Cybersecurity Education in the Age of Artificial Intelligence: A Novel Proactive and Collaborative Learning Paradigm, 2022 IEEE Frontiers in Education Conference (FIE), Uppsala, Sweden, pp. 1-5. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9962643

[10] Tim Femister - Forbes. Education Under Siege: The Rising Threat Of Cyberattacks On K-12 Schools. Internet: https://www.forbes.com/councils/forbestechcouncil/2024/05/06/education-under-siege-the-rising-threat-of-cyberattacks-on-k-12-schools, May 6, 2024. [Nov. 02, 2024]

[11] Kyle Chin - UpGuard. Why is the Education Sector a Target for Cyber Attacks? Internet: https://www.upguard.com/blog/education-sector-cyber-attacks, November 18, 2024. [Nov. 20, 2024]

[12] Microsoft Threat Intelligence. Cyber Signals Issue 8 | Education under siege: How cybercriminals target our

schools. Internet: https://www.microsoft.com/en-us/security/blog/2024/10/10/cyber-signals-issue-8-education-under-siege-how-cybercriminals-target-our-schools/, October 10, 2024. [Nov. 02, 2024]

[13] Ani Petrosyan - Statista. Average weekly number of cyberattacks in organizations worldwide in 2022 and 2023, by industry. Internet: https://www.statista.com/statistics/1377217/average-weekly-number-attacks-global-by-industry/, December 10, 2024. [Dec. 28, 2024]

[14] David Angerdina. Top Cyberattacks on Schools in 2024 and How to Prevent Them. Internet: https://charterts.com/insights/top-cyberattacks-on-schools-in-2024-and-how-to-prevent-them, December 9, 2024. [Dec. 14, 2024]

[15] Check Point Team. A Closer Look at Q3 2024: 75% Surge in Cyber Attacks Worldwide. Internet: https://blog.checkpoint.com/research/a-closer-look-at-q3-2024-75-surge-in-cyber-attacks-worldwide, October 18, 2024. [Nov. 02, 2024]

[16] Rob Sobers. 157 Cybersecurity Statistics and Trends. Internet: https://www.varonis.com/blog/cybersecurity-statistics, September 13, 2024. [Nov. 03, 2024]

[17] IBM. Cost of a Data Breach Report 2024. Internet: https://www.ibm.com/reports/data-breach, August 13, 2024. [Nov. 03, 2024]

[18] Tomer Ronen. 31 Must-Know Education Cybersecurity Statistics. Internet: https://www.varonis.com/blog/education-cybersecurity-statistics, October 08, 2024. [Nov. 03, 2024]

[19] Intelecis. Education Sector Faces Growing Cybersecurity Challenges in 2024. Internet: https://www.intelecis.com/education-sector-faces-growing-cybersecurity-challenges-in-2024, November 20, 2024. [Nov. 23, 2024]

[20] M. S. Rana and A. H. Sung. Advanced Deepfake Detection using Machine Learning Algorithms: A Statistical Analysis and Performance Comparison. 2024 7th International Conference on Information and Computer Technologies (ICICT), Honolulu, HI, USA, pp. 75-81. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10541828

[21] S. Jawhar, J. Miller and Z. Bitar. AI-Based Cybersecurity Policies and Procedures. *2024 IEEE 3rd International Conference on AI in Cybersecurity (ICAIC),* Houston, TX, USA, 2024, pp. 1-5. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10433845

[22] Aldawood H, Skinner G. Reviewing Cyber Security Social Engineering Training and Awareness Programs—Pitfalls and Ongoing Issues. *Future Internet*, 11(3):73. 2019. [Online]. Available: https://doi.org/10.3390/fi11030073

[23] R. Sasaki. AI and Security - What Changes with Generative AI. 2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security Companion (QRS-C), Chiang Mai, Thailand, pp. 208-215. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10430001

[24] Maddy Ell – UK Gov. - Department for Science, Innovation & Technology. Cyber security breaches survey 2024: education institutions annex . Internet: https://www.gov.uk/government/statistics/cyber-security-breaches-survey-2024/cyber-security-breaches-survey-2024-education-institutions-annex, April 09, 2024. [Nov. 11, 2024]

[25] L. Tang and Q. H. Mahmoud. A Deep Learning-Based Framework for Phishing Website Detection, in IEEE Access, vol. 10, pp. 1509-1521. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9661323

[26] Norliza Katuk, Adi Badiozaman Ruhani, Muzdalini Malik, Abdul Kadir Mahamood, and Mohd Shamshul Anuar Omar. Intelligent Systems of Computing and Informatics - Protecting Higher Learning Institutions from

Phishing Attacks: A Staff Awareness Program. USA. CRC Press. 2024

[27] Sterling Ideas. Phishing Attacks in Schools: How to Recognize and Prevent Them. https://www.sterlingideas.com/phishing-cyberattacks-schools, August 14, 2024. [Dec. 02, 2024]

[28] U.S. Department of Education. K-12 Cybersecurity. https://www.ed.gov/teaching-and-administration/safe-learning-environments/school-safety-and-security/k-12-cybersecurity, December 01, 2024. [Dec. 27, 2024]

[29] Xinyi Zhou and Reza Zafarani. A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities. ACM Comput. Surv. 53, 5, Article 109. 2020. [Online]. Available: https://doi.org/10.1145/3395046

[30] Kai Shu, Suhang Wang, and Huan Liu. Beyond News Contents: The Role of Social Context for Fake News Detection. In Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19). Association for Computing Machinery, New York, NY, USA, 312–320. 2019. [Online]. Available: https://doi.org/10.1145/3289600.3290994

[31] Priyanka Meel, Dinesh Kumar Vishwakarma. Fake news, rumor, information pollution in social media and web: A contemporary survey of state-of-the-arts, challenges and opportunities, Expert Systems with Applications, Volume 153, 112986, ISSN 0957-4174. 2020. [Online]. Available: https://doi.org/10.1016/j.eswa.2019.112986

[32] Rogerson, A.M. Student Peer-to-Peer File Sharing as an Academic Integrity Issue. In: Eaton, S.E. (eds) Second Handbook of Academic Integrity. Springer International Handbooks of Education. Springer, Cham. 2024. [Online]. Available: https://doi.org/10.1007/978-3-031-54144-5_55

[33] Danni Xu, Shaojing Fan, and Mohan Kankanhalli. Combating Misinformation in the Era of Generative AI Models. In Proceedings of the 31st ACM International Conference on Multimedia (MM '23). Association for Computing Machinery, New York, NY, USA, 9291–9298. 2023. [Online]. Available: https://doi.org/10.1145/3581783.3612704

[34] AbuJarour, S., Qarariah, A., Saadeh, N., Salem, M. AI, Misinformation, and Fake News: A Literature Review of Ethical and Technical Approaches. In: Mansour, N., Bujosa Vadell, L.M. (eds) Finance and Law in the Metaverse World. Contributions to Finance and Accounting. Springer, Cham. 2024. [Online]. Available: https://doi.org/10.1007/978-3-031-67547-8_55

[35] Caled, D., Silva, M.J. Digital media and misinformation: An outlook on multidisciplinary strategies against manipulation. J Comput Soc Sc 5, 123–159. 2022. [Online]. Available: https://doi.org/10.1007/s42001-021-00118-8

[36] Soga K, Yoshida S, Muneyasu M. Graph-Based Interpretability for Fake News Detection through Topic- and Propagation-Aware Visualization. Computation, 12(4):82. 2024. [Online]. Available: https://doi.org/10.3390/computation12040082

[37] A. Kundu and U. T. Nguyen. Automated Fact Checking Using A Knowledge Graph-based Model, International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Osaka, Japan, 2024, pp. 709-716, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10463196

[38] Savoy, J. Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling. 2020.

[39] I. Alsmadi et al. Adversarial NLP for Social Network Applications: Attacks, Defenses, and Research Directions, in IEEE Transactions on Computational Social Systems, vol. 10, no. 6, pp. 3089-3108. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9942953

[40] A. M. B. Flandoli and J. M. S. Eguiguren. Media and digital literacy: From particularities to encounters and

possibilities," 2021 16th Iberian Conference on Information Systems and Technologies (CISTI), Chaves, Portugal, pp. 1-6. 2021. Online]. Available: https://ieeexplore.ieee.org/document/9476428

[41] Alon, A.T., Rahimi, I.D. & Tahar, H. Fighting fake news on social media: a comparative evaluation of digital literacy interventions. Curr Psychol 43, 17343–17361. 2024. [Online]. Available: https://doi.org/10.1007/s12144-024-05668-4

[42] Kolić-Vehovec, S., Pahljina-Reinić, R. & Rončević Zubković, B. Effects of collaboration and informing students about overconfidence on metacognitive judgment in conceptual learning. Metacognition Learning 17, 87–116. 2022. [Online]. Available: https://doi.org/10.1007/s11409-021-09275-7

[43] Jia, L., Jin, H. The role of relevance in the continued influence effect of misinformation under different retraction methods. Curr Psychol 43, 21437–21447.2024. [Online]. Available: https://doi.org/10.1007/s12144-024-05967-w

[44] Ecker, U.K.H., Lewandowsky, S., Swire, B. et al. Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. Psychon Bull Rev 18, 570–578. 2011. [Online]. Available: https://doi.org/10.3758/s13423-011-0065-1

[45] Yan, Y., Liu, H. Ethical framework for AI education based on large language models. *Educ Inf Technol*. 2024. [Online]. Available: https://doi.org/10.1007/s10639-024-13241-6

74[46] D. Chapagain, N. Kshetri and B. Aryal. Deepfake Disasters: A Comprehensive Review of Technology, Ethical Concerns, Countermeasures, and Societal Implications, 2024 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Windhoek, Namibia, pp. 1-9. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10767452

[47] S. A. Laczi and V. Póser, "Impact of Deepfake Technology on Children: Risks and Consequences," 2024 IEEE 22nd Jubilee International Symposium on Intelligent Systems and Informatics (SISY), Pula, Croatia, pp. 215-220. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10737593

[48] Jen A. Mille. Ed Tech - 3 Ways Artificial Intelligence Can Improve Campus Cybersecurity. https://edtechmagazine.com/higher/article/2020/03/3-ways-artificial-intelligence-can-improve-campus-cybersecurity-perfcon, March 30, 2020. [Dec. 02, 2024]

[49] Lauren Coffey. Inside Higher Ed - AI Voice Clones and Deepfakes: The Latest Presidents' Engagement Tools. https://www.insidehighered.com/news/tech-innovation/artificial-intelligence/2023/11/14/presidents-use-ai-voice-clones-and, November 14, 2023. [Dec. 02, 2024]

[50] S. A. Shahzad, A. Hashmi, S. Khan, Y. -T. Peng, Y. Tsao and H. -M. Wang. Lip Sync Matters: A Novel Multimodal Forgery Detector, Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Chiang Mai, Thailand, pp. 1885-1892. 2022. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9980296

[51] SD. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi and L. Verdoliva. Are GAN-Generated Images Easy to Detect? A Critical Analysis of the State-Of-The-Art, 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, pp. 1-6. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9428429

[52] Hedge, A.S., Vinutha, M.N., Supriya, K., Nagasundari, S., Honnavalli, P.B. CLH: Approach for Detecting Deep Fake Videos. In: Abdullah, N., Manickam, S., Anbar, M. (eds) Advances in Cyber Security. ACeS 2021. Communications in Computer and Information Science, vol 1487. Springer, Singapore. 2021.

https://doi.org/10.1007/978-981-16-8059-5_33

[53] Rani, A., Jain, A. Digital Image Forensics-Image Verification Techniques. In: Dash, S.S., Das, S., Panigrahi, B.K. (eds) Intelligent Computing and Applications. Advances in Intelligent Systems and Computing, vol 1172. Springer, Singapore. 2021. https://doi.org/10.1007/978-981-15-5566-4_19

[54] S. G and P. Vijaybaskar. DeepFake Detection by Prediction of Mismatch Between Audio and Video Lip Movement, 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, pp. 01-08. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10533515

[55] Kolić-L. Verdoliva. Media Forensics and DeepFakes: An Overview, in IEEE Journal of Selected Topics in Signal Processing, vol. 14, no. 5, pp. 910-932, Aug. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9115874

[56] E. Preu, M. Jackson and N. Choudhury. Perception vs. Reality: Understanding and Evaluating the Impact of Synthetic Image Deepfakes over College Students, 2022 IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, NY, USA, pp. 0547-0553. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9965697

[57] Lauraine Langreo, Education Week, Bethesda, Md. How Should U.S. Schools Confront Deepfakes? https://www.govtech.com/education/k-12/how-should-u-s-schools-confront-deepfakes, September 27, 2024. [November. 05, 2024]

[58] Deepfakes and Higher Education: A Research Agenda and Scoping Review of Synthetic Media. Journal of University Teaching and Learning Practice, 21(10). 2024. https://doi.org/10.53761/2y2np178

[59] Matteo Wong. High School Is Becoming a Cesspool of Sexually Explicit Deepfakes. https://www.theatlantic.com/technology/archive/2024/09/ai-generated-csam-crisis/680034, September 26, 2024. [Nov. 02, 2024]

[60] Kara Arundel. Schools lack supports for victims of sexually explicit deepfake and real images. https://www.k12dive.com/news/schools-deepfake-images-student-supports/728107, September 26, 2024. [Nov. 05, 2024]

[61] Dana Nickel. AI is shockingly good at making fake nudes — and causing havoc in schools. https://www.politico.com/news/2024/05/28/ai-deepfake-nudes-schools-states-00160183, May 29, 2024. [Nov. 05, 2024]

[62] Dhingra, M., Goyal, R., Goyal, S.J. Cybersecurity Challenges and Opportunities in Education 4.0. In: Abdul Karim, S.A. (eds) Intelligent Systems Modeling and Simulation III. Studies in Systems, Decision and Control, vol 553. Springer, Cham. 2024. [Online]. Available: https://doi.org/10.1007/978-3-031-67317-7_20

[63] B. Desai, K. Patil, I. Mehta and A. Patil, "Explainable AI in Cybersecurity: A Comprehensive Framework for enhancing transparency, trust, and Human-AI Collaboration," 2024 International Seminar on Application for Technology of Information and Communication (iSemantic), Semarang, Indonesia, pp. 135-150. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10762690

[64] Zhixin Pan, Prabhat Mishra. Explainable AI for Cybersecurity. USA. Springer. 2023

[65] Stansfield, C., Dickson, K. & Bangpan, M. Exploring issues in the conduct of website searching and other online sources for systematic reviews: how can we be systematic? *Syst Rev* 5, 191. 2016. [Online]. Available: https://doi.org/10.1186/s13643-016-0371-9

[66] Bellefontaine, S.P., Lee, C.M. Between Black and White: Examining Grey Literature in Meta-analyses of

Psychological Research. *J Child Fam Stud 23*, 1378–1388. 2014. [Online]. Available: https://doi.org/10.1007/s10826-013-9795-1

[67] Coleman, S., Wright, J.M., Nixon, J. et al. Searching for Programme theories for a realist evaluation: a case study comparing an academic database search and a simple Google search. *BMC Med Res Methodol 20*, 217. 2020. [Online]. Available: https://doi.org/10.1186/s12874-020-01084-x

[68] Jones, M.A.E. LibGuides: Systematic Reviews: Step 8: Write the Review. [Online]. Available: https://guides.lib.unc.edu/systematic-reviews/write

[69] Vahid Garousi, Michael Felderer, Mika V. Mäntylä. Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Inf. Softw. Technol* 106, 101-121. 2019.[Online]. Available: https://doi.org/10.1016/j.infsof.2018.09.006

[70] Vahid Garousi, Michael Felderer, Mika V. Mäntylä. The need for multivocal literature reviews in software engineering. *ACM Int. Conf. Proceeding Ser* 26, 1-6. 2019. [Online]. Available: https://doi.org/10.1145/2915970.2916008

[71] Kitchenham, Barbara & Charters, Stuart. Guidelines for performing Systematic Literature Reviews in Software Engineering 2. 2007. [Online]. Available: https://www.researchgate.net/publication/302924724_Guidelines_for_performing_Systematic_Literature_Reviews_in_Software_Engineering

[72] Desai, D., & Hegde, R. Zscaler Blog - Phishing Attacks Rise: ThreatLabz 2024 Phishing Report. https://zerotrust.cio.com/wp-content/uploads/sites/64/2024/09/threatlabz-phishing-report-2024.pdf, April 24, 2024. [December. 15, 2024]

[73] Mackenzie Tatananni. The U.S Sun - RISE OF AI Flawless 'deepfakes' you can't spot and 'manipulative' robots that cheat you – most terrifying AI breakthroughs revealed. https://www.the-sun.com/tech/11904094/artificial-intelligence-dangerous-breakthroughs-deepfakes-weapons, July 13, 2024. [December. 15, 2024]

[74] Matt Burgess & Lily Hay Newman. Wired - Pig Butchering Scams Are Going High Tech. https://www.wired.com/story/pig-butchering-scams-go-high-tech, October 12, 2024. [Dec. 15, 2024]

[75] Enzo Cervini & Maria Carro. ISPI - An Overview of the Impact of GenAI and Deepfakes on Global Electoral Processes. https://www.ispionline.it/en/publication/an-overview-of-the-impact-of-genai-and-deepfakes-on-global-electoral-processes-167584, October 25, 2024. [Dec. 15, 2024]

[76] Melissa Heikkilä. MIT Technology Review - Why detecting AI-generated text is so difficult (and what to do about it). https://www.technologyreview.com/2023/02/07/1067928/why-detecting-ai-generated-text-is-so-difficult-and-what-to-do-about-it, February 07, 2023. [Dec. 15, 2024]

[77] Nathan Schmidt. News AU - Australian schools grappling with AI-driven deepfake crisis. https://www.news.com.au/technology/online/proposed-deepfake-bill-not-enough-to-combat-unprecedented-crisis/news-story/ffb8e5f9a661ef9270b222c5ff4e0dec, February 07, 2023. [Dec. 16, 2024]

[78] Stu Sjouwerman. Forbes - Deepfake Phishing: The Dangerous New Face Of Cybercrime. https://www.forbes.com/councils/forbestechcouncil/2024/01/23/deepfake-phishing-the-dangerous-new-face-of-cybercrime, January 23, 2023. [Dec. 17, 2024]

[79] N. Bontridder and Y. Poullet. The role of artificial intelligence in disinformation, Data & Policy, vol. 3, p. e32. 2021. [Online]. Available: https://www.cambridge.org/core/journals/data-and-policy/article/role-of-

artificial-intelligence-in-disinformation/7C4BF6CA35184F149143DE968FC4C3B

[80] Jason Risch. Greylock - Deepfakes and the New Era of Social Engineering. https://greylock.com/greymatter/deepfakes-and-the-new-era-of-social-engineering, October 09, 2023. [Dec. 17, 2024]

[81] Nguyen, A., Ngo, H.N., Hong, Y. *et al*. Ethical principles for artificial intelligence in education. *Educ Inf Technol*. 2023. [Online]. Available: https://doi.org/10.1007/s10639-022-11316-w

[82] Brooke Kato. New York Post - Gmail, Outlook and Apple users urged to watch out for this new email scam: Cybersecurity experts sound alarm. https://nypost.com/2025/01/04/tech/gmail-outlook-and-apple-users-urged-to-watch-out-for-this-new-email-scam-cybersecurity-experts-sound-alarm, January 04, 2025. [Jan. 08, 2025]

[83] Abdulaziz Almaslukh. World Economic Forum - AI could empower and proliferate social engineering cyberattacks. https://www.weforum.org/stories/2024/10/ai-agents-in-cybersecurity-the-augmented-risks-we-all-need-to-know-about/, October 25, 2024. [Dec. 17, 2024]

[84] Christine Wong. CSO - What is phishing? Examples, types, and techniques. https://www.csoonline.com/article/514515/what-is-phishing-examples-types-and-techniques.html, October 11, 2024. [Dec. 17, 2024]

[85] Tanya Milberg. World Economic Forum - The future of learning: How AI is revolutionizing education 4.0. https://www.weforum.org/stories/2024/04/future-learning-ai-revolutionizing-education-4-0/, April 28, 2024. [Dec. 18, 2024]

[86] Neil Lappage. ISACA - The Role of Deepfake Technology in the Landscape of Misinformation and Cybersecurity Threats. https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2023/the-role-of-deepfake-technology-in-the-landscape-of-misinformation-and-cybersecurity-threats, August 09, 2023. [Dec. 17, 2024]

[87] CISA. Avoiding Social Engineering and Phishing Attacks. https://www.cisa.gov/news-events/news/avoiding-social-engineering-and-phishing-attacks, February 01, 2021. [Dec. 17, 2024]

[88] Benjamin Chou. Forbes - The Digital Sentry: How AI Will Revolutionize Financial Fraud Investigation. https://www.forbes.com/councils/forbestechcouncil/2023/07/03/the-digital-sentry-how-ai-will-revolutionize-financial-fraud-investigation, July 03, 2023. [Dec. 20, 2024]

[89] Matthew Tyson. CSO - 7 guidelines for identifying and mitigating AI-enabled phishing campaigns. https://www.csoonline.com/article/574745/7-guidelines-for-identifying-and-mitigating-ai-enabled-phishing-campaigns.html, March 20, 2023. [Dec. 20, 2024]

[90] Christopher Beam. Wired – The AI Detection Arms Race Is On. https://www.wired.com/story/ai-detection-chat-gpt-college-students/, September 14, 2023. [Dec. 20, 2024]

[91] Kate Knibbs. Wired – Researchers Tested AI Watermarks—and Broke All of Them. https://www.wired.com/story/artificial-intelligence-watermarking-issues/, October 03, 2023. [Dec. 20, 2024]

[92] Steve Durbin. Forbes - The Risks And Rewards Of AI: Strategies For Mitigation And Containment. https://www.forbes.com/councils/forbesbusinesscouncil/2024/06/05/the-risks-and-rewards-of-ai-strategies-for-mitigation-and-containment/, June 05, 2024. [Dec. 20, 2024]

[93] Cathy Ross. Forbes - AI And Cybercrime: Is Fraud Detection The Final Backstop? https://www.forbes.com/councils/forbesbusinesscouncil/2024/10/04/ai-and-cybercrime-is-fraud-detection-the-

final-backstop/, October 04, 2024. [Dec. 20, 2024]

[94] Sheena Vasani. The Verge - Google open-sourced its watermarking tool for AI-generated text https://www.theverge.com/2024/10/23/24277873/google-artificial-intelligence-synthid-watermarking-open-source, October 23, 2024. [Dec. 20, 2024]

[95] Reece Rogers. Wired – Real-Time Video Deepfake Scams Are Here. This Tool Attempts to Zap Them. https://www.wired.com/story/real-time-video-deepfake-scams-reality-defender, October 15, 2024. [Dec. 20, 2024]

[96] Will Knight. MIT Technology Review – Facebook is making its own AI deepfakes to head off a disinformation disaster. https://www.technologyreview.com/2019/09/05/65353/facebook-is-making-ai-deepfakes-to-head-off-a-disinformation-disaster/, September 05, 2019. [Dec. 20, 2024]

[97] Will Knight. MIT Technology Review – A new deepfake detection tool should keep world leaders safe—for now. https://www.technologyreview.com/2019/06/21/134815/a-new-deepfake-detection-tool-should-keep-world-leaders-safefor-now, June 21, 2019. [Dec. 21, 2024]

[98] Will Knight. Wired – Deepfakes Are Evolving. This Company Wants to Catch Them All. https://www.wired.com/story/deepfake-detection-get-real-labs, June 27, 2024. [Dec. 21, 2024]

[99] Mateus-Coelho, Nuno, and Manuela Cruz-Cunha, editors. Exploring Cyber Criminals and Data Privacy Measures. IGI Global, 2023. https://doi.org/10.4018/978-1-6684-8422-7

[100] Mateus-Coelho, Nuno Ricardo, et al. "POSMASWEB: Paranoid Operating System Methodology for Anonymous and Secure Web Browsing." Handbook of Research on Cyber Crime and Information Privacy, edited by Maria Manuela Cruz-Cunha and Nuno Mateus-Coelho, IGI Global, 2021, pp. 466-497. https://doi.org/10.4018/978-1-7998-5728-0.ch023

[101] Mateus-Coelho, N. (2021). A New Methodology for the Development of Secure and Paranoid Operating Systems. Procedia Computer Science, 181, 1207-1215. https://doi.org/10.1016/j.procs.2021.01.318