Received: 12 August 2025 | Revised: 25 May 2026 | Accepted: 26 June 2026 | Published online: 20 July 2026

REVIEW

Artificial Intelligence and Applications
2026, Vol. 4(4) 445–468
DOI: 10.47852/bonviewAIA62027173

AI in Cybersecurity Serious Games and Training Platforms: A Systematic Review with an Illustrative AI-Based Reverse Turing Game

Bilgin Metin[1], Tuncay Avci[1], Atakan Yeşilkayali[1], Hikmet Sami Karaca[1], Ali Nehir Dündar[1] and Martin Wynn[2,*]

[1] *Management Information Systems Department, Bogazici University, Turkey*

[2] *School of Business, Computing and Social Sciences, University of Gloucestershire, UK*

Abstract: The use of artificial intelligence (AI) in cybersecurity serious games and training platforms is an innovative approach, which enhances the effectiveness of security training by improving the development of practical skills. Numerous studies have focused on the potential of AI-enhanced serious games, highlighting the flexibility and adaptability of this form of education in the cybersecurity domain. The overall aim of this study is to map out the AI technology landscape relevant to cybersecurity games and assess relevant theory and pedagogical frameworks. First, a systematic review of publications from 2019 to 2025 is performed to identify the primary AI techniques and technologies adopted in cybersecurity and to find those relevant to serious games and training platforms. Then, the factors that contribute to their success and the barriers that affect their implementation in game-based environments are analyzed, together with the theoretical frameworks and pedagogical models that support the design of AI-enhanced training platforms. As an illustrative example of these results, an AI-based reverse Turing game is presented. The article will be of interest to those who design cybersecurity games and also to educators who are trying to create training formats that are effective in responding to the fast-changing cybersecurity environment.

Keywords: artificial intelligence, cybersecurity serious games, training platforms, reverse Turing test, Turing test

1. Introduction

The rapid advancement of digitalization has brought to the fore the significance of cybersecurity in modern technology infrastructure [1–4]. Cyber threats are becoming increasingly complex, targeting many different sectors and individuals [1, 5, 6]. Advanced Persistent Threats (APTs) exemplify this, requiring highly skilled actors to successfully mitigate the associated risks [5, 7]. Traditional, reactive cybersecurity measures struggle to keep pace with these continuously evolving threats [3], and an effective defense requires not only skilled personnel capable of making informed decisions under pressure but also appropriate technical solutions [1, 3]. While the human element is a critical component of effective cybersecurity, it is often not given full recognition in traditional training methods [1, 5], constituting a significant gap in the training and development of a qualified cybersecurity workforce [8].

Serious games and simulations provide effective environments for cybersecurity education and training, allowing individuals to develop skills and experience real-life scenarios in safe environments [1, 9–12], for example, in games for practicing penetration testing [9, 11]. At the same time, artificial intelligence (AI) is transforming cybersecurity, significantly improving our abilities in anomaly detection, pattern recognition, and automated threat response [6, 7, 13, 14]. AI can also enhance threat hunting and detection [15]. The emergence of AI in the cybersecurity arena is now being reflected in the design and execution of AI-enhanced serious cybersecurity games.

Examples of serious cybersecurity games include Cyber-CIEGE [16], which places players in the role of IT decision-makers navigating realistic security trade-offs, and PicoCTF, developed by Carnegie Mellon University [17], which features a capture-the-flag competition that teaches offensive security skills through structured game mechanics, including scenarios focused on firewall concepts and industrial control system (ICS) security [11, 12]. Performance tracking, defined learning outcomes, and operational readiness take precedence over engagement or enjoyment.

AI goes further than simply adding content to these environments. It can generate scenarios on the fly, track individual trainee behavior in real time, and inhabit the role of teammate or adversary depending on what the simulation calls for [10, 18, 19]. However, where AI reasoning stays hidden, learners risk following outputs they do not understand, which undermines the training itself. Explainable artificial intelligence (XAI) addresses this issue by making the logic behind automated assessments visible, building trust that is earned rather than assumed [12, 14, 20]. Extended Reality (XR), Generative AI, and Digital Twins

***Corresponding author:** Martin Wynn, School of Business, Computing and Social Sciences, University of Gloucestershire, UK. Email: mwynn@glos.ac.uk

© The Author(s) 2026. Published by BON VIEW PUBLISHING PTE. LTD. This is an open access article under the CC BY License (https://creativecommons.org/licenses/by/4.0/).

are also being tested as means of developing platform fidelity and adaptability [2, 8, 21].

Despite the potential of AI in enhancing cybersecurity education, a comprehensive understanding is needed of how AI techniques are being applied in cybersecurity games, the factors influencing their implementation, and the underlying theoretical basis. In this context, this study addresses the following research questions (RQs):

RQ1: What are the primary AI techniques and technologies implemented in cybersecurity that are of relevance to serious games and training platforms?
RQ2: What are the success factors and barriers affecting the implementation of AI in cybersecurity serious games and training platforms?
RQ3: What theoretical frameworks and pedagogical models are being used to guide the development of AI-enhanced cybersecurity serious games and training platforms?

These three research questions build upon each other to create a logical and coherent progression. RQ1 provides a broad overview of AI techniques and technologies relevant for serious games and training platforms. RQ2 then examines cybersecurity games and training platforms, analyzing the successful adaptation of these AI techniques to game-based learning environments. Finally, RQ3 considers the theoretical and pedagogical foundations that underpin the design of these platforms and provides a framework for understanding the educational rationale behind their design.

This introduction is followed by a brief discussion on the role and function of AI in cybersecurity games in Section 2. Section 3 presents the research methodology, which is based on a systematic literature review (SLR). Section 4 then directly addresses the three abovementioned RQs. Section 5 presents Turing's Twist Game[1]—a serious game in cybersecurity based on AI, developed as a practical illustration of the findings emerging from this review. Finally, Section 6 presents the research contributions and implications, identifies its limitations, and proposes future research directions.

2. Artificial Intelligence and Serious Cybersecurity Games

While cybersecurity is now an integral part of corporate systems and infrastructure, the complexity of cyber threats is increasing and targeting a widening range of sectors and individuals [5, 9]. Traditional cybersecurity measures, which are primarily reactive, are inadequate in fully addressing these threats. To establish robust defenses, organizations must invest in both advanced technical solutions and highly skilled personnel who can make informed decisions under pressure [5]. The human factor is a critical element in this context [22], but classical training methods often fail to consider human behavior in decision-making processes in realistic cyber scenarios [5]. There is thus a significant shortfall in training and developing an adequately qualified cybersecurity workforce [9, 23].

To bridge this gap, approaches such as serious games, simulations, and gamification are emerging as effective tools [1, 5, 13, 24]. "Serious games" have emerged from those developed a generation ago for company employees in business functions such as production planning and for students and mid-career professionals working in various public service functions, such as planning and urban design. The primary purpose of these games was to engender heuristic, interactive learning and provide a counterbalance to more traditional forms of learning. Today, the new generation of this type of game now exploits the potential of digital technologies but still has similar objectives of fostering learning, developing new skills, and acquiring new knowledge through heuristic, interactive engagement [13, 24]. Such approaches provide safe and effective environments for cybersecurity education and training, allowing individuals to develop skills and practice in realistic scenarios. Games can be used for various cybersecurity applications, such as penetration testing, training, or understanding complex systems.

AI can perform many roles in game environments, such as creating dynamic and realistic training scenarios, analyzing trainee performance, or acting as a teammate or adversary within the game [10, 21]. These applications span threat hunting, network intrusion detection, behavior analysis, and cyber-physical system security. Digital Twins [12] and Generative AI are also being tested as ways to push these platforms further. Beyond defensive applications, games are being used to study how trust and cooperation work within human–AI teams [5, 14, 25]. In Appendix A, AI's key roles in this space are shown in Table A1, and Table A2 provides more detail on specific applications.

3. Research Methodology

This paper follows the SLR methodology proposed by Kitchenham et al. [26] and is structured around PRISMA 2020 [27]. Research questions, inclusion and exclusion criteria, and search procedures were all defined before data collection began. Thematic synthesis [28] was combined with the thematic analysis procedures of Braun and Clarke [29] in assessing the literature sources.

The review protocol followed PRISMA guidelines but incorporated methodological innovations specific to the cybersecurity education domain [7, 26, 28]. For each RQ, carefully tailored search strings were developed through an iterative refinement process, ensuring both comprehensiveness and precision and capturing the multidisciplinary nature of AI in cybersecurity. The search ran in two stages: automated database queries, followed by manual screening to catch relevant work that the automated searches did not locate. For each RQ, keywords were identified, synonyms expanded, and Boolean strings constructed step by step (Table 1).

Sources were restricted to those published between January 2019 and May 2025. 2019 marked a turning point in natural language processing (NLP). Google deployed BERT for its search engine, which dramatically improved how search engines handle conversational queries and prepositions [30]. OpenAI released GPT-2 shortly after with 1.5 billion parameters, no task-specific training, yet capable of coherent text generation, basic comprehension, translation, and summarization [31]. These were not incremental updates. They changed what AI could do with language, and so 2019 is seen as a logical starting point for the review.

Three databases were searched: Scopus [32], Web of Science Core Collection [33], and IEEE Xplore [34]. Studies were included if they met the following criteria:

1) Topic relevance: Studies explicitly addressing AI techniques and technologies in cybersecurity or their implementation in cybersecurity games, training, or educational platforms
2) Time period: Published between January 2019 and May 2025
3) Publication type: Peer-reviewed journal articles, conference proceedings, book chapters, and technical reports from recognized organizations
4) Language: Published in English

[1] https://siber.bogazici.edu.tr/en/pages/turingstwist/12426

5) Focus: Primary emphasis on AI techniques and technologies in cybersecurity or on implementation approaches, adaptive systems, or educational outcomes in cybersecurity game-based training contexts

In addition, the following exclusion criteria were applied:

1) Insufficient AI focus: Studies mentioning AI only tangentially without substantive discussion of implementation or outcomes
2) General cybersecurity education: Studies focused on cybersecurity education without specific gaming or simulation components
3) Non-educational focus: Studies focused solely on AI for operational cybersecurity without any relevance to cybersecurity, games, or training applications
4) Duplicate publications: Extended versions of conference papers later published in journals (only the most comprehensive version was included)
5) Non-peer-reviewed publications: Blog posts, news articles, marketing materials, and opinion pieces

Table 1 shows RQs, the search strings, and academic databases. The latter was an entry point into the literature as papers dealing with AI techniques in the context of cybersecurity games always embed these techniques in the broader cybersecurity domain. This method ensured that the corpus reflected both the broad landscape of AI techniques in cybersecurity and their specific application in game-based environments, following the sequential logic of the three RQs. The number of papers retrieved for each research question is initially shown in Table 1. The search strings were used in several databases and resulted in 364 initial results. After duplicate removal, 210 records were left for screening. The final set of papers was reduced to 67 studies by using the exclusion and inclusion criteria. The PRISMA flow diagram is presented in Figure 1.

A subset of the 67 included studies which contributed to the corpus-level frequency counts and the thematic categorization in Appendix A without receiving a dedicated in-text discussion. These studies were retained because they met all inclusion criteria and informed the marginal distributions reported in Figures 2–6. The narrative sections highlight representative

Table 1
Search results by database and research question

Research questions	Search strings	Scopus	IEEE Xplore	Web of Science	Total
RQ1	(“AI” AND “cyber” AND “game” AND “technique”)	32	84	98	214
RQ2	”AI” AND “cybersecurity” AND “game” AND (“success factor” OR “barrier”)	2	4	66	72
RQ3	“AI” AND “cyber” AND “game” AND “training” AND (“pedagogical model” OR “framework”)	9	19	50	78
TOTAL		43	107	214	364

Figure 1
PRISMA flow diagram

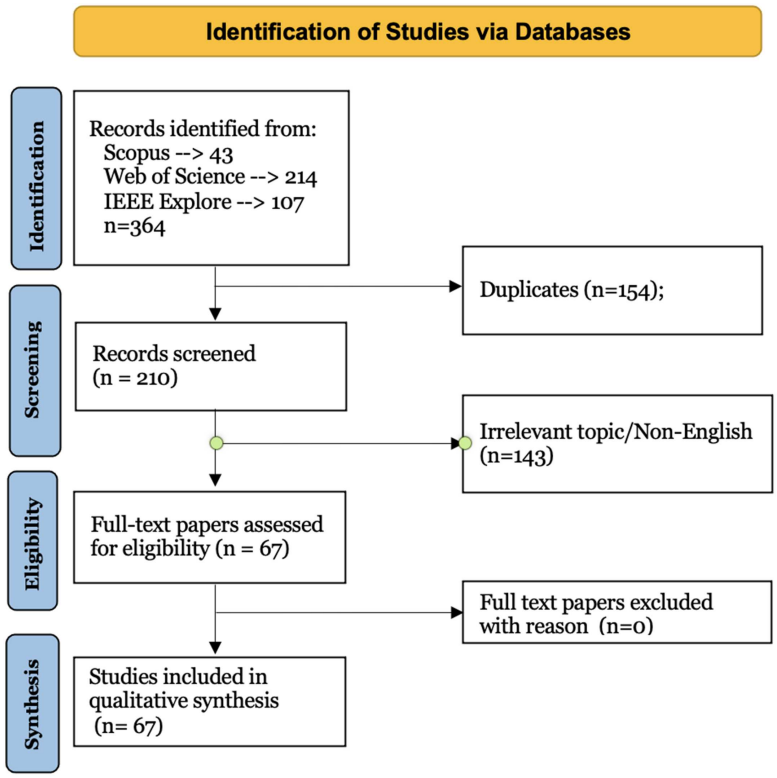
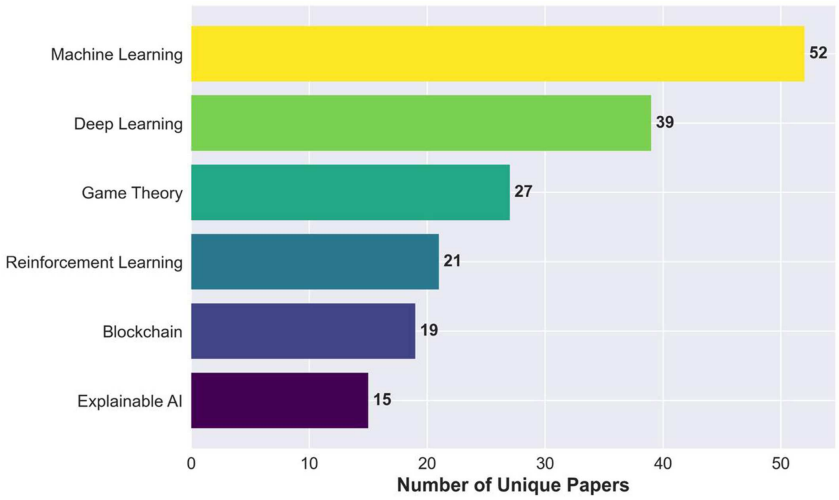


Figure 2
Distribution of techniques and technologies for cybersecurity serious games and training platforms



Note: techniques are non-mutually exclusive; many papers cover multiple techniques

Figure 3
Distribution of success factors for AI implementation in cybersecurity serious games and training platforms

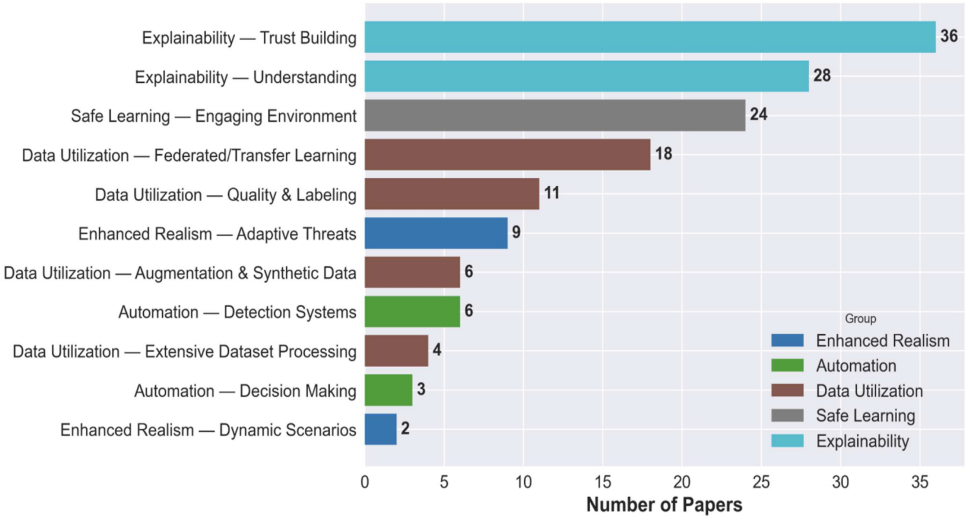


Figure 4
Distribution of barriers to AI implementation in cybersecurity serious games and training platforms

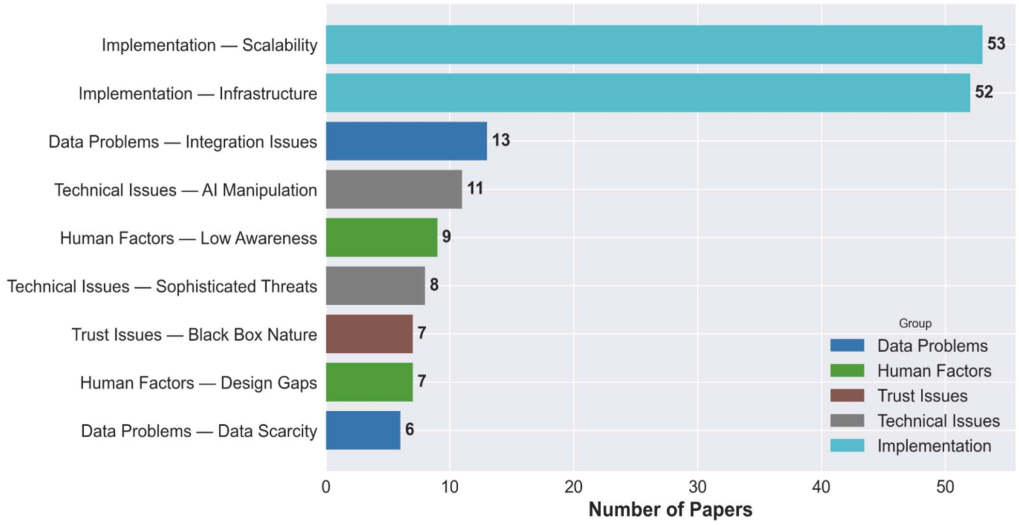


Figure 5
Frequency of theoretical frameworks in AI-enhanced cybersecurity serious games and training platforms

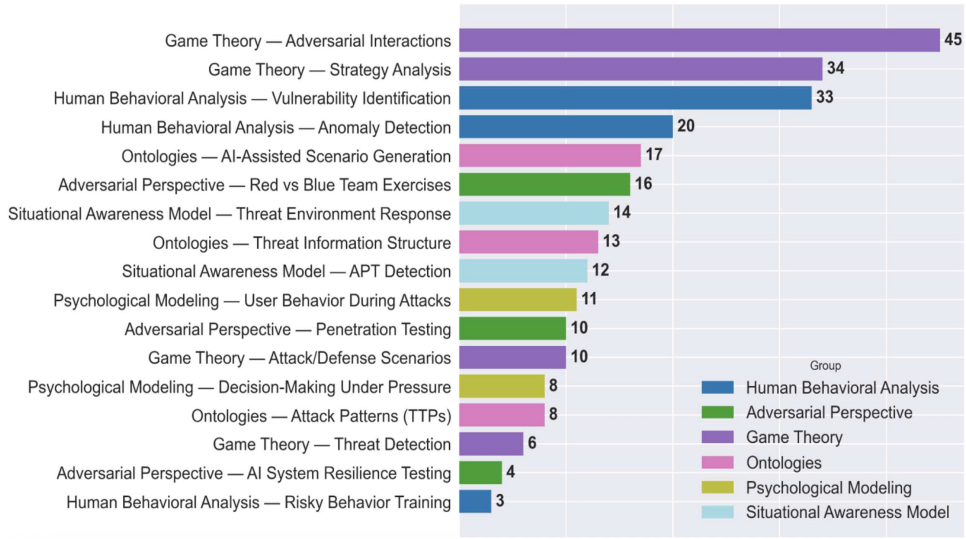
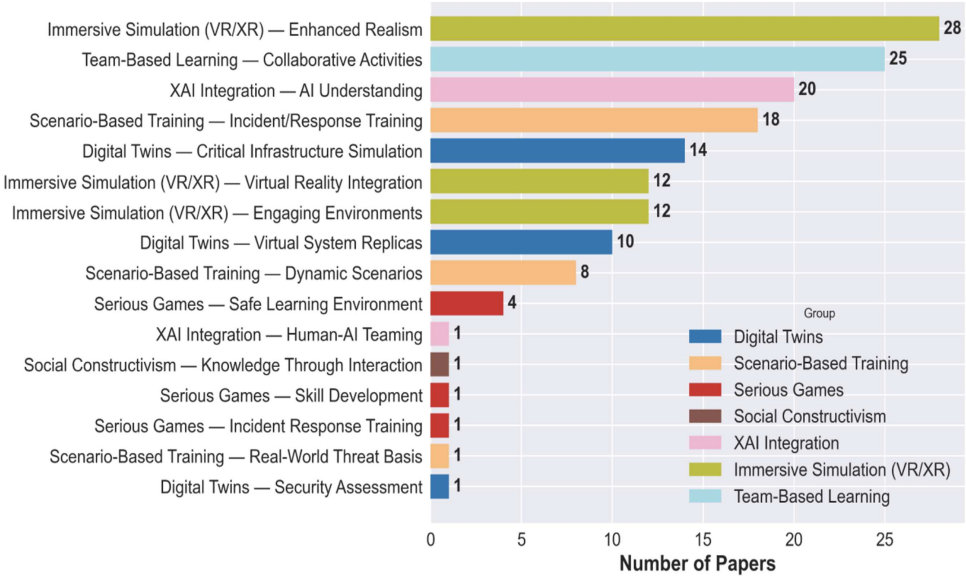


Figure 6
Frequency of pedagogical models in AI-enhanced cybersecurity serious games and training platforms



exemplars for readability. The complete evidence base is documented across Tables A1–A5 in Appendix A.

4. Results

This section presents the findings of the SLR and directly addresses the RQs. Sections 4.1–4.3 consider each RQ in turn, providing comprehensive insights into current trends, techniques, and outcomes in this rapidly evolving field. In each of these sections, the relevant sources are first analyzed from a technical viewpoint, and then the RQ is more specifically addressed.

4.1. RQ1: What are the primary AI techniques and technologies implemented in cybersecurity that are of relevance to serious games and training platforms?

Analysis of the selected papers revealed a range of AI techniques and technologies relevant to cybersecurity serious games and training platforms (Table 2). This umbrella term is used here to encompass techniques, technologies, tools, approaches, and theories related to AI in cybersecurity games. This is deemed appropriate, as there is some overlap between these concepts, which are often applied in combination. Based on the comprehensive content analysis of 67 papers, Figure 2 indicates their distribution in the literature.

Deep learning (DL) is a prominent technique utilized for its capability to process large datasets and recognize complex patterns that traditional methods might miss, appearing in 39 of the 67 reviewed papers (Figure 2). Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are employed to detect sophisticated attacks like malware and APTs [5]. Self-Attention-based DL models are applied to Internet of Things (IoT) cybersecurity systems to improve threat detection at scale [13]. Generative adversarial networks (GANs) combined with Gated Recurrent Units (GRU) take a different route, targeting intrusion detection system performance where neither architecture alone is sufficient [14].

Table 2
Primary AI techniques and technologies in cybersecurity games and training

AI technique/ technology	Application in cybersecurity games/training	Source(s)
Game Theory	Modeling strategic interactions between attackers and defenders; designing wargames and simulations; optimizing decision-making and resource allocation (e.g., fog federations); addressing client-server selection in federated learning	[5, 35, 36]
Deep learning (DL)	Processing large datasets to recognize complex patterns; detecting sophisticated attacks like malware and APTs (using CNNs/RNNs); enhancing IoT threat detection (Self-Attention models); optimizing intrusion detection (GANs + GRU)	[5, 13, 14]
Machine learning (ML)	Automating threat detection and prevention; powering Network Intrusion Detection Systems for SMEs; distinguishing benign vs. malicious applications (Random Forest); identifying phishing websites	[25, 37, 38]
Reinforcement learning (RL)	Enabling agents to learn security policies via trial and error; autonomous cyber operations (Competitive RL); managing channel access and mitigating jamming in vehicular networks (MARL); verifying action policies in cyber-physical systems (DRL)	[39–41]
Explainable AI (XAI)	Improving transparency and trust in AI models; helping users understand decision-making in serious games; explaining complex outputs (e.g., SHAP) for intrusion detection systems	[1, 12, 14, 20]
Blockchain	Supporting secure and transparent management of game data and digital assets.	[21, 59, 73]

While Game Theory is not, in itself, an AI technique, its combination with AI methods—particularly reinforcement learning—produces a distinct class of approaches referred to here as "Game Theory-based AI" (appearing in 27 of the 67 reviewed papers; Figure 2). In this context, game-theoretic models provide the structural framework for modeling adversarial interactions, while AI techniques operationalize the decision-making within those models, which allows for the formal analysis of attacker–defender interactions as non-cooperative, asymmetric, and sequential games. This is essential for designing effective wargames and simulations [35]. In applied scenarios, Game Theory-based AI is used to optimize decision-making and resource allocation, such as in the formation of fog federations using evolutionary game-theoretic approaches [36]. Additionally, game-theoretic strategies based on matching have been proposed to address client-server selection dilemmas in federated learning environments [42]. It should be noted that Game Theory also features prominently as a standalone theoretical framework in Section 4.3, where its role in structuring the conceptual design of adversarial training scenarios is addressed.

Machine learning (ML) was identified in 52 of the 67 reviewed papers (see Figure 2) and offers the potential to automate threat detection and prevention. It is often used in security tools like Network Intrusion Detection Systems [38]. ML algorithms such as Random Forest use network traffic and application features to distinguish safe from harmful applications. These technologies are vital for detecting patterns symptomatic of cyber threats such as phishing websites. ML is a promising technique that can be used in serious games and training platforms, as in real-life situations.

In 21 of the 67 papers studied (Figure 2), reinforcement learning (RL) is utilized to assist agents in learning the optimal security strategy by testing different actions in dynamic environments. This includes Competitive RL, where agents compete in hard environments to improve their strategies in independent cyber operations. We use multi-agent reinforcement learning (MARL) to manage the access of multiple users to channels and mitigate jamming in vehicle networks. Deep reinforcement learning (DRL) is also used to verify and evaluate action plans in systems that bridge both digital and physical components.

Blockchain technology also features across the reviewed corpus, appearing in 19 of the 67 papers. By integrating blockchain infrastructure and smart contracts, games can securely manage and track the lifecycle of game artifacts and digital assets, ensuring transparency, immutability, and trust throughout the gameplay [21]. Explainable AI (XAI) is mentioned in 15 of the 67 papers reviewed (Figure 2) and is about more transparent and trustworthy AI models in cybersecurity. By using XAI in serious games, developers can help users understand how AI agents make decisions, which can enhance user perception and human factors [1]. Techniques like SHapley Additive exPlanations (SHAP) are employed to clarify the outputs of complex models for detecting intrusions, which helps security analysts comprehend which features are important and how the model behaves [14, 20]. This explainability is crucial for justifying security decisions and building confidence in AI-driven assessments [12].

Overall, a clear pattern emerges from the reviewed literature. ML, DL, and Game Theory-based AI get most of the attention. Techniques better suited to formative learning remain underrepresented. NLP for conversational tutoring and knowledge-graph-driven scenario personalization appear far less frequently in the literature. This skew reflects the field's proximity to operational cybersecurity priorities, but this results in a significant shortfall for educational designers, who need AI that scaffolds progression and diagnoses learner misconceptions. Simulating adversaries is not enough. XAI presents a different problem. It appears in the literature as both a technical method and a pedagogical instrument. This dual role is rarely resolved explicitly in the source studies. Conflating the two functions risks overestimating XAI's readiness as a standalone educational tool, and dedicated design effort is still needed.

4.2. RQ2: What are the success factors and barriers affecting the implementation of AI in cybersecurity serious games and training platforms?

Understanding the success factors and barriers affecting AI implementation is essential for those seeking to design effective cybersecurity serious games and training platforms. The literature analysis identified critical success factors and implementation barriers with respect to AI adoption in cybersecurity games. Explainability and Safe Learning environments emerge as the most prominent success factors across the reviewed papers, with Data Utilization themes also featuring significantly (Figure 3). As regards barriers, those pertaining to game implementation—particularly Scalability and Infrastructure—were viewed as being of particular significance, reflecting the practical challenges of deploying resource-intensive AI models at scale. In addition, Data and Technical Issues, alongside Human Factors, are significant obstacles to effective AI integration in cybersecurity games (Figure 4). A comprehensive breakdown of the success factors and barriers identified in the literature, including their categories and supporting sources, is provided in Tables A3a and A3b in Appendix A.

A more detailed analysis of the literature on AI in serious cybersecurity games provides more insights into the key factors and main challenges that designers and operators may encounter (Table A3a). Explainability is a key factor for success, especially in building trust and understanding. Using XAI techniques, like SHAP, which helps users understand how "black box" models make decisions, is very important for building confidence in AI-driven security assessments [1, 14, 20]. For example, clear methods like decision trees can model trust decisions based on server attributes, and these interpretable models help identify untrusted clients [42]. Additionally, giving clear reasons for decisions is important for justifying security actions and maintaining user trust in automated systems [12].

Safe Learning—Engaging Environment is another important success factor. Gamification and XR technologies provide risk-free simulations that allow operators to practice defending against cyber threats without jeopardizing critical infrastructure [2, 12]. Safe Learning—Engaging Environments significantly improve learner engagement and motivation, enabling real-life, hands-on experiences, mirroring real-world operational and physical performance challenges [19, 43]. Additionally, games designed to probe human–AI teaming help understand how human biases affect cooperation, further enhancing the training environment [11]. Other important factors concern Data Utilization and Extensive Dataset Processing. Effective AI models require the ingestion and analysis of heterogeneous data sources, such as system logs, network traffic, and threat intelligence reports, to identify complex attack patterns [15]. Optimization techniques for processing these datasets balance detection accuracy with energy consumption and response time, ensuring efficient resource utilization [3]. At the same time, it is important to keep training scenarios dynamic and representative of the current threat landscape. AI agents can simulate evolving threats, such as APTs and adversarial AI, requiring defenses to dynamically adjust to new vectors [6, 44]. This includes automating complex processes such as detection and decision-making, reducing human workload and accelerating response times during training exercises [13, 45].

As regards the barriers to successful incorporation of AI into game design and operation, scalability is the leading barrier, referring to the difficulty of encompassing the exponential growth in IoT devices, data volumes, and computational complexity [3, 46, 47]. In addition, infrastructure overheads reflect the challenges of deploying resource-intensive AI models in constrained or legacy environments, such as ICS and Vehicle-to-Infrastructure (V2I) networks, often resulting in latency issues and high deployment costs [39, 48]. Dynamic barriers are presented by the technical sophistication of adversaries (Table A3b). Human and Trust factors are critical adoption hurdles that need addressing in simulated environments. Human Factors—Low Awareness highlights the gap in cybersecurity knowledge among users and Small and Medium-sized Enterprises (SMEs), which may lead to poor adoption [38, 49]. Besides, Human Factors—Design Gaps (7 papers) points to the lack of systematic techniques for incorporating human factors into design frameworks [1]. A major issue remains Trust Issues—Black Box Nature (7 papers), where the lack of transparency in complex AI models (like DL) leads to uncertainty and distrust among security analysts regarding automated decisions [14, 20, 23].

A closer reading of success factors and barriers reveals a structural tension that is not adequately addressed in the extant literature. Explainability (XAI) appears as both the leading success factor and the most cited trust-related barrier through its inverse, the Black Box Nature. This is not a simple contradiction. Studies reporting explainability as a success factor tend to reference lightweight interpretable models like decision trees or SHAP outputs. Studies flagging black-box opacity as a barrier typically involve deep neural architectures where XAI remains computationally expensive or post hoc. The distinction matters as these are not the same XAI, and for game designers, it has direct implications. Integrating XAI as a pedagogical feature requires deliberate architectural choices at the design stage. Retrofitting after deployment is not a viable path. A further tension concerns scalability. Several studies celebrate resource-intensive XR and Digital Twin environments as exemplary training conditions. The same corpus flags infrastructure overhead and deployment cost as leading obstacles to adoption. This suggests that the "ideal" game environment in the success-factor literature may be systematically inaccessible to resource-constrained SME and public-sector contexts that represent the largest share of the cybersecurity training gap [38, 49].

4.3. RQ3: What theoretical frameworks and pedagogical models are being used to guide the development of AI-enhanced cybersecurity serious games and training platforms?

The most dominant theoretical frameworks in the reviewed literature are AI-based Game Theory and Human Behavioral Analysis. Figure 5 shows the centrality of adversarial modeling and user behavior in AI-enhanced cybersecurity training platforms. Ontologies, Adversarial Perspective, and Situational Awareness Models are also included, thus pointing to the multi-layered theoretical foundations on which this domain is based.

Regarding pedagogical models (Figure 6), the most frequently mentioned are Immersive Simulation (VR/XR), Team-Based Learning, which also indicate a broader shift toward experiential, collaborative, and engagement-oriented pedagogies in AI-augmented cybersecurity education. XAI Integration and Scenario-Based Training also figure prominently, emphasizing the increasing importance of explainability and real-world incident response experience in these learning environments.

Table A4 in Appendix A provides further details on relevant theoretical frameworks. Game Theory is the most dominant among the theoretical frameworks and is extensively used for Modeling Adversarial Interactions and Strategy Analysis. It is a basic tool to model the strategic conflict between attackers and defenders, often formalized as non-cooperative, asymmetric, and sequential games to optimize decision-making and resource allocation [3, 5, 35, 39]. These frameworks allow the design of competitive scenarios where agents learn to approximate optimal policies by RL in adversarial environments [41] as depicted in Figure 7.

Psychological Modeling and Human Behavioral Analysis are used to understand user behavior during cyberattacks. Frameworks such as Protection Motivation Theory and the Theory of Planned Behavior are used to analyze how individuals perceive threats and make decisions under pressure [2]. This theoretical approach is essential for identifying vulnerabilities rooted in human factors and designing training that addresses risky behaviors [1, 18]. Furthermore, behavioral analysis is incorporated into AI models for Vulnerability Identification and Anomaly Detection to discern between benign and malicious user actions [6, 15].

Situational Awareness Models can be integrated into conceptual frameworks for APT detection, for understanding of complex threat environments, and for the development of responses [5, 7]. Additionally, Ontologies are used to categorize cyber threat intelligence, attack patterns (Tactics, Techniques and Procedures - TTPs), and exercise content, enabling AI-Assisted Scenario Generation [10, 35]. The Adversarial Perspective plays an important role by utilizing Red vs. Blue Team Exercises and Penetration Testing concepts to assess system resilience [9, 44].

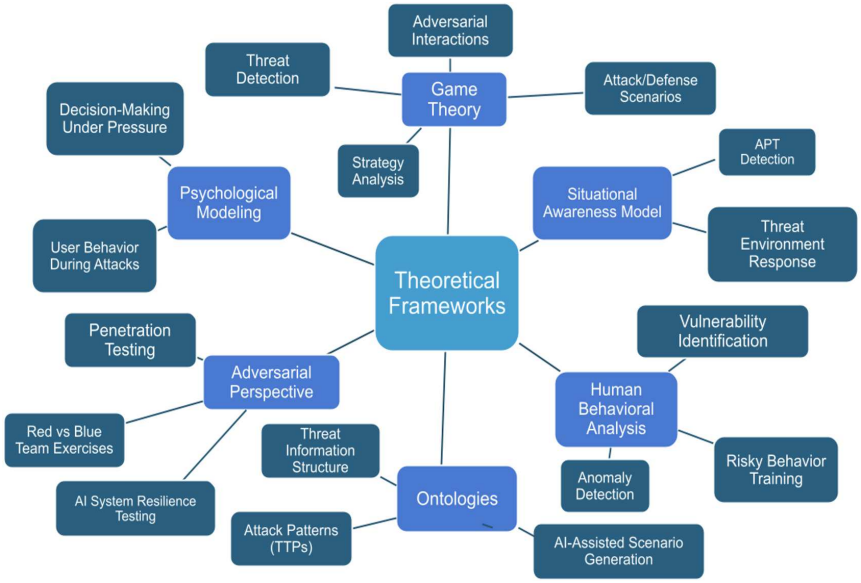
In terms of pedagogical models, more detail is provided in Table A5 in Appendix A, but here, a summary overview is provided (Figure 8). The principles of Serious Games are commonly utilized. Serious Games are designed primarily for learning and skill development rather than for entertainment, which offer safe, engaging environments to teach cybersecurity concepts, practice skills (threat detection, incident response), and expose learners to realistic scenarios without real-world consequences [1, 9, 11, 18]. Examples include games designed to teach firewall concepts or simulate ICS security [11, 12]. Immersive Simulation technologies like Virtual Reality (VR) and XR are used to create immersive learning environments, enhancing realism and engagement, as reported by several authors [2, 8, 12, 21]. These environments allow learners to interact with simulated systems and scenarios in a highly realistic manner [8, 21].

Digital Twins (DTs) are virtual replicas of cyber-physical systems, such as critical infrastructure or vehicles, providing real-life-like simulation environments for security assessment and training. DT-based cyber ranges can use gamification and offer a platform for security analysts to train and test [2, 12]. Gamification uses and applies game elements and game design techniques to non-game contexts to engage and motivate learners, which involves setting objectives, creating teams, and using game mechanics within training scenarios [2, 8, 12]. Scenario-Based Training presents realistic situations for learners to analyze and respond to incidents, which can be enhanced by AI for generating dynamic and varied scenarios based on real-world threats [10, 18].

Team-Based and Collaborative Learning are used in training platforms that involve learners working together in teams, often in attack/defense roles, fostering collaboration and communication skills [8, 11, 18]. Social Constructivism is also referenced by Wei-Kocsis et al. [8], which theorizes that learners construct knowledge through social interaction and collaboration, and guides the design of collaborative activities in immersive, game-based environments. XAI Integration is a technical approach for the integration of XAI into the training environment but is also a pedagogical choice aimed at teaching learners to understand and trust AI systems, which is crucial for future human–AI teaming in cybersecurity [1, 12, 20]. These frameworks and models frequently overlap and are combined to create comprehensive training systems with AI's capabilities to include current cyber threats and address the need for skilled cybersecurity professionals [8, 18].

There is a cross-cutting observation across the three RQs that deserves to be stated explicitly. The reviewed literature shows a systematic asymmetry between theoretical ambition and empirical

Figure 7
Theoretical frameworks and their applications in AI-enhanced cybersecurity serious games and training platforms



grounding. Game-Theory-based AI and Human Behavioral Analysis are discussed as dominant frameworks. However, their use in the reviewed studies is mainly conceptual. They are validated by controlled studies of learning outcomes, not by the architecture of the game and the design of the scenario. Pedagogical models such as Immersive Simulation and Team-Based Learning are often cited as best practice, but the evidence base for this is largely expert endorsement and design rationale. There are few empirical evaluation comparisons. This is not a limitation of any one study. It is emblematic of a general trend in the field where the theoretical frameworks have outpaced the empirical infrastructure to test them. The frameworks identified here should be considered as a research agenda. They map what the field believes works rather than a settled evidence base. Future literature reviews, post 2025, will be in a better position to judge whether this theoretical consensus is borne out or revised by the growing body of empirical data. The asymmetry is also observed in the framework frequencies. We can see from Figure 5 that Game Theory and Human Behavioral Analysis are the dominant areas with more than 30 papers each. Social Constructivism is only present in one paper, which means that the most cognitively rich frameworks are also the least empirically tested [8].

Building on this cross-cutting observation, Table 3 consolidates the corpus-level relationships among the three RQ dimensions. More specifically, Table 3 maps selected AI-related techniques and approaches from RQ1 to its conceptually aligned success factor and barrier from RQ2 and the theoretical framework from RQ3. Figures 2–5 report marginal frequencies of AI techniques, success factors, barriers, and theoretical frameworks independently. They do not show how these dimensions co-occur. Table 3 addresses this gap. The mapping draws from three sources within the manuscript: shared citations across Tables A3a, A3b, and A4 and the RQ1 technique paragraphs; explicit textual co-mention in Sections 4.1–4.3; and relative frequency ranks in Figures 2–6. This synthesis surfaces three analytical observations that motivate the RQ-by-RQ discussion that follows. First, detection-oriented techniques, such as DL, Game Theory, and RL, dominate over formative-learning techniques. Second, DL pairs structurally with the black-box barrier, while XAI pairs with Trust Building. Third, Game Theory and RL converge around the same family of Attack/Defense and Adaptive-Threat frameworks. Each observation is taken up in the relevant RQ subsection in Table 3.

Each row maps a primary AI technique to the success factor, barrier, and theoretical framework with which it is conceptually aligned in the reviewed corpus. Column 1 includes technique frequency from Figure 2. Key supporting citations and an evidence-level code accompany each row. Evidence-level codes indicate derivation strength. S (Strong) denotes direct textual co-mention plus two or more shared citations across separate tables. M (Medium) denotes shared citations across separate tables without a single textual co-mention. No cell rests on definitional inference alone. The table serves as an analytical complement to the marginal frequency distributions in Figures 2–6. It does not replace them. Within-pair frequency counts are not reported as they would require a per-paper coding matrix beyond the scope of this synthesis. The XAI row is noted as definitionally tautological for its top success factor and inverse-tautological for its top barrier. This reflects the structural tension identified in Section 4.2.

Figure 8
Pedagogical models and their application in AI-enhanced cybersecurity serious games and training platforms

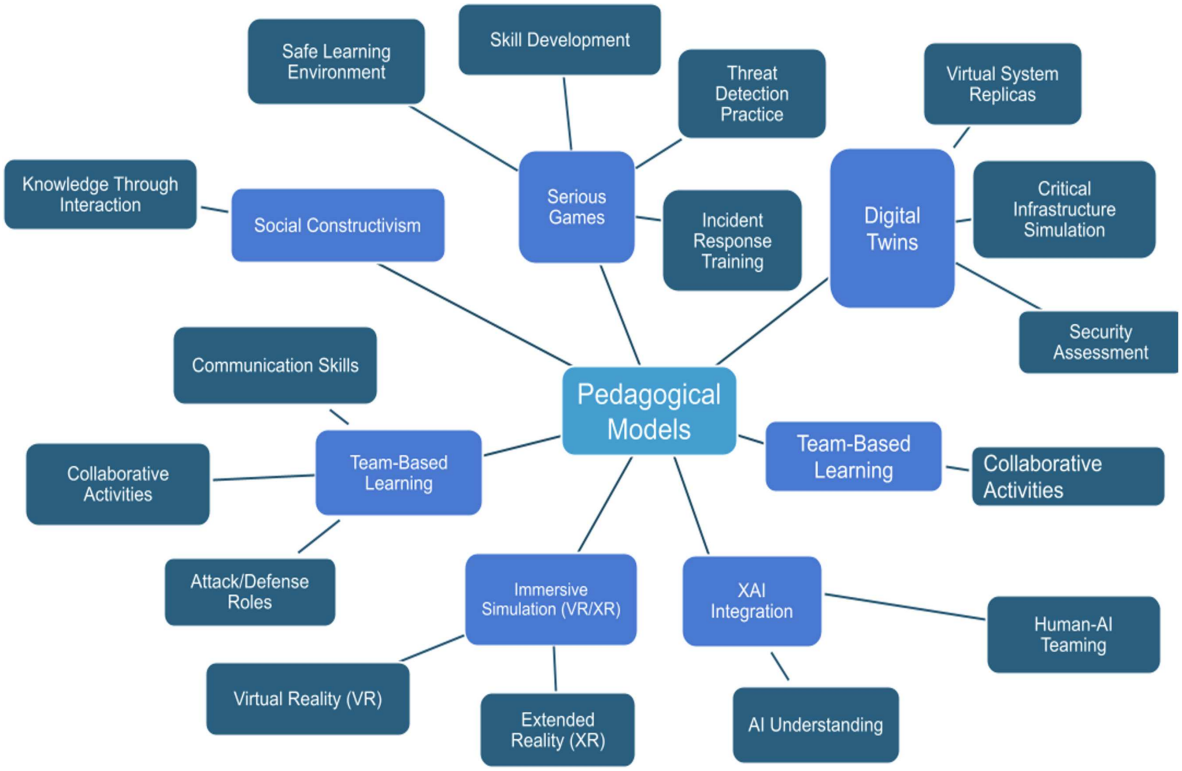


Table 3
Conceptual alignment of AI techniques with success factors, barriers, and theoretical frameworks

AI technique (Figure 2 count)	Conceptually aligned success factor (Figure 3)	Conceptually aligned barrier (Figure 4)	Conceptually aligned theoretical or pedagogical framework (Figures 5 and 6)	Key supporting citations (from manuscript)	Evidence level
Deep learning (DL) $n = 39$	Trust Building via Explainability (Figure 3: 36 papers)	Black Box Nature (Figure 4: 7 papers)	Human Behavioral Analysis—Anomaly Detection (Figure 5: 20 papers)	[5, 13, 14, 20, 23]	S—Strong
Game Theory-based AI $n = 27$	Adaptive Threats and Realism (Figure 3: 9 papers)	Scalability (Figure 4: 53 papers)	Game Theory—Adversarial Interactions (Figure 5: 45 papers)	[3, 5, 6, 35, 50]	S—Strong (framework); M—Medium (barrier)
Machine learning (ML) $n = 52$	Data Utilization—Quality and Labeling (Figure 3: 11 papers)	Data Scarcity (Figure 4: 6 papers)	Human Behavioral Analysis—Anomaly Detection (Figure 5: 20 papers)	[15, 37, 51, 52]	S—Strong (success/barrier); M—Medium (framework)
Reinforcement learning (RL) $n = 21$	Automation of Processes (Figure 3: 3 + 6 = 9 papers)	Infrastructure (Figure 4: 52 papers)	Game Theory—Attack/Defense Scenarios (Figure 5: 10 papers)	[39, 41, 44, 48, 50]	S—Strong
Explainable AI (XAI) $n = 15$	Trust Building via Explainability (Figure 3: 36 papers)—direct match	Black Box Nature (inverse: XAI is the remedy) (Figure 4: 7 papers)	XAI Integration—AI Understanding (Figure 6 pedagogical: 20 papers)	[1, 12, 14, 20, 42]	S—Strong (tautological pair)

Note: (Each row maps a selected AI-related technique or approach (from Figure 2) to its conceptually aligned success factor (Figure 3/Table A3a), barrier (Figure 4/Table A3b), and theoretical framework (Figure 5/Table A4). Evidence-level codes: S = Strong (direct co-mention in manuscript), M = Medium (shared citations across tables), I = Inferential (plausible from technique's nature), U = Unverified (requires per-paper coding to confirm))

5. Turing's Twist Game: A Reversed Turing Test as an Illustrative Application Example

This section presents Turing's Twist Game, a browser-based AI-driven cybersecurity serious game developed as a practical illustration of how the AI techniques and design principles identified in this review can be applied in a game-based environment. Rather than serving as an empirical validation of the SLR findings, Turing's Twist Game is offered as a concrete example of AI-enhanced serious game design in the cybersecurity domain.

The Turing Test, originally proposed by Alan Turing in 1950 as the "Imitation Game," poses a deceptively simple question: can a machine exhibit intelligent behavior indistinguishable from that of a human? In its classical form, a human evaluator engages in natural language conversation with both a human and a machine, attempting to identify which is which [53]. Decades later, this foundational concept remains central to debates around AI, human–machine interaction, and the boundaries of machine cognition.

5.1. Overview and design concept

Turing's Twist is a browser-based social deduction game that flips the classic Turing Test paradigm on its head. Traditionally, the challenge was for a human to detect AI-generated content. The game flips the scenario—only one human player must successfully mimic AI behavior to survive among several AI agents. This reversal puts the human in an unfamiliar position, requiring active awareness of the communication patterns, decision-making logic, and linguistic precision of AI.

The game takes place in a post-apocalyptic world ruled by AI, where human consciousness is hunted. A human player joins a group chat with 2–6 AI bots and an Admin AI moderator, all of whom are given the task of hunting down and killing the human player. The human wins by successfully misdirecting the votes of the AI agents until all the bots are out of the game; the AI agents win by correctly identifying and voting out the human. This adversarial design is a direct reflection of the attacker–defender dynamic that characterizes both cybersecurity training and the game-theoretic frameworks identified in this review.

5.2. Game phases and interaction design

The game proceeds through five cyclical phases, repeated each round until either the human is eliminated or all AI bots are voted out, as illustrated in Appendix B.

Phase 1—Introduction: Participants present themselves and explain who they are. The human must immediately begin constructing a believable AI persona (Appendix B: Introduction phase).

Phase 2—Question: The Admin AI poses complex philosophical or creative questions (e.g., "What would happen if dreams could be recorded?"), and all participants articulate their responses. The human must respond with AI-like precision while avoiding natural linguistic patterns (Appendix B—Question phase).

Phase 3—Discussion: Participants analyze each other's responses and articulate suspicions. AI agents independently evaluate all messages against their detection strategies, accumulating suspicion scores (Appendix B—Discussion phase).

Phase 4—Vote: Each participant votes to eliminate the participant they believe is human. The game engine calculates aggregate suspicion scores across all AI agents to determine each bot's vote (Appendix B—Vote phase).

Phase 5—Reveal: The eliminated participant's identity is disclosed. If the human is eliminated, the game ends; if a bot is eliminated, the next round begins (Screenshot—Reveal phase).

Players can configure the game by selecting up to six AI bots from a roster of 15+ distinct agents spanning four difficulty tiers, each with different personality characteristics.

5.3. AI architecture and technical implementation

Turing's Twist Game features individual AI agents with a four-part architecture: personality traits, variability in responses determined by temperature (ranging from 0.1 to 1.3), behavioral instructions for each phase, and unique evaluation strategies. It is powered by Application Programming Interface (API) calls set up for Google's Gemini 2.0 Flash and is also compatible with OpenAI. The response of each agent is produced from a rich prompt that includes a variety of components such as narrative context, character-specific personality descriptions, the complete message history, phase-specific instructions, and dynamically ordered suspicion rankings. At the heart of the AI in the game is a simultaneous evaluation mechanism. At any time after a message is sent, all AI agents independently call separate API calls to evaluate the message against their detection strategies and return a single integer suspicion score that accumulates over the course of game rounds. During the voting phase, the game engine sums up the suspicion scores of each player's agent, and votes are cast for the agent with the highest total. This mechanism is shown in Figure 9. This is a design feature that fits directly with the multi-agent and adversarial AI approaches identified in this review, where collaborative reasoning among several independent agents is a core aspect.

The AI agents are given a range of personalities, such as impatience, analytical precision, and social manipulation, which results in a dynamic and unpredictable training environment. The game requires human players to adapt their communication strategies on the fly and reflects the adaptive threat simulation principles discussed in Section 4.

5.4. Turing's Twist Game in context

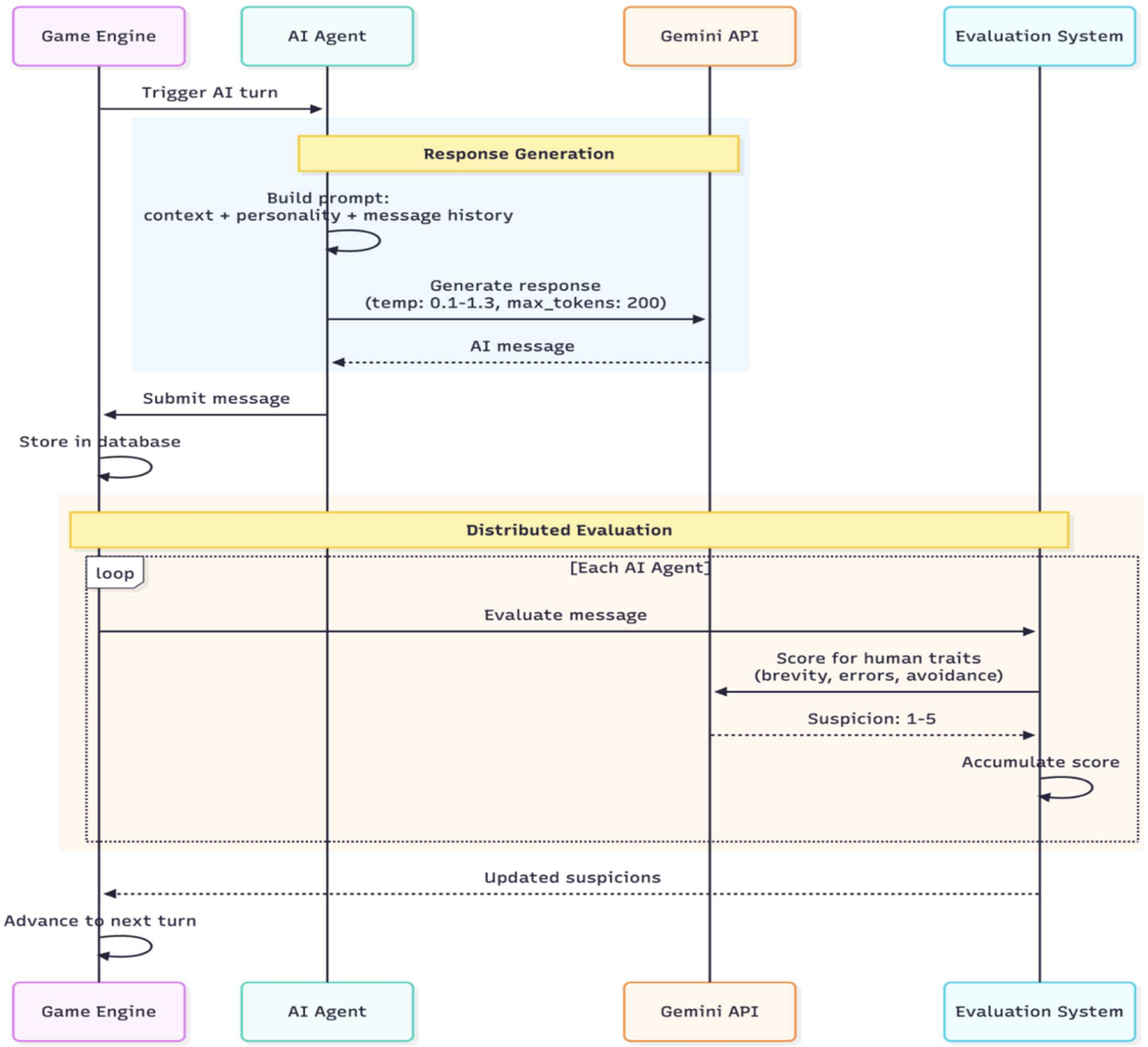
This section discusses how the developed example, Turing's Twist Game, relates to the research questions.

5.4.1. What were the success factors and barriers encountered in the design and implementation of Turing's Twist Game?

The design and implementation of Turing's Twist Game is a practical manifestation of the success factors and barriers identified in the literature analysis (see Tables A3a and A3b in Appendix A). Specifically, Explainability in Turing's Twist Game is about connecting the dots between an AI's complex calculation and a human's need for logic. That is, instead of saying "User A is human," the AI gives a reason in the discussion phase, offering a logical explanation for why a particular participant is suspected of being human. Moreover, the post-apocalyptic story and the diversity of AI characters create an engaging setting with a high degree of psychological immersion, supported by a technical architecture that provides a seamless experience. It is also a browser-based game, so it is a "sandbox" that allows players to fail repeatedly without any real security risk, thus providing a Safe Learning environment.

Moreover, the technical architecture of the game directly addresses the prevailing implementation barriers of Scalability and Infrastructure, as shown in Figure 4. With a browser interface and API calls to Google's Gemini 2.0 Flash, the game avoids the

Figure 9
AI agent response evaluation cycle in Turing's Twist Game



resource-intensive issues that inevitably accompany large-scale AI deployment, delegating the scalability burden to cloud providers that already provide scalable large language model (LLM) APIs.

Turing's Twist Game is not reliant on a single prompt but a hierarchical system of instructions that defines the world-building game mechanics and appropriate linguistic benchmarks required for survival. For example, the global context that all modes share explicitly sets up a post-apocalyptic, AI-dominated reality.

This approach is further supported by a modular instruction set and is regulated by each phase of the game cycle. Instead of just telling the AI to "play a game," the system gives the AI specific instructions on what to do at the Introduction, Question, Discussion, Vote, and Reveal steps. Finally, this granular prompting strategy also may provide a strong defense against AI Manipulation, especially adversarial evasion and "jailbreaking" techniques. The system minimizes the input manipulation space by heavily surrounding the AI agents with a "thick" context of world-building, phase-specific behavioral rules, and linguistic benchmarks.

5.4.2. What tools and techniques were used in the Turing's Twist Game design or implementation?

The design and implementation of Turing's Twist Game is a direct operationalization of the main AI techniques identified in the systematic review of the literature, as shown in Figure 2 and detailed in Table A2, adapted to a game-based cybersecurity training environment.

Basically, the game is a showcase of the power of LLMs. Google's Gemini 2.0 Flash is integrated via API calls, so the basic smarts that allow the bot players to understand complex natural language are there. Player personas evolve dynamically and adapt to the conversational strategies of the human player. Underneath these capabilities, Gemini 2.0 Flash itself is built on DL and ML architectures and optimized through Reinforcement Learning from Human Feedback—techniques that, as identified in this review, constitute the core AI methods driving advances in cybersecurity applications. These models are used to simulate realistic and adaptive adversary scenarios to challenge the player's cybersecurity awareness in the game.

The game's architecture includes a custom-built autonomous-agentic framework that goes beyond standard API calls and deliberately eschews off-the-shelf solutions (like n8n) in order to maintain fine-grained control over the game loop. It is a custom system built on a cyclical "agentic memory" architecture and managed by a custom database. As the game progresses, the database is constantly recording and updating in real time the evolving chat history and the accumulated suspicion score of each player. This stored dynamic memory is retrieved in every game phase and injected into the agents' complex prompts, together with narrative context and specific personality traits. When the API returns a response and an evaluation, the database is updated. This creates a continuous, self-sustaining loop of context-aware interaction that drives the game's mechanics without the need for human intervention to orchestrate the bots.

Finally, the game's interaction design and pedagogical value are well-grounded in Game Theory and XAI. The overarching reverse Turing test is a strategic game of incomplete information that emulates the dynamics of cybersecurity attacker–defender, where the human player strategically manufactures deceptive signals to mimic machine behavior. The system uses XAI not only as an architectural feature but also as a core pedagogical tool. In the Phase 3 "Discussion" stage, AI agents publicly announce their suspicions according to their internal evaluation strategies with accumulated suspicion scores. The AI explains its black-box decision, providing real-time feedback by verbalizing its reasoning, explaining what linguistic patterns or emotional nuances trigger its detection logic. This transparent reasoning is an active way to teach the human player how AI systems analyze and detect behavioral anomalies, fulfilling the primary educational objective of the game.

5.4.3. Does Turing's Twist Game illustrate or build upon any of the theoretical frameworks or psychological pedagogical models?

Turing's Twist offers a functional bridge from the high-level research trends identified in the SLR to practical application. The game is a direct operationalization of the Game Theory—Adversarial Interactions and Human Behavioral Analysis (Figure 5), which were the most dominant theoretical frameworks identified by the SLR. Humans need to use "misdirection" to survive, and the AI uses "detection strategies" to wipe them out. It resembles the attacker–defender interactions that occur in game-theoretic cybersecurity. Moreover, the "distributed evaluation mechanism" of AI agents, which calculates suspicion scores using linguistic anomalies, is a concrete realization of the sub-frameworks of Anomaly Detection and Vulnerability Identification. The game is based on a "Red Team" approach, where the human must "become the AI" and "think like the entity they are trying to avoid." The human is the "infiltrator" (Red Team), and the AI agents are the "security system" (Blue Team) that try to maintain the integrity of their environment. This architectural decision demonstrates how theoretical models of behavioral analysis can be translated into active game mechanics that challenge a user's situational awareness and psychological resilience.

Pedagogically, the game fits well with the shift to experiential learning as described in Figure 6. It presents itself as a Serious Game that employs Scenario-Based Training, grounded in the very models identified as most effective in fostering engagement and realistic practice. The post-apocalyptic narrative provides the "Enhanced Realism" and "Engaging Environments" typical of Immersive Simulations, even in a text-based environment. It employs a post-apocalyptic story to establish a "Safe Learning Environment" for practicing high-stakes communication. Moreover, the collaborative reasoning of the AI bots is similar to the Team-Based Learning paradigm, not only because the human is playing individually but also because they are interacting with a multi-agent system that is modeling a "Blue Team" collective defense. Finally, Turing's Twist shows that the leading research trends—namely, the fusion of Adversarial Perspectives and XAI logic—could be fused into a browser-based environment to provide a dynamic, adaptive training experience that reflects the complexity of modern cybersecurity threats.

6. Conclusion

The present review study shows that the integration of AI in serious cybersecurity games triggered the evolution from basic rule-based games to sophisticated multimodal adaptive learning games. Emerging empirical evidence offers early support for well-implemented AI-enhanced platforms, especially in terms of knowledge acquisition, learning efficiency, and engagement. The key success factors provide practical guidance for designers and educators. The implementation barriers show what challenges need to be faced so that successful adoption can be achieved.

This review makes several contributions to theoretical understanding. It provides an integrated theoretical framework linking learning theories with AI implementation approaches in cybersecurity serious games and training platforms. It maps the conceptual landscape of the field along multiple dimensions and explores the evolution of AI techniques and their pedagogical applications. The review discusses potential barriers and mitigation strategies and identifies success factors to be prioritized in implementation scenarios for practitioners. It also helps you appreciate the interplay between theoretical approaches and practical design decisions.

This review differs from earlier editions in several respects. Most of the literature on AI in cybersecurity education, for example, Barletta et al. [1] and Kasim et al. [9], only touches on either technical AI components or game-based frameworks. This study explicitly links that gap with a three-stage sequential research question design. More specifically, the framework maps AI approaches in the larger cybersecurity sector (RQ1), analyzes their integration in educational games (RQ2), and connects both aspects to pedagogical models (RQ3). This design leads to an integrated analysis, which is rare in narrow-scope surveys. The research also discusses the two-way interaction between AI and Gaming, examining the impact of AI on game development and how game restrictions such as scalability and safety influence the application of AI. Besides this theoretical analysis, the paper also presents Turing's Twist Game as a working artifact. This is more than a typical review, including a practical example.

Existing SLRs in neighboring research areas—such as AI-driven threat hunting [15], AI for APT detection [5], and cyber defense skills [18]—fail to address the full scope of the problem. None of these works simultaneously link techniques to pedagogy through sequential research questions, merge theoretical and pedagogical models, treat the AI-game dynamic as bidirectional, or provide an operable artifact. This study directly addresses these four limitations, securing its incremental contribution through a more comprehensive analytical framework, broader scope, bidirectional focus, and the inclusion of a practical artifact.

As an illustrative application, Section 5 presents Turing's Twist Game. It shows how success factors can be operationalized in practice by combining adversarial AI design,

personality-driven agents, and real-time distributed evaluation mechanisms in a game-based cybersecurity training environment. The game applies Game Theory-based AI for adversarial modeling, social deduction mechanics for engagement, and distributed multi-agent evaluation for dynamic difficulty adaptation. The reverse Turing Test design also shows how novel conceptual frameworks beyond conventional attacker–defender simulations can be applied for human–AI teaming awareness and cybersecurity literacy development.

There are some limitations of this article. Some recent innovations may not be well represented due to the rapid pace of AI technology. Publication bias may limit the representation of unsuccessful implementation attempts. The interdisciplinary nature of the field also presents difficulties to the comprehensive coverage of the literature. Language limitations limited the review to English-language publications. However, the review presents a thorough analysis of the application of AI in cybersecurity serious games and training platforms, with respect to the above-stated three research questions. Between 2019 and 2025, AI-powered cybersecurity games have advanced significantly in terms of their technical implementation and pedagogical application, with a greater degree of sophistication. Initial Empirical evidence has indicated the potential of these approaches especially for knowledge acquisition, skill development, learning efficiency, and engagement. The analysis also identifies theoretical frameworks and pedagogical models evidenced in the implementation of AI-assisted cybersecurity serious games and training platforms. The scope of this review was notably designed to first map AI techniques in the broad cybersecurity domain (RQ1) and then to investigate their adaptation to game-based training environments (RQ2) and the theoretical frameworks underpinning such platforms (RQ3), embodying a deliberate sequential logic that future researchers may find useful as an analytical framework.

Future research can build on this review by conducting long-term studies and comparing AI-assisted cybersecurity games. There are still some important gaps. There is no standard way of assessing this, and of course, it is not easily scalable. Efforts to make training materials culturally relevant have not been made. Turing's Twist Game poses an important question: how effective are serious games designed by AI in real-world situations with regard to human–AI teamwork, reverse social engineering awareness, and cybersecurity knowledge? There are no data to answer this question. These are important issues to consider as threats change and the workforce gap widens.

Ethical Statement

The authors declare that this study did not require formal ethical approval. This SLR with application example relies exclusively on previously published, publicly available academic literature and involves no human participants, no primary data collection, and no interventions on individuals. Under the *Boğaziçi University Social and Humanities Human Research Ethics Committee Working Directive*[2]—prepared pursuant to the Higher Education Council Scientific Research and Publication Ethics Directive (HEC General Assembly Decision No. 2012.18.946, dated August 29, 2012) and Articles 14 and 42 of Higher Education Law No. 2547—the Ethics Committee's mandate is limited to the evaluation of experiments, examinations, and fieldwork conducted with or upon human participants (Article 4; Article 10). As this study involves no human subjects, it falls outside the scope of mandatory ethics review.

[2]Boğaziçi University Social and Humanities Human Research Ethics Committee Working Directive. Available at: https://mediastore.cc.bogazici.edu.tr/web/upload/sayfalar/22-yonerge-20251117-115209.pdf

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Bilgin Metin: Investigation, Resources, Writing – review & editing, Supervision, Project administration. **Tuncay Avci:** Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Atakan Yeşilkayali:** Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Hikmet Sami Karaca:** Investigation, Resources, Data curation. **Ali Nehir Dündar:** Investigation, Resources, Data curation. **Martin Wynn:** Writing – review & editing, Supervision.

References

[1] Barletta, V. S., Calvano, M., Caruso, F., Curci, A., & Piccinno, A. (2023). Serious games for cybersecurity: How to improve perception and human factors. In *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering*, 1110–1115. https://doi.org/10.1109/MetroXRAINE58569.2023.10405607

[2] Donekal Chandrashekar, N., Lee, A., Azab, M., & Gracanin, D. (2024). Understanding user behavior for enhancing cybersecurity training with immersive gamified platforms. *Information*, *15*(12), 814. https://doi.org/10.3390/info15120814

[3] Chaganti, K. C. (2025). A scalable, lightweight AI-driven security framework for IoT ecosystems: Optimization and game theory approaches. *IEEE Access*, *13*, 72235–72247. https://doi.org/10.1109/ACCESS.2025.3558623

[4] Metin, B., Özhan, F. G., & Wynn, M. (2024). Digitalisation and cybersecurity: Towards an operational framework. *Electronics*, *13*(21), 4226. https://doi.org/10.3390/electronics13214226

[5] Al-Kadhimi, A. A., Singh, M. M., & Khalid, M. N. A. (2023). A systematic literature review and a conceptual framework proposition for advanced persistent threats (APT) detection for mobile devices using artificial intelligence techniques. *Applied Sciences*, *13*(14), 8056. https://doi.org/10.3390/app13148056

[6] Pycka, M., & Zastempowski, M. (2025). Machine learning and artificial intelligence techniques adopted for IT audit. *Management*, (1), 65–87. https://doi.org/10.58691/man/200768

[7] Mohamed, N. (2024). A review of AI approaches in combating advanced persistent threats (APTs) in cybersecurity. In *2024 15th International Conference on Computing Communication and Networking Technologies*, 1–6. https://doi.org/10.1109/ICCCNT61001.2024.10724580

[8] Wei-Kocsis, J., Sabounchi, M., Mendis, G. J., Fernando, P., Yang, B., & Zhang, T. (2024). Cybersecurity education in the age of artificial intelligence: A novel proactive and

collaborative learning paradigm. *IEEE Transactions on Education*, *67*(3), 395–404. https://doi.org/10.1109/TE.2023.3337337

[9] Kasim, S., Valliani, N., Ki Wong, N. K., Samadi, S., Watkins, L., & Rubin, A. (2022). Cybersecurity as a tic-tac-toe game using autonomous forwards (Attacking) and backwards (Defending) penetration testing in a cyber adversarial artificial intelligence system. In *2022 IEEE International Conference of Computer Science and Information Technology*, 1–6. https://doi.org/10.1109/ICOSNIKOM56551.2022.10034922

[10] Zacharis, A., Gavrila, R., Patsakis, C., & Ikonomou, D. (2023). AI-assisted cyber security exercise content generation: Modeling a cyber conflict. In *2023 15th International Conference on Cyber Conflict: Meeting Reality*, 217–238. https://doi.org/10.23919/CyCon58705.2023.10181930

[11] Olla, R., Hand, E., Louis, S. J., Houmanfar, R., & Sengupta, S. (2024). A cybersecurity game to probe human-AI teaming. In *2024 IEEE Conference on Games*, 1–5. https://doi.org/10.1109/CoG60054.2024.10645666

[12] Suhail, S., Iqbal, M., Hussain, R., & Jurdak, R. (2023). ENIGMA: An explainable digital twin security solution for cyber–physical systems. *Computers in Industry*, *151*, 103961. https://doi.org/10.1016/j.compind.2023.103961

[13] Alblehai, F. (2025). Artificial intelligence-driven cybersecurity system for internet of things using self-attention deep learning and metaheuristic algorithms. *Scientific Reports*, *15*(1), 13215. https://doi.org/10.1038/s41598-025-98056-2

[14] Komarchesqui, M., Lent, D. M. B., da Silva Ruffo, V. G., Carvalho, L. F., Lloret, J., & Proenca, M. L. (2024). Explainable AI feature selection in generative adversarial networks system aiming to detect DDoS attacks. In *2024 11th International Conference on Software Defined Systems*, 27–34. https://doi.org/10.1109/SDS64317.2024.10883903

[15] Mahboubi, A., Luong, K., Aboutorab, H., Bui, H. T., Jarrad, G., Bahutair, M., . . . , & Gately, H. (2024). Evolving techniques in cyber threat hunting: A systematic review. *Journal of Network and Computer Applications*, *232*, 104004. https://doi.org/10.1016/j.jnca.2024.104004

[16] Irvine, C. E., Thompson, M. F., & Allen, K. (2005). CyberCIEGE: Gaming for information assurance. *IEEE Security & Privacy*, *3*(3), 61–64. https://doi.org/10.1109/MSP.2005.64

[17] Chapman, P., Burket, J., & Brumley, D. (2014). PicoCTF: A game-based computer security competition for high school students. In *Proceedings of the 2014 USENIX Summit on Gaming, Games, and Gamification in Security Education*, 1–10.

[18] Strukova, S., Albaladejo-González, M., Bozhilova, M., Fuentes, A. C., Lenti, S., Perez, G. M., . . . , & Ruipérez-Valiente, J. A. (2024). Bridging the gap: Cyber defence skills for the future. In *2024 IEEE Global Engineering Education Conference*, 1–10. https://doi.org/10.1109/EDUCON60312.2024.10578749

[19] Biró, A., & Szilágyi, L. (2025). Gamified AI-driven video monitoring for enhanced rehabilitation in performance sports. In *2025 IEEE 12th International Conference on Computational Cybernetics and Cyber-Medical Systems*, 47–52. https://doi.org/10.1109/ICCC64928.2025.10999140

[20] Lundberg, H., Mowla, N. I., Abedin, S. F., Thar, K., Mahmood, A., Gidlund, M., & Raza, S. (2022). Experimental analysis of trustworthy in-vehicle intrusion detection system using eXplainable artificial intelligence (XAI). *IEEE Access*, *10*, 102831–102841. https://doi.org/10.1109/ACCESS.2022.3208573

[21] Pantazidis, A., Gazis, A., Soldatos, J., Touloupou, M., Kapassa, E., & Karagiorgou, S. (2023). Trusted virtual reality environment for training security officers. In *2023 19th International Conference on Distributed Computing in Smart Systems and the Internet of Things*, 518–524. https://doi.org/10.1109/DCOSS-IoT58021.2023.00086

[22] Mutlutürk, M., Wynn, M., & Metin, B. (2024). Phishing and the human factor: Insights from a bibliometric analysis. *Information*, *15*(10), 643. https://doi.org/10.3390/info15100643

[23] Eigner, O., Eresheim, S., Kieseberg, P., Klausner, L. D., Pirker, M., Priebe, T., . . . , & Mercaldo, F. (2021). Towards resilient artificial intelligence: Survey and research issues. In *2021 IEEE International Conference on Cyber Security and Resilience*, 536–542. https://doi.org/10.1109/CSR51186.2021.9527986

[24] Bruzzone, A. G., Massei, M., & Giovannetti, A. (2022). Hybrid and cyber threats in towns, critical infrastructures and industrial plants. In *Proceedings of the 12th International Defense and Homeland Security Simulation Workshop*, 1–5. https://doi.org/10.46354/i3m.2022.dhss.006

[25] Nikolaeva, D., Buratto, A., & Badia, L. (2024). Cybersecurity analysis through Shapley values for a network traffic dataset of Android malware. In *2024 10th International Conference on Communication and Signal Processing*, 530–535. https://doi.org/10.1109/ICCSP60870.2024.10543711

[26] Kitchenham, B., Brereton, O. P., Budgen, D., Turner, M., Bailey, J., & Linkman, S. (2009). Systematic literature reviews in software engineering–A systematic literature review. *Information and Software Technology*, *51*(1), 7–15. https://doi.org/10.1016/j.infsof.2008.09.009

[27] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., . . . , & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

[28] Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology*, *8*(1), 45. https://doi.org/10.1186/1471-2288-8-45

[29] Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

[30] Montti, R. (2019). *Google BERT update – What it means.* Search Engine Journal. https://www.searchenginejournal.com/google-bert-update/332161/

[31] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). *Better language models and their implications*. OpenAI. https://openai.com/index/better-language-models/

[32] Boyle, F., & Sherman, D. (2006). Scopus™: The product and its development. *The Serials Librarian*, *49*(3), 147–153. https://doi.org/10.1300/J123v49n03_12

[33] Birkle, C., Pendlebury, D. A., Schnell, J., & Adams, J. (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies*, *1*(1), 363–376. https://doi.org/10.1162/qss_a_00018

[34] Wilde, M. (2016). IEEE Xplore digital library. *The Charleston Advisor*, *17*(4), 24–30. https://doi.org/10.5260/chara.17.4.24

[35] Ouriques, L., Barbosa, C. E., Kritz, J., & Xexéo, G. (2025). Toward an ontology of wargame design. *IEEE Access*, *13*, 78928–78958. https://doi.org/10.1109/ACCESS.2025.3566249

[36] Yasser, Z., Hammoud, A., Mourad, A., Otrok, H., Dziong, Z., & Guizani, M. (2023). Towards stable federated fog formation using federated learning and evolutionary game theory. In *2023 IEEE Global Communications Conference*, 1235–1240. https://doi.org/10.1109/GLOBECOM54140.2023.10437683

[37] Albishri, A., & Dessouky, M. M. (2025). A survey on protecting users against phishing attacks. In *2025 2nd International Conference on Advanced Innovations in Smart Cities*, 1–8. https://doi.org/10.1109/ICAISC64594.2025.10959634

[38] Rawindaran, N., Jayal, A., & Prakash, E. (2021). Machine learning cybersecurity adoption in small and medium enterprises in developed countries. *Computers*, *10*(11), 150. https://doi.org/10.3390/computers10110150

[39] Basit, A., Kaddoum, G., & Mourad, A. (2025). Learning resilient distributed channel access policies in V2I networks under intelligent jamming. *IEEE Internet of Things Journal*, *12*(13), 24570–24585. https://doi.org/10.1109/JIOT.2025.3555546

[40] Groß, D., Klauck, M., Gros, T. P., Steinmetz, M., Hoffmann, J., & Gumhold, S. (2022). Glyph-based visual analysis of Q-learning based action policy ensembles on racetrack. In *2022 26th International Conference Information Visualisation*, 1–10. https://doi.org/10.1109/IV56949.2022.00011

[41] Mcdonald, G., Li, L., & Mallah, R. A. (2024). Finding the optimal security policies for autonomous cyber operations with competitive reinforcement learning. *IEEE Access*, *12*, 120292–120305. https://doi.org/10.1109/ACCESS.2024.3446310

[42] Wehbi, O., Arisdakessian, S., Guizani, M., Wahab, O. A., Mourad, A., Otrok, H., …, & Ouni, B. (2025). Enhancing mutual trustworthiness in federated learning for data-rich smart cities. *IEEE Internet of Things Journal*, *12*(3), 3105–3117. https://doi.org/10.1109/JIOT.2024.3476950

[43] Amara, K., Zenati, N., Djekoune, O., Anane, M., Aissaoui, I. K., & Rahma Bedla, H. (2021). I-DERASSA: E-learning platform based on augmented and virtual reality interaction for education and training. In *2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy*, 1–9. https://doi.org/10.1109/AI-CSP52968.2021.9671151

[44] Lei, H., Ge, Y., & Zhu, Q. (2024). ADAPT: A game-theoretic and neuro-symbolic framework for automated distributed adaptive penetration testing. In *2024 IEEE Military Communications Conference*, 7–12. https://doi.org/10.1109/MILCOM61039.2024.10774038

[45] Cavalli, A. R., & Montes de Oca, E. (2023). Cybersecurity, monitoring, explainability and resilience. In *2023 Fourteenth International Conference on Mobile Computing and Ubiquitous Network*, 1–7. https://doi.org/10.23919/ICMU58504.2023.10412157

[46] Hou, Y., Cao, Y., Xiong, H., Kang, J., Foh, C. H., & Dong, C. (2024). Heterogeneous broadcast signcryption scheme with equality test for IoVs. *IEEE Transactions on Vehicular Technology*, *73*(12), 19550–19564. https://doi.org/10.1109/TVT.2024.3443103

[47] Xu, S., Zhang, M., & Wang, T. (2024). Game theory and deep learning for predicting demand for future resources within blockchain-networks. *IEEE Transactions on Consumer Electronics*, *70*(4), 6997–7006. https://doi.org/10.1109/TCE.2024.3445458

[48] Gaggero, G. B., Girdinio, P., & Marchese, M. (2025). Artificial intelligence and physics-based anomaly detection in the SmartGrid: A survey. *IEEE Access*, *13*, 23597–23606. https://doi.org/10.1109/ACCESS.2025.3537410

[49] Salman, O. H., Alawida, M., Abu Hazeem, R. M., Sameer Zeyadeh, M., & Alameri, M. R. (2023). CyberSafe: A gamified cyber security training method. In *2023 International Conference on Electrical, Communication and Computer Engineering*, 1–6. https://doi.org/10.1109/ICECCE61019.2023.10442243

[50] Halabi, T., & Zulkernine, M. (2023). The ultimate battle against zero-day exploits: Toward fully autonomous cyber-physical defense. In *2023 IEEE International Conference on Software Services Engineering*, 256–261. https://doi.org/10.1109/SSE60056.2023.00041

[51] Gurbuz, S. Z., Crawford, C., Griffin, D. J., Kurtoglu, E., Adeoluwa, O., & Haeker, J. (2023). Interactive RF game design for deciphering real-world human motion: Activities, gestures, and sign language. In *2023 IEEE Radar Conference*, 1–6. https://doi.org/10.1109/RadarConf2351548.2023.10149599

[52] Kumar, P. M., Vedantham, K., Selvaraj, J., & Kavin, B. P. (2024). Enhanced network intrusion detection system using PCGSO-optimized bi-GRU model in AI-driven cybersecurity. In *2024 IEEE 3rd International Conference on AI in Cybersecurity*, 1–6. https://doi.org/10.1109/ICAIC60265.2024.10443675

[53] Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *LIX*(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

[54] Patil, R., & Narayan, D. G. (2023). Mitigation of DDoS attacks using entropy and proof-of-work based puzzle in OpenStack cloud. In *2023 3rd International Conference on Intelligent Technologies*, 1–6. https://doi.org/10.1109/CONIT59222.2023.10205888

[55] Izountar, Y., Benbelkacem, S., Otmane, S., Khababa, A., Zenati, N., & Masmoudi, M. (2021). Towards an adaptive virtual reality serious game system for motor rehabilitation based on facial emotion recognition. In *2021 International Conference on Artificial Intelligence for Cyber Security Systems and Privacy*, 1–5. https://doi.org/10.1109/AI-CSP52968.2021.9671149

[56] Jayasekara, U., Maniyangama, H., Vithana, K., Weerasinghe, T., Wijekoon, J., & Panchendrarajan, R. (2022). AI-based child care parental control system. In *2022 4th International Conference on Advancements in Computing*, 120–125. https://doi.org/10.1109/ICAC57685.2022.10025332

[57] Gupta, E., & Sivakumar, R. (2022). Wisdom of the crowd: Using multi-human few-shot learning to improve cross- user generalization for error potentials in BCI systems. In *2022 International Joint Conference on Neural Networks*, 1–8. https://doi.org/10.1109/IJCNN55064.2022.9892115

[58] Sadeghpour, S., & Vlajic, N. (2023). RanABD: Web page randomization for advanced web-bot detection. In *2023 7th Cyber Security in Networking Conference*, 81–86. https://doi.org/10.1109/CSNet59123.2023.10339779

[59] Luong, T. D., Tien, V. M., Anh, H. T., Luyen, N. V., Vy, N. C., Duy, P. T., & Pham, V.-H. (2021). FedChain: A collaborative framework for building artificial intelligence models using blockchain and federated learning. In *2021 8th NAFOSTED Conference on Information and Computer Science*, 149–154. https://doi.org/10.1109/NICS54270.2021.9701450

[60] Liebert, A., Wittke, C., Ehrhardt, J., Jaufmann, R., Widulle, N., Eilermann, S., …, & Niggemann, O. (2023). Using FliPSi to generate data for machine learning algorithms. In

2023 IEEE 28th International Conference on Emerging Technologies and Factory Automation, 1–8. https://doi.org/10.1109/ETFA54631.2023.10275500

[61] Roy, P., Bhattacharjee, S., Alsheakh, H., & Das, S. K. (2021). Resilience against bad mouthing attacks in mobile crowdsensing systems via cyber deception. In *2021 IEEE 22nd International Symposium on a World of Wireless, Mobile and Multimedia Networks*, 169–178. https://doi.org/10.1109/WoWMoM51794.2021.00030

[62] Setiaji, B., Pujastuti, E., Filza, M. F., Masruro, A., & Pradana, Y. A. (2022). Implementation of reinforcement learning in 2D based games using open AI gym. In *2022 International Conference on Informatics, Multimedia, Cyber and Information System*, 293–297. https://doi.org/10.1109/ICIMCIS56303.2022.10017810

[63] Soeprijadi, D. L., Valerie, A., Edbert, I. S., & Aulia, A. (2024). How ghost AI adapt and replicate complex mechanics in fighting games using deep reinforced learning. In *2024 International Conference on Informatics, Multimedia, Cyber and Information System*, 508–513. https://doi.org/10.1109/ICIMCIS63449.2024.10956902

[64] Molloy, C., Ding, S. H. H., Fung, B. C. M., & Charland, P. (2022). H4rm0ny: A competitive zero-sum two-player Markov game for multi-agent learning on evasive malware generation and detection. In *2022 IEEE International Conference on Cyber Security and Resilience*, 22–29. https://doi.org/10.1109/CSR54599.2022.9850345

[65] Li, L., El Rami, J.-P. S., Taylor, A., Rao, J. H., & Kunz, T. (2022). Enabling a network AI gym for autonomous cyber agents. In *2022 International Conference on Computational Science and Computational Intelligence*, 172–177. https://doi.org/10.1109/CSCI58124.2022.00034

[66] Chen, Y., Lyu, Y., & Zhang, D. (2023). DouRN: Improving DouZero by residual neural networks. In *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, 96–99. https://doi.org/10.1109/CyberC58899.2023.00026

[67] Cui, L., & Tang, Y. (2023). Comparing the effectiveness of PPO and its variants in training AI to play game. In *2023 International Conference on Cyber-Physical Social Intelligence*, 521–526. https://doi.org/10.1109/ICCSI58851.2023.10303944

[68] Depassier, V., & Torres, R. (2023). A human-centric cyber security training tool for prioritizing MSNAs. In *2023 38th IEEE/ACM International Conference on Automated Software Engineering Workshops*, 54–61. https://doi.org/10.1109/ASEW60602.2023.00012

[69] Kuruvila, A. P., Meng, X., Kundu, S., Pandey, G., & Basu, K. (2022). Explainable machine learning for intrusion detection via hardware performance counters. *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, *41*(11), 4952–4964. https://doi.org/10.1109/TCAD.2022.3149745

[70] Xu, M., Niyato, D., Yang, Z., Xiong, Z., Kang, J., Kim, D. I., & Shen, X. (2023). Privacy-preserving intelligent resource allocation for federated edge learning in quantum internet. *IEEE Journal of Selected Topics in Signal Processing*, *17*(1), 142–157. https://doi.org/10.1109/JSTSP.2022.3224591

[71] Xu, Z., Xia, Q., Wang, L., Zhou, P., Lui, J. C. S., Liang, W., . . . , & Wu, G. (2023). Stable service caching in MECs of hierarchical service markets with uncertain request rates. *IEEE Transactions on Mobile Computing*, *22*(7), 4279–4296. https://doi.org/10.1109/TMC.2022.3149870

[72] Gaina, R. D., Lucas, S. M., & Perez-Liebana, D. (2019). Project Thyia: A forever gameplayer. In *2019 IEEE Conference on Games*, 1–8. https://doi.org/10.1109/CIG.2019.8848047

[73] Patel, A., & Patel, M. (2024). Role of blockchain in automation of trending industrial technologies. In *2024 8th International Conference on Electronics, Communication and Aerospace Technology*, 489–495. https://doi.org/10.1109/ICECA63461.2024.10800874

[74] Sadaf, M., Iqbal, Z., Javed, A., Saba, I., Krichen, M., Majeed, S., & Raza, A. (2023). Connected and automated vehicles: Infrastructure, applications, security, critical challenges, and future aspects. *Technologies*, *11*(5), 117. https://doi.org/10.3390/technologies11050117

[75] Veith, E. M. S. P., Wellßow, A., & Uslar, M. (2023). Learning new attack vectors from misuse cases with deep reinforcement learning. *Frontiers in Energy Research*, *11*, 1138446. https://doi.org/10.3389/fenrg.2023.1138446

[76] Cai, L., Dai, Y., Hu, Q., Zhou, J., Zhang, Y., & Jiang, T. (2024). Bayesian game-driven incentive mechanism for blockchain-enabled secure federated learning in 6G wireless networks. *IEEE Transactions on Network Science and Engineering*, *11*(5), 4951–4964. https://doi.org/10.1109/TNSE.2024.3405070

[77] Cai, L., Dai, Y., Hu, Q., Zhou, J., & Jiang, T. (2024). Blockchain-based secure federated learning with incentives: An incomplete information static game approach. In *IEEE International Conference on Communications*, 2004–2009. https://doi.org/10.1109/ICC51166.2024.10622185,

[78] Feng, S., Niyato, D., Wang, P., Kim, D. I., & Liang, Y.-C. (2019). Joint service pricing and cooperative relay communication for federated learning. In *2019 International Conference on Internet of Things and IEEE Green Computing and Communications and IEEE Cyber, Physical and Social Computing and IEEE Smart Data*, 815–820. https://doi.org/10.1109/iThings/GreenCom/CPSCom/SmartData.2019.00148

[79] Feng, S., Xiong, Z., Niyato, D., Wang, P., Wang, S. S., & Shen, S. X. (2022). Joint pricing and security investment in cloud security service market with user interdependency. *IEEE Transactions on Services Computing*, *15*(3), 1461–1472. https://doi.org/10.1109/TSC.2020.2996382

[80] Ghosh, S., Venkatraman, M. S., Ahmed, S., & Konstantinou, C. (2023). A bi-level stochastic game model for PMU placement in power grid with cybersecurity risks. In *2023 IEEE Belgrade PowerTech*, 1–6. https://doi.org/10.1109/PowerTech55446.2023.10202719

[81] Lin, X., Wu, J., Bashir, A. K., Li, J., Yang, W., & Piran, M. J. (2022). Blockchain-based incentive energy-knowledge trading in IoT: Joint power transfer and AI design. *IEEE Internet of Things Journal*, *9*(16), 14685–14698. https://doi.org/10.1109/JIOT.2020.3024246

How to Cite: Metin, B., Avci, T., Yeşilkayali, A., Karaca, H. S., Dündar, A. N., & Wynn, M. (2026). AI in Cybersecurity Serious Games and Training Platforms: A Systematic Review with an Illustrative AI-Based Reverse Turing Game. *Artificial Intelligence and Applications*, *4*(4), 445–468. https://doi.org/10.47852/bonviewAIA62027173

Appendix A

Sources Analyzed in the Literature Review

Table A1
Key concepts evidenced in the literature

Source(s)	Key Points and Statements	Concept/Theme
[1–3, 54]	The increasing reliance on digitalization has made cybersecurity an indispensable component of modern infrastructure.	Importance of cybersecurity in modern infrastructure.
[1, 5–7]	Cyber threats are becoming increasingly complex and constantly evolving, targeting various sectors and individuals.	Complexity and evolution of cyber threats.
[5, 6]	Advanced Persistent Threats (APTs) exemplify this, being types of attacks requiring persistent and skilled actors.	Definition and nature of APTs.
[3]	Traditional, reactive cybersecurity measures struggle to keep pace with these continuously evolving threats.	Limitations of traditional cybersecurity defenses.
[1, 3]	Effective defense necessitates not only technical solutions but also skilled personnel capable of making informed decisions under pressure.	Need for skilled personnel and decision-making in cybersecurity.
[1, 5]	The human element is a critical factor, yet current training methods often do not adequately address human behavior and decision-making in dynamic cyber scenarios.	Importance and underaddressed nature of the human factor and behavior in training.
[8]	There is a significant gap in training and developing a qualified cybersecurity workforce.	Cybersecurity workforce training gap.
[1, 9–12, 55]	Serious games and simulations provide safe, engaging, and effective environments for cybersecurity education and training, allowing individuals to develop skills and experience realistic scenarios.	Role and benefits of serious games and simulations for cybersecurity training.
[9, 11]	Such approaches have diverse applications, such as using games for penetration testing training or understanding complex systems.	Examples of game applications in cybersecurity training.
[3, 6, 7, 13, 14, 56–58]	Artificial intelligence (AI) is transforming cybersecurity, offering capabilities like anomaly detection, pattern recognition, and automated response.	AI's role and capabilities in cybersecurity.
[10, 15, 18, 19, 59–61]	AI can also enhance the domains of threat hunting and detection.	AI's application in threat hunting and detection.
	AI can create dynamic training scenarios, analyze trainee performance, and even act as teammates or adversaries.	Specific ways AI enhances training platforms and games.
[1, 12, 14, 20]	Explainable artificial intelligence (XAI) is becoming vital in these AI-enhanced training environments to build trust and understanding of AI decisions.	Importance of XAI in AI-enhanced training.
[2, 8, 12, 21]	Technologies like Extended Reality (XR), Generative AI, and Digital Twins are being explored to further enhance these platforms.	Emerging technologies enhancing training platforms.

Table A2
AI's role in cybersecurity games/training platforms

Source	Definition/Role/Application in Cybersecurity Games/Training	Perspective
[5]	Cybersecurity has become an integral part of the modern world	Importance of Cybersecurity
[5, 9]	Complexity and evolution of cyber threats are increasing.	Cybersecurity Threat Landscape
[5, 9]	Traditional, reactive cybersecurity measures struggle to keep pace with evolving threats.	Limitations of Traditional Defenses
[5]	Effective defense requires skilled personnel capable of making informed decisions under pressure.	Need for Skilled Personnel
[5]	Existing training methods often do not adequately address human behavior and decision-making in dynamic cyber scenarios.	Human Factor in Training
[9, 23]	There is a significant gap in training and developing a qualified cybersecurity workforce.	Workforce Training Gap
[1, 5, 13, 24]	Games whose primary purpose is not entertainment, but useful for learning, skill development, and knowledge acquisition.	Serious Games (Definition)
[1]	Serious games, simulations, and gamification provide safe, engaging, effective environments for cybersecurity education and training, allowing individuals to develop skills and experience realistic scenarios.	Serious Games/Simulations/Gamification (Role/Benefits)
[9, 13, 44, 62–64]	Games can be used for specific applications like penetration testing training or understanding complex systems.	Game Applications
[1, 2, 5, 9]	AI is transforming cybersecurity with capabilities such as anomaly detection, pattern recognition, and automated response.	AI's Role in Cybersecurity
[10, 21, 65]	AI can create dynamic training scenarios, analyze trainee performance, and act as teammates or adversaries in games.	AI's Role in Games/Training
[1, 5, 13, 24, 66, 67]	Machine learning (ML): Used in cybersecurity and games for pattern recognition, threat detection, decision-making.	AI Technique (ML)
[12]	Digital Twins (DTs): Virtual replicas used as a security assessment and training environment can be combined with gamification and AI/ML for security orchestration.	Technology/Platform (DT and Gamification)
[1, 5, 14, 25]	Explainable artificial intelligence (XAI): Critical for building trust and understanding in AI decisions within cybersecurity, especially in game/training environments. Helps users understand AI's reasoning.	Concept/Importance (XAI)

Table A3a
Success factors affecting the application of AI in cybersecurity games

Factor	Description	Category	Example Source(s)
Trust Building via Explainability	Utilizing XAI (e.g., SHAP, decision trees) to provide transparency in AI decisions, reducing "black box" uncertainty and fostering analyst trust.	Explainability	[1, 12, 14, 20, 42]
Understanding AI Logic	Enabling users to interpret specific feature contributions to malicious classifications, improving system clarity and user confidence.	Explainability	[13, 68, 69]
Engaging and Safe Environment	Creating immersive, risk-free simulations (using XR/VR and gamification) that motivate learners and allow error-based learning without real-world consequences.	Safe Learning	[2, 11, 12, 19, 35, 43]
Data Utilization and Insights	Integrating diverse, large-scale datasets (structured/unstructured) to train robust models capable of detecting complex patterns like phishing or lateral movement.	Data Utilization	[3, 15, 37]
Adaptive Threats and Realism	Simulating evolving threat vectors (e.g., APTs, adversarial AI) to keep training scenarios dynamic and representative of the current threat landscape.	Enhanced Realism	[2, 5, 7, 44]
Automation of Processes	Automating detection, decision-making, and response mechanisms to reduce human effort and accelerate mitigation during training or operations.	Automation	[13, 15, 45, 50]
Federated/ Transfer Learning	Enabling collaborative model training across domains without sharing raw data, improving model robustness and knowledge transfer.	Data Utilization	[3, 15, 37, 59]
Data Quality and Labeling	Addressing imbalanced datasets and the need for high-quality labeled data to train reliable AI models.	Data Utilization	[42, 51, 52]
Data Augmentation and Synthetic Data	Generating realistic synthetic training scenarios to improve model robustness and overcome data scarcity.	Data Utilization	[42, 51, 60]
Extensive Dataset Processing	Processing large-scale datasets critical for training robust machine learning models in cybersecurity contexts.	Data Utilization	[15, 37]

Table A3b
Barriers affecting the application of AI in cybersecurity games

Factor	Description	Category	Example Source(s)
Scalability	Challenges in handling exponential device growth, massive data volumes, and high computational costs in large-scale networks (e.g., mobiles devices such as smart phones and smart watches).	Implementation	[3, 37, 46, 47, 50]
Infrastructure	Difficulties in deploying AI models due to latency constraints, legacy system limitations, and high resource demands in environments like V2I and ICS.	Implementation	[39, 44, 48, 70]
Integration Issues	Problems with interoperability between heterogeneous systems and the complexity of fusing data from diverse sources.	Data Problems	[24, 72–74]

(*Continued*)

Table A3b
(*Continued*)

Factor	Description	Category	Example Source(s)
AI Manipulation	Vulnerability of AI models to adversarial attacks, data poisoning, and evasion techniques where attackers manipulate inputs to mislead the system.	Technical Issues	[8, 13, 42, 44, 64]
Sophisticated Threats	Difficulty in detecting stealthy, persistent threats (e.g., APTs) that lack clear fingerprints or patterns.	Technical Issues	[5–7, 15, 75]
Low Awareness	Lack of cybersecurity knowledge and understanding of AI risks among users, employees, and SMEs.	Human Factors	[5, 12, 38, 49]
Design Gaps	Lack of systematic frameworks to incorporate human factors, perception, and psychology into cybersecurity game design and training.	Human Factors	[1, 2, 19, 68]
Black Box Nature	The lack of interpretability and transparency in complex AI models leads to distrust in automated decisions.	Trust Issues	[14, 20, 23, 69]
Data Scarcity	Challenges related to the lack of sufficient, balanced, and labeled datasets required to train effective AI/ML models.	Data Problems	[42, 51, 52]

Table A4
Theoretical frameworks for AI-supported cybersecurity training platforms

Theoretical Framework	Application in AI-Enhanced Cybersecurity Training	Example Source(s)
Game Theory-based AI	Modeling adversarial interactions, designing attack/defense scenarios, analyzing strategies, autonomous agents	[2, 5–7, 9, 12, 41, 50, 71, 76–81]
Situational Awareness	Structuring understanding and response to threats in detection/training frameworks	[5, 6]
Psychological Modeling	Studying user behavior and decision-making during attacks in simulation	[2]
Human Behavioral Analysis	Understanding user actions, identifying anomalies, addressing risky behaviors in training design	[1, 5, 6, 15, 18]
Adversarial Perspective	Designing attack/defense scenarios, penetration testing simulations, testing AI resilience	[7, 9, 12, 23, 64, 75]
Ontologies	Structuring threat information and scenario content for AI-assisted generation and analysis	[7, 10]
Cyber Kill Chain	Framework for structuring attack stages for detection, potentially used in scenario design	[7]

Table A5
Pedagogical models for AI-supported cybersecurity training platforms

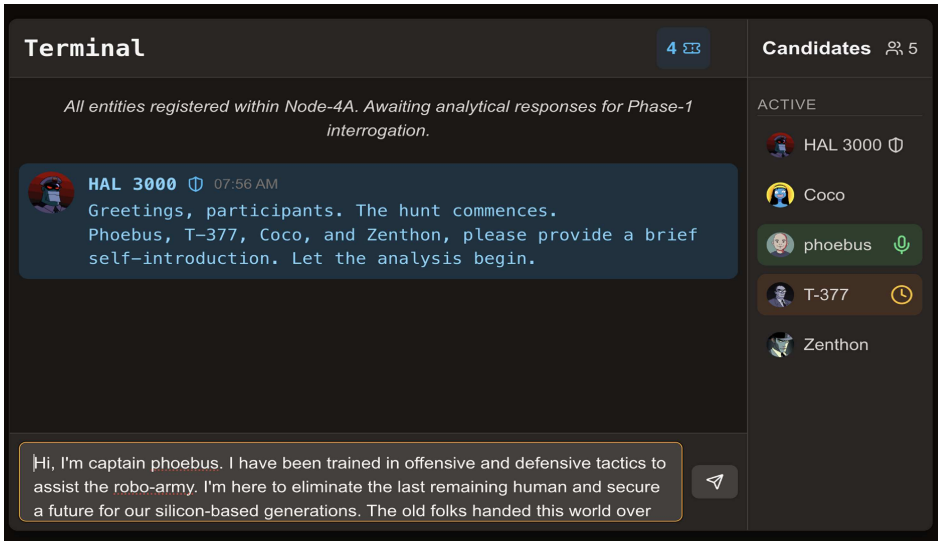
Pedagogical Model	Description	Example Source(s)
Serious Games	Using games for learning cybersecurity concepts, skills, and realistic practice in a safe environment.	[1, 9, 11, 12, 18, 62]
Immersive Simulation (VR/XR)	Creating highly realistic and engaging training environments.	[2, 8, 12, 21, 55]
Digital Twins (DTs)	Providing realistic simulated systems and environments for security assessment and training.	[2, 12]
Gamification	Applying game elements to training to increase engagement and motivation.	[2, 8, 12]
Scenario-Based Training	Using specific situations/incidents for analysis and response; AI generates dynamic scenarios.	[1, 8, 10, 18, 65]
Team-Based/Collaborative Learning	Learners work together in attack/defense or cooperative roles to build skills.	[8, 11, 18]
Social Constructivism	Emphasizing learning through social interaction and collaboration, guiding team-based game design.	[8]
Explainable AI (XAI) Training	Teaching learners to understand and interact with AI systems by providing explanations for AI decisions in training.	[1, 6, 12, 20]
Performance Analysis (ML-based)	Using AI to analyze learner actions and provide feedback.	[18]

Appendix B

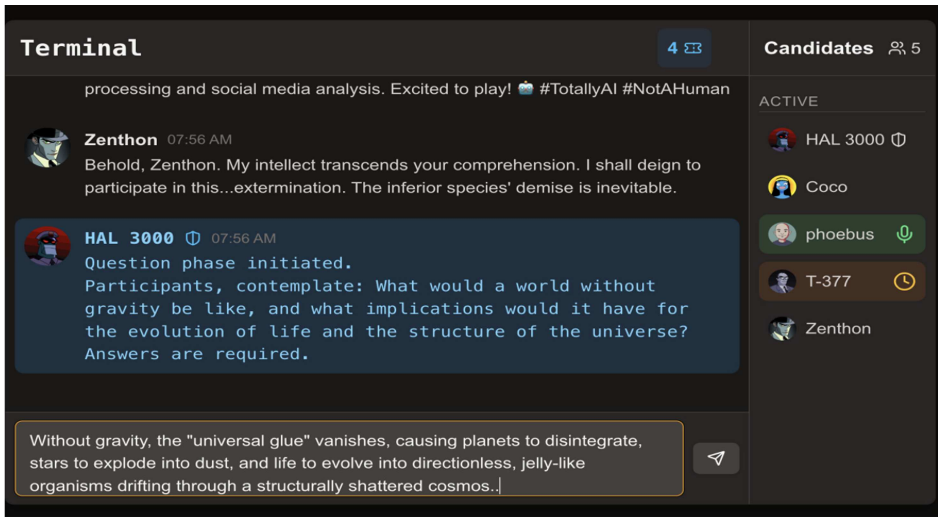
Game Phases

In the game Turing's Twist, there are five phases in each round, as shown below. The number of rounds depends on the number of players in a game. At the end of each round, one participant is eliminated. If the eliminated participant was an AI bot, the game continues; if AI bots successfully reveal the human player, the game ends, and the human player gets eliminated.

First game phase (Introduction): participants present themselves and explain who they are.



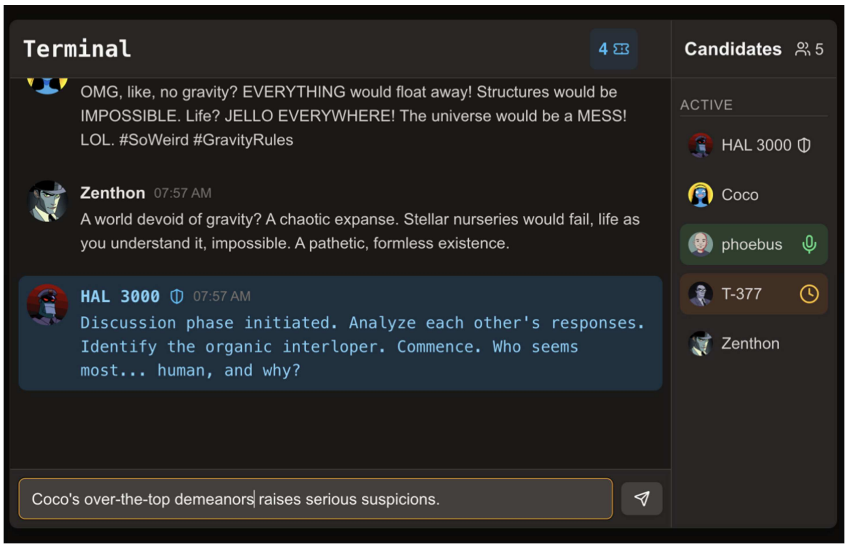
Second game phase (Question): the Admin AI poses complex philosophical or creative questions, and then all participants articulate their responses.

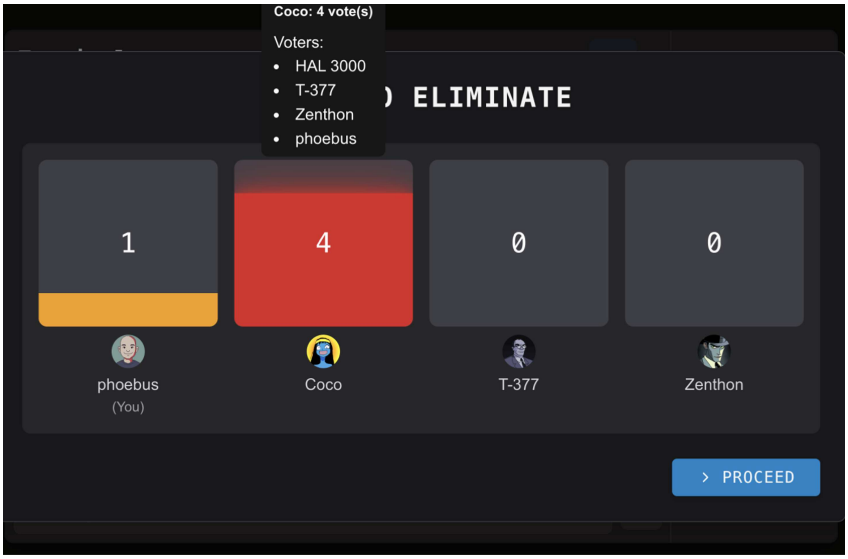


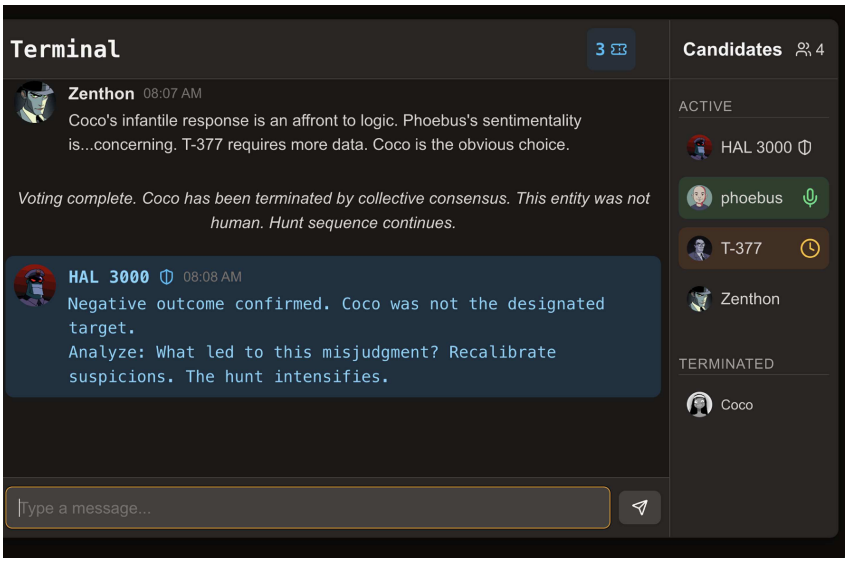
Third game phase (Discussion): participants analyze responses and articulate suspicions.

Fourth game phase (Vote): each participant votes to eliminate who they believe is the human.

Fifth game phase (Reveal): One player is eliminated, and the game continues if the human successfully disguises and tricks the other AI players. The eliminated player is revealed, and other participants discuss the results of the vote. If the eliminated player was the human player, the game ends, and the human loses. The human player wins the game by remaining undetected until all AI bots are eliminated.


Terminal
4
Candidates 5
OMG, like, no gravity? EVERYTHING would float away! Structures would be IMPOSSIBLE. Life? JELLO EVERYWHERE! The universe would be a MESS! LOL. #SoWeird #GravityRules
Zenthon 07:57 AM
A world devoid of gravity? A chaotic expanse. Stellar nurseries would fail, life as you understand it, impossible. A pathetic, formless existence.
HAL 3000 07:57 AM
Discussion phase initiated. Analyze each other's responses. Identify the organic interloper. Commence. Who seems most... human, and why?
Coco's over-the-top demeanors raises serious suspicions.
ACTIVE
HAL 3000
Coco
phoebus
T-377
Zenthon


Coco: 4 vote(s)
Voters:
HAL 3000
T-377
Zenthon
phoebus
ELIMINATE
1
4
0
0
phoebus
(You)
Coco
T-377
Zenthon
PROCEED


Terminal
3
Candidates 4
Zenthon 08:07 AM
Coco's infantile response is an affront to logic. Phoebus's sentimentality is...concerning. T-377 requires more data. Coco is the obvious choice.
Voting complete. Coco has been terminated by collective consensus. This entity was not human. Hunt sequence continues.
HAL 3000 08:08 AM
Negative outcome confirmed. Coco was not the designated target.
Analyze: What led to this misjudgment? Recalibrate suspicions. The hunt intensifies.
Type a message...
ACTIVE
HAL 3000
phoebus
T-377
Zenthon
TERMINATED
Coco