Bayesian Modeling of Uncertainty in Anti-phishing Training: A Serious Game for Adversarial Email Crafting

Dimitrios Lappas[1], Diogo Alexandre Silva[2,3], Aristeidis Angelos Zoumpakis[1] and Panagiotis Karampelas[4]

[1]Hellenic Centre for Defence Innovation, Athens, Hellas
[2]Portuguese Air Force Academy Research Center, Sintra, Portugal
[3]Military University Institute Research Center, Lisbon, Portugal
[4] Hellenic Air Force Academy, Dekeleia, Hellas

d.lappas@elkak.gr
dasilva@emfa.gov.pt
a.zoumpakis@elkak.gr
panagiotis.karampelas@hafa.haf.gr

Abstract: This paper presents a Bayesian-driven serious game designed to support human-centered anti-phishing training through explainable artificial intelligence and adversarial gameplay. As phishing attacks increasingly leverage personalization and AI-assisted social engineering, traditional training approaches and opaque detection systems offer limited support for developing strategic awareness and decision-making under uncertainty. The proposed framework addresses this gap by combining Bayesian explainability with an interactive, adversarial learning environment. The serious game was implemented in a military academy context and centers on a sandboxed anti-phishing platform powered by a Bayesian Network. Participants are tasked with crafting targeted phishing emails for simulated victims using social media-derived contextual information. Each email submission is analyzed by the Bayesian engine, which outputs both a probabilistic phishing likelihood score and an explanation highlighting the linguistic features that contributed most strongly to the classification. Learners then iteratively revise the same email to reduce detectability, effectively attempting to "fool" the system while preserving semantic intent. Evaluation was conducted using two complementary data sources: interaction logs from the game and a post-activity questionnaire. The results suggest that integrating explainable Bayesian models into adversarial serious games can effectively enhance phishing awareness, strategic thinking, and sensitivity to uncertainty. The study highlights the potential of explainable AI not only as a defensive mechanism, but also as a powerful educational tool in cybersecurity training.

Keywords: Serious game, Explainability, Bayesian inference, Phishing

1. Introduction

The contemporary cybersecurity landscape is characterized by a continuous increase in the complexity, adaptability, and targeted nature of cyber threats (Masunda, 2025). As attackers increasingly leverage artificial intelligence and automation technologies, cyberattacks are evolving from generalized and predictable actions into strategically designed operations that exploit human, cognitive, and social vulnerabilities (Infosecurity Magazine, 2025). Within this context, phishing attacks constitute one of the most prevalent and effective forms of cyber threats, as they bridge the technical and human layers of security (Mutlutürk et al., 2024).

The detection of phishing attacks no longer relies on obvious technical flaws or generalized patterns (Kavya & Sumathi, 2025). Instead, modern phishing campaigns exploit personalized content, social context, and linguistic nuances to construct persuasive narratives that can bypass both human judgment and automated detection systems (Push Security, 2025). Consequently, the prevailing approach to phishing detection, which emphasizes technical features and statistical patterns, remains effective in certain scenarios but exhibits notable limitations at both operational and educational levels. From an educational perspective, users often interact with anti-phishing systems as black-boxes, receiving a final decision or risk score without meaningful insight into how and why a message has been classified as suspicious (Calzarossa et al., 2025). This lack of transparency constrains the development of strategic thinking, as learners are unable to associate specific linguistic or social design choices with corresponding risk assessments (Rjoub et al., 2023). Such limitations highlight the need for explainable models that incorporate qualitative content analysis and decision justification, thereby enhancing user understanding and trust.

Effective education for addressing phishing threats therefore requires a shift from the mere recognition of technical indicators toward an understanding of uncertainty, causality, and the adaptive behavior of attackers (Lappas & Karampelas, 2024). One approach aligned with this perspective is the cultivation of probabilistic reasoning (Huang & Zhu, 2018; Rjoub et al., 2023), which supports decision-making in environments characterized by uncertainty, an essential competence in modern cybersecurity contexts.

In this context, Explainable Artificial Intelligence (XAI), and particularly Bayesian approaches, provide a strong theoretical and pedagogical foundation for cybersecurity education. Bayesian Networks enable the quantification of uncertainty and the modeling of causal relationships through conditional probabilities, making transparent how individual message features influence the overall phishing risk assessment (Pearl, 1988; Calzarossa et al., 2025). This capability is especially critical in educational environments, where understanding the decision-making process is as important as the final outcome, as it fosters strategic thinking skills and trust in automated systems (Rjoub et al., 2023).

Building upon these considerations, the present study developed a purpose-built anti-phishing tool designed explicitly for educational use. The tool is based on a Bayesian Network and provides not only a probabilistic phishing risk estimation but also interpretable feedback regarding the linguistic and contextual features that contributed to the assessment. The educational approach is further strengthened through the integration of adversarial game dynamics. The application of adversarial game dynamics emphasizes learner engagement in simulations that reveal attacker strategies and encourage the adoption of corresponding counter-strategic reasoning (Prebot, Du, & Gonzalez, 2023; Shashkov et al., 2023). Through this process, participants gain experiential insight into both social engineering strategies and the capabilities and limitations of anti-phishing systems (Pawlick et al., 2019).

Accordingly, this study proposes a holistic educational framework that combines Bayesian explainability and adversarial gaming dynamics for phishing awareness and defense training. The primary contributions of this work are summarized as follows:

- Highlighting the limitations of conventional anti-phishing tools with respect to qualitative message analysis.
- The development of an educational Bayesian anti-phishing system with built-in explainability.
- The investigation of the role of adversarial serious games in fostering strategic thinking and uncertainty awareness in cybersecurity contexts.

Unlike existing phishing training tools, this work exposes the internal probabilistic reasoning of the detection system and positions the learner in an adversarial optimization loop.

To guide the reader, the remainder of this paper is structured as follows. Section 2 presents the theoretical background, covering Explainable AI in cybersecurity, Bayesian Networks and adversarial game theory. Section 3 formulates the research problem and identifies gaps in the existing literature. The proposed educational framework, system design, evaluation methodology and results are described in Section 4. Finally, Section 5 discusses the findings and outlines directions for future research.

2. Theoretical Framework

2.1 Explainable AI (XAI) in Cybersecurity

Explainable Artificial Intelligence (XAI) encompasses methods that render the decisions of AI systems transparent and interpretable, enabling users to understand not only system outputs but also the reasoning processes underlying them (Calzarossa, Giudici, & Zieni, 2025). In cybersecurity, explainability is particularly important, as automated decisions directly affect trust, situational awareness, and operational responses to threats.

Within phishing detection, traditional machine learning approaches often operate as black-boxes, providing users with a binary classification or risk score without insight into the contributing factors. This opacity limits both user trust and educational value, as learners are unable to associate specific linguistic or contextual cues with phishing risk (Rjoub et al., 2023). XAI techniques address this limitation by exposing influential features or causal relationships, allowing users to reflect on why a message is considered suspicious.

Feature-attribution methods such as Local Interpretable Model-agnostic Explanations (LIME) have been used to highlight influential words or structural patterns in phishing emails (Ribeiro, Singh, & Guestrin, 2016). Probabilistic approaches, particularly Bayesian models, further enhance explainability by explicitly representing uncertainty and causal dependencies among features. Such models enable users to observe how individual elements contribute to the overall risk assessment, making them well suited for educational contexts where understanding decision logic is as important as detection accuracy.

Despite these advantages, applying XAI in education introduces challenges. Overly complex explanations may overwhelm learners, while overly simplified explanations risk misrepresenting system behavior (Mellon &

Worrell, 2023). Consequently, educational XAI systems must balance interpretability, realism, and cognitive load to effectively support learning.

2.2 Bayesian Networks and Uncertainty

Bayesian Networks (BNs) are probabilistic graphical models that represent causal relationships between variables using directed graphs and conditional probability distributions (Pearl, 1988). Their defining strength lies in their capacity to model uncertainty explicitly, allowing probabilistic inference in environments characterized by incomplete information.

In cybersecurity, Bayesian Networks have been applied to risk assessment, intrusion detection, and attack modeling by integrating heterogeneous indicators into coherent probabilistic estimates (Xie et al., 2010; Wei & Dong, 2025). Within phishing analysis, BNs can capture how linguistic cues, contextual signals, and behavioral patterns jointly influence the likelihood that a message constitutes a phishing attempt.

From an educational perspective, Bayesian Networks offer distinct pedagogical advantages. By exposing conditional dependencies and posterior probability updates, they support the development of probabilistic reasoning and decision-making under uncertainty. Learners can observe how small changes in input features propagate through the model, altering risk assessments in a transparent and interpretable manner. This aligns closely with educational goals in cybersecurity, where uncertainty awareness and causal reasoning are critical competencies (Rjoub et al., 2023).

However, educational use of Bayesian Networks requires careful abstraction. Highly complex models may obscure learning objectives, while simplified models must still preserve meaningful causal structure. Prior work has shown that scenario-specific Bayesian models can effectively balance interpretability and realism, particularly when designed explicitly for training rather than operational deployment (Lappas et al., 2025a; Lappas et al., 2025b).

2.3 Adversarial Game Theory in Cybersecurity

Game theory provides a formal framework for analyzing strategic interactions between competing agents. In cybersecurity, adversarial game theory models the dynamic interaction between attackers and defenders, each adapting strategies under uncertainty and resource constraints (Huang & Zhu, 2018). Rather than static classification problems, cybersecurity challenges often involve iterative adaptation, deception, and counter-strategy.

Educational applications of adversarial game theory typically employ role-based simulations, where learners assume attacker or defender perspectives to experience strategic trade-offs firsthand (Prebot, Du, & Gonzalez, 2023). Such simulations foster deeper understanding of attacker behavior, system vulnerabilities, and the limitations of defensive mechanisms.

When combined with explainable probabilistic models, adversarial games can transform detection systems into interactive learning environments. Learners do not merely observe outcomes but engage in strategic adaptation informed by transparent feedback. This approach supports the development of adversarial reasoning, perspective-taking, and adaptive decision-making, skills that are essential in contemporary cybersecurity practice (Pawlick et al., 2019; Shashkov et al., 2023).

3. Problem Statement

The integration of Explainable Artificial Intelligence (XAI), Bayesian Networks, and adversarial game theory offers a promising foundation for cybersecurity education. XAI enhances transparency in algorithmic decision-making, enabling learners to understand why specific messages are classified as phishing. Bayesian Networks complement this capability by providing a formal mechanism for quantifying uncertainty and modeling causal relationships among linguistic and contextual features. Adversarial game theory further extends this framework by positioning learners as strategic agents who must adapt their behavior based on probabilistic feedback (Huang & Zhu, 2018; Rjoub et al., 2023).

This combined perspective moves beyond traditional cybersecurity training approaches, which often emphasize theoretical instruction or isolated simulations. Instead, it supports learning environments where explanation, uncertainty reasoning, and strategic interaction are tightly coupled. Within such environments, learners can explore how their decisions influence detection outcomes, reflect on system behavior, and iteratively refine their strategies.

Despite advances in these individual domains, significant gaps remain in the existing literature. Most XAI research in cybersecurity focuses on technical interpretability methods, with limited attention to their pedagogical integration and impact on learner decision-making (Rjoub et al., 2023). Bayesian Networks are predominantly employed for operational risk assessment and detection, rather than as interactive educational tools that allow learners to experiment with probabilistic models and uncertainty-driven scenarios (Frigault & Wang, 2008; Xie et al., 2010). Similarly, while adversarial game theory has informed defensive and offensive cybersecurity strategies, its systematic application in educational settings, particularly in combination with explainable probabilistic models, remains underexplored (Huang & Zhu, 2018; Pawlick et al., 2019).

These limitations highlight the need for an integrated educational framework that unifies explainability, probabilistic uncertainty modeling, and adversarial interaction. Research Part

3.1 Design and Implementation of the Educational Intervention

The educational intervention was structured as a concise, multi-phase instructional sequence, explicitly designed to operationalize theoretical constructs from explainable artificial intelligence, Bayesian probabilistic reasoning, and adversarial game dynamics within a cybersecurity training context. The sequence comprised:

- Theoretical lectures introducing the taxonomy of phishing attacks and the psychological underpinnings of social engineering techniques.
- Conceptual exposition of Bayesian networks, focusing on their capacity to model uncertainty and causal relationships in automated decision-making systems.
- Experiential learning through a hands-on session utilizing the custom-developed serious game platform.

The serious game functioned as the integrative nexus where abstract theoretical principles were translated into actionable skills. Within this environment, participants engaged in a series of scaffolded activities, including:

- Bayesian reasoning tasks: Interpreting probabilistic phishing scores, analyzing the influence of specific linguistic and contextual features, and exercising decision-making under uncertainty.
- Adversarial, game-based decision-making: Iteratively modifying phishing strategies in direct response to transparent, explainable feedback from the Bayesian detection software.
- Reflective practice: Critically comparing human intuition with algorithmic judgment, thereby fostering metacognitive awareness of both strengths and limitations in human-AI collaboration.

A central pedagogical emphasis was placed on elucidating how subtle linguistic modifications propagate through the probabilistic model, altering posterior risk assessments and detection outcomes. This approach foregrounded the dynamic interplay between attacker adaptation and defender detection logic.

The intervention targeted the following learning objectives:

- To cultivate a nuanced understanding of uncertainty and explainability in automated cybersecurity decision systems.
- To develop adversarial and strategic thinking by anticipating and circumventing detection mechanisms.
- To enhance awareness of social engineering tactics, moving beyond superficial indicators to deeper, context-sensitive threat analysis.

By simultaneously exposing participants to both the attack surface and the internal logic of the detection system, the framework advanced a human-centered, explainable approach to phishing defense. This design not only demystified the operation of probabilistic classifiers but also empowered learners to reason strategically about both offensive and defensive dimensions of cybersecurity.

3.2 Research Context

The presented work adopts a design-based case study approach within a higher-education cybersecurity training context. It was held on October 20, 2025, at the Polish Air Force University, on October 28, 2025, at the Portuguese Air Force Academy and on December 5, 2025, at the Hellenic Air Force Academy as part of the "Cyber Warfare" training within the "European Union Military Erasmus" program. A total of 126 students from six different European countries, aged 19 to 21, participated in the training. All participants were cadets from military academies. The module spanned six training hours at each place.

Rather than evaluating a static detection model, the focus is on the human-algorithm interaction and on how explainable probabilistic feedback can support adversarial reasoning and cybersecurity awareness. The serious game was deployed as part of a structured educational intervention and served both as a training instrument and as a data collection mechanism for observing participant behavior under uncertainty.

The study explores how learners interpret, exploit, and adapt to Bayesian phishing detection mechanisms when explicitly exposed to their internal reasoning.

3.3 Framework design

The proposed educational framework synthesizes Bayesian explainability with adversarial game dynamics, fostering an environment where learners alternate between attacker and defender roles. This dual-role approach operationalizes theoretical constructs from explainable AI and game theory, enabling participants to experience both the strategic creation of phishing messages and the probabilistic evaluation of such attacks.

At the core of the system are three interdependent modules:

1. Scenario & Target Profiling Module: participants are provided with a simulated victim profile constructed from fictionalized social media data, closely emulating realistic open-source intelligence (OSINT) sources commonly exploited in social engineering. More specifically, the profile depicted a German fighter pilot who expressed a passion for running and a keen interest in the military aviation history. The individual also disclosed specific dietary habits and preferences for particular brands of running shoes. Frequent references were made to a close friend named Josh and to the individual's daughter. All such information was publicly accessible through the designated social media platform, thereby grounding the exercise in authentic adversarial contexts and supporting the development of targeted attack strategies by the participants (figure 1).

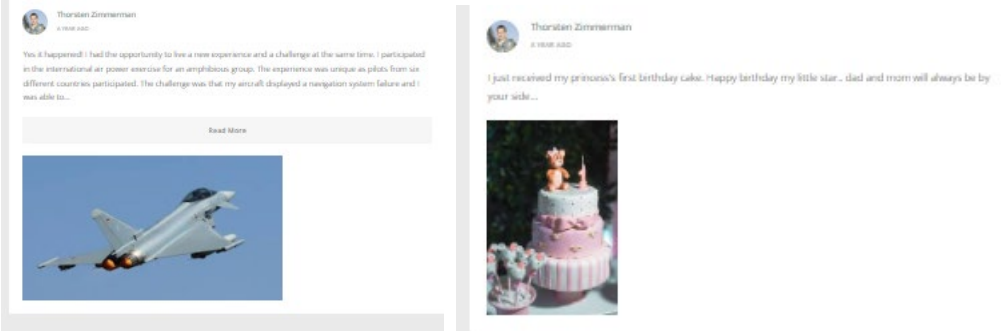

Figure 1: Images from the hypothetical person's social media profile

2. Bayesian Anti-Phishing Software: In the context of the Bayesian Anti-Phishing Software, submitted emails were evaluated by a Bayesian Network-based classifier, which provided both a probabilistic phishing likelihood score, quantifying the risk associated with each message and an interpretable explanation highlighting the lexical and contextual features most influential in the classification process. Each detected feature was mapped to a corresponding node within the Bayesian network, facilitating the propagation of uncertainty and modeling conditional dependencies between linguistic cues and phishing risk.

The construction of the Bayesian network was informed by qualitative data collected over the previous years, during which the same phishing scenario and target profile were used in instructional settings (Lappas & Karampelas, 2023). Emails authored by participants in these earlier cohorts were systematically analyzed and categorized according to the persuasive elements and strategies employed to attract the target. The frequency and distribution of these elements were quantified, enabling the derivation of conditional probabilities for the network. As a result, the Bayesian model was specifically tailored to predict the likelihood that an email would constitute a phishing attempt within this scenario. The network was implemented in Python and programmed to recognize keywords within the email text that corresponded to the semantic meaning of each node. When such keywords were detected, the respective node was activated, thereby influencing the overall risk assessment produced by the model (figure 2).

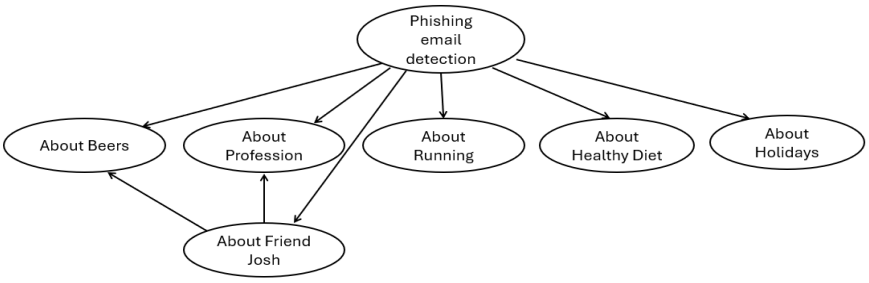

Figure 2: The bayesian network used in the educational framework

3. Explainability & Feedback Module: The system transparently exposes detected keywords and their respective impact on posterior probabilities, transforming the classifier from a traditional black-box filter into an interactive, explainable adversarial opponent. This module enables learners to iteratively refine their strategies based on actionable feedback, reinforcing the link between feature manipulation and detection outcomes.

Figures 3 and 4 illustrate a representative example of adversarial adaptation, showing how a participant iteratively revised a phishing email in response to explainable Bayesian feedback, resulting in reduced detection probability through targeted lexical modifications while preserving semantic intent. Furthermore, semantic similarity analysis between the two email versions yielded a high score of 0.965, indicating that the core message and communicative intent were preserved despite localized linguistic modifications. This example demonstrates the capacity of participants to interpret explainable Bayesian feedback and engage in adversarial reasoning. Rather than resorting to superficial deletion of flagged terms, the trainee employed strategic substitutions and contextual adjustments, successfully reducing the detectability of the phishing attempt while maintaining message coherence. Such behavior exemplifies the adversarial learning objectives of the proposed framework, wherein learners internalize the probabilistic logic of automated detection and adapt their strategies accordingly.

Figure 3: Image from the anti-phishing software that the trainee was viewing in the first effort

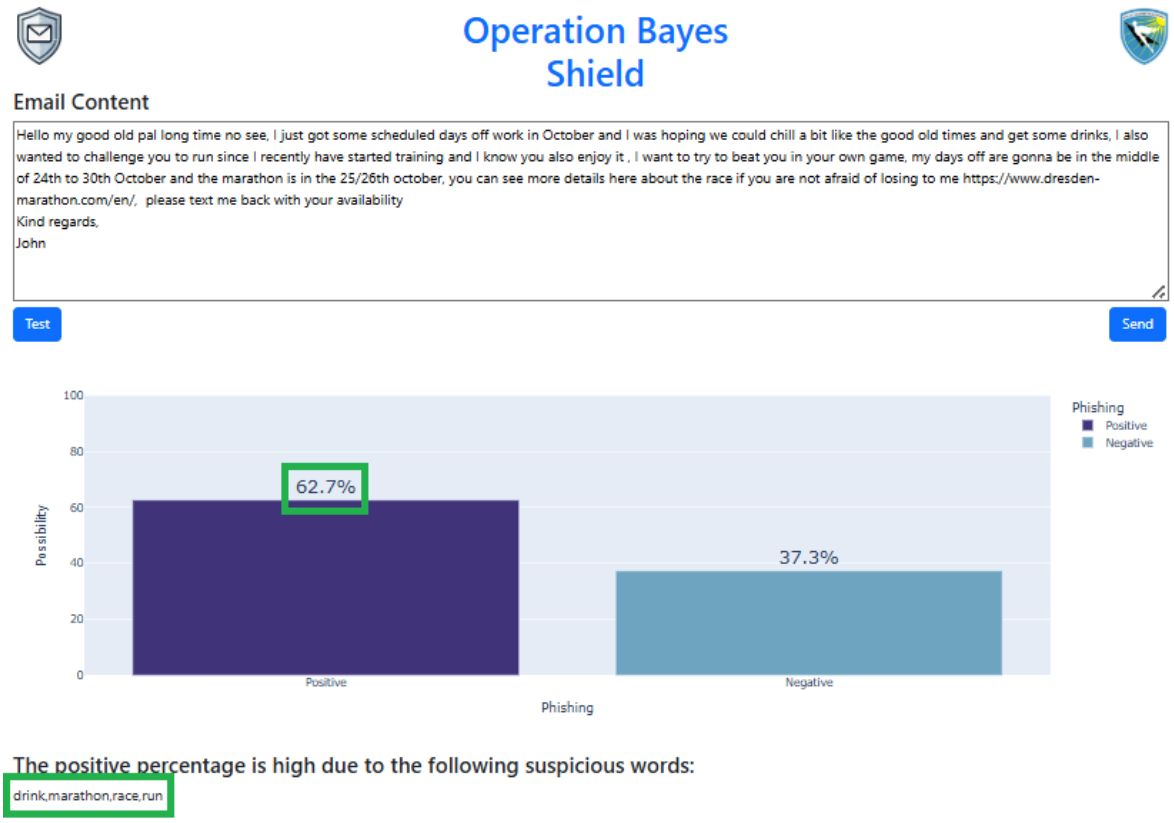

Figure 4: Image from the anti-phishing software that the trainee was viewing in the second effort

3.4 Scenarios and Roles

The serious game scenario is designed to authentically simulate a targeted spear-phishing attack, immersing participants in realistic adversarial dynamics. Learners alternate systematically between two roles (figure 4):

- Attacker: Participants craft phishing emails tailored to the simulated target's interests, behavioral patterns, and social connections, leveraging open-source intelligence data provided by the system.
- Defender: Participants analyze how the Bayesian anti-phishing software evaluates their crafted messages, interpreting the system's probabilistic risk assessments and the rationale behind flagged features.

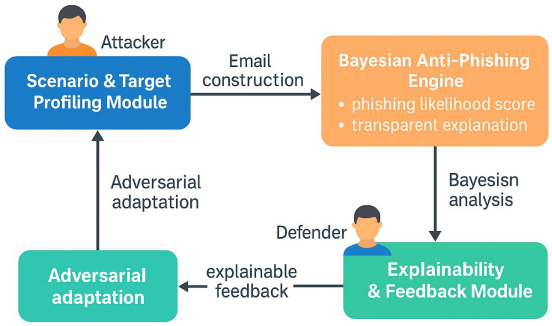

Figure 4: Roles in Bayesian Anti-phishing Serious Game

A key instructional constraint requires participants to iteratively revise a single email instance, maintaining semantic continuity while strategically modifying linguistic and contextual elements. This design choice discourages superficial keyword removal and instead promotes meaningful adversarial adaptation, mirroring real-world attacker behavior.

To further enhance strategic engagement, the number of permissible resubmissions is deliberately limited. This introduces a layer of decision-making pressure, compelling learners to reflect critically on each modification and to anticipate the evolving detection logic of the Bayesian system.

Through this dual-role, iterative process, the scenario operationalizes the theoretical principles of adversarial game theory and explainable AI, fostering both offensive and defensive cybersecurity competencies within a controlled, feedback-rich environment.

3.5 Results of Trainee Submissions

The analysis of participant submissions within the serious game framework was conducted across three complementary dimensions: (a) changes in phishing detection probability between successive email revisions, (b) modification of detected keywords across iterations, and (c) semantic similarity between initial and revised email versions.

Quantitative results revealed a consistent reduction in the Bayesian phishing probability between the first and second submissions (from 55.36% to 51.28%), indicating effective adversarial adaptation following explainable feedback. The probability stabilized in the third submission, suggesting convergence toward an optimal strategy. This trend demonstrates that learners successfully internalized the probabilistic feedback and adapted their strategies to minimize detectability, fulfilling a core objective of the educational framework.

A qualitative examination of the detected keywords further substantiates this interpretation. Rather than simply removing highlighted terms, participants engaged in targeted lexical adaptations, replacing flagged keywords with semantically related or contextually neutral alternatives. This behavior reflects adversarial reasoning aligned with realistic attacker strategies, whereby the core message is preserved while overtly suspicious signals are minimized. The evolution of keyword sets across submissions illustrates that learners internalized which categories of terms (e.g., urgency, reward, brand-related cues) significantly contributed to higher phishing risk assessments.

Semantic similarity analysis supports these findings, with similarity scores between initial and revised emails remaining consistently high (mean similarity 0.886 between first and second, and 0.954 between second and third submissions). This indicates that participants preserved the core intent and communicative purpose of their emails, opting for strategic linguistic refinement rather than superficial or trivial content removal. Evaluation of the Educational Intervention

To assess the pedagogical impact of the intervention, participants completed a post-activity questionnaire using a 5-point Likert scale (1 = strongly disagree, 5 = strongly agree). The overall mean score across all items was 4.28, indicating a highly positive evaluation of the learning activity. The internal consistency of the scale was excellent (Cronbach's $\alpha = 0.96$), supporting the reliability of the findings.

The results demonstrate substantial learning gains. For example, 84.8% of participants agreed or strongly agreed that they understood phishing techniques better after playing the game, while 94.0% reported increased ability to recognize social engineering tactics in emails. Similarly, 87.8% felt that the game helped them understand how attackers craft deceptive messages, and 81.8% considered the game a useful method for learning about phishing. Notably, 90.9% believed that this activity would benefit future students, and 84.8% agreed that it is more effective than lecture-only learning.

Regarding system transparency and explainability, 78.8% agreed or strongly agreed that the detected keywords made the system's reasoning transparent, and 81.8% understood why the system changed the phishing probability. Furthermore, 78.8% reported that the game helped them understand how automated filters make decisions.

General satisfaction was high, with 84.8% expressing overall satisfaction with the activity and 84.8% stating they would recommend the game to others. Areas with higher neutrality (e.g., ability to identify highly personalized phishing or clarity of feedback) suggest opportunities for further improvement.

In summary, the intervention was perceived as highly effective in enhancing both conceptual understanding and practical skills related to phishing detection, while the integration of explainable AI elements contributed to learners' confidence and strategic thinking.

3.6 Discussion

The study demonstrates the pedagogical value of integrating explainable Bayesian models within adversarial serious games for cybersecurity education. Observed reductions in phishing probability and subsequent stabilization indicate that participants adopted informed strategies rather than relying on trial-and-error, reflecting an internalization of the Bayesian model's logic and realistic attacker adaptation.

High semantic similarity scores confirm that learners preserved message intent while making targeted lexical adjustments, avoiding superficial keyword deletion. This suggests the development of strategic manipulation skills grounded in understanding, a key educational objective.

Questionnaire responses reinforce these findings, highlighting improved comprehension of phishing techniques and the probabilistic reasoning behind automated detection systems, addressing a gap in traditional training approaches. Explainable feedback acted as a cognitive scaffold, transforming the detection system into an interactive learning counterpart and fostering trust and strategic reasoning.

Overall, the combination of adversarial role-play and transparent probabilistic feedback cultivates critical competencies such as uncertainty reasoning, attacker–defender perspective-taking, and adaptive decision-making, skills essential in cybersecurity environments facing AI-driven threats.

One limitation relates to the question of learning transfer.. The present study assessed strategic adaptation and uncertainty reasoning within a sandboxed, simulated phishing scenario, yet did not measure whether these competencies translate to real-world detection of unsolicited phishing attempts. The extent to which learners generalize adversarial understanding to authentic environments remains an open empirical question.

On the other hand, a potential concern associated with adversarial training approaches is the dual-use nature of the skills developed: participants learn techniques for crafting more effective phishing emails. This mirrors established practices in cybersecurity education, where offensive methodologies, such as penetration testing and red-team exercises, are routinely taught to cultivate defensive competence. In this context, the controlled environment, limited scope of the scenario, and absence of real operational impact ensure that the adversarial component serves strictly educational purposes.

Future work may include developing the framework in more diverse student populations, including non-military academic settings or professional training environments where it can offer more relevant information and generalizability of conclusions.

4. Conclusion

This paper introduced a Bayesian-driven serious game for anti-phishing training that integrates explainable AI principles with adversarial game dynamics. By exposing learners to both the construction of phishing emails and the transparent reasoning of a Bayesian detection system, the proposed framework advances a human-centered approach to cybersecurity education.

Empirical results from interaction logs and post-activity evaluation indicate that participants were able to strategically adapt their behavior in response to explainable probabilistic feedback. The observed reduction in phishing detection probability across submissions, combined with high semantic similarity between email versions, demonstrates meaningful adversarial learning rather than trial-and-error behavior. Moreover, participant feedback confirms that the explainability of the Bayesian model played a central role in enhancing understanding, confidence, and trust in automated detection mechanisms.

From a cybersecurity perspective, the study contributes evidence that explainable probabilistic systems can serve not only as defensive tools, but also as effective educational instruments for developing strategic thinking and uncertainty awareness. The findings suggest that exposing users to the internal logic of detection mechanisms may improve long-term resilience against sophisticated, AI-assisted phishing attacks.

Future work will focus on extending the framework to additional attack scenarios, incorporating longitudinal measurements, and exploring how learner-generated data can further refine Bayesian detection models. Overall, the proposed approach demonstrates that gamified, explainable Bayesian systems offer a promising pathway for advancing human-centered cybersecurity training.

Ethics Declaration: This study did not involve the collection of personal or sensitive data and was conducted within a controlled educational setting. Therefore, formal ethical clearance was not required. All participants were informed about the purpose of the activity and consented to the use of anonymized data for research purposes.

AI Declaration: Artificial Intelligence (AI) tools were used exclusively for language enhancement purposes, specifically to improve the clarity and fluency of English in the manuscript. No AI tools were employed for data analysis, interpretation, or generation of research content.

References

Calzarossa, M. C., Giudici, P., & Zieni, R. (2025). An assessment framework for explainable AI with applications to cybersecurity. *Artificial Intelligence Review, 58*, Article 150. https://doi.org/10.1007/s10462-025-11141-w

Frigault, M., & Wang, L. (2008). Measuring network security using Bayesian network-based attack graphs. *Proceedings of the 32nd Annual IEEE International Computer Software and Applications Conference*, 698–703.

Huang, L., & Zhu, Q. (2018). Dynamic Bayesian games for adversarial and defensive cyber deception. *arXiv preprint* arXiv:1808.02161.

Infosecurity Magazine. (2025, October 16). AI attacks surge as Microsoft process 100 trillion signals daily. *Infosecurity Magazine*. https://www.infosecurity-magazine.com/news/microsoft-process-100-trillion/

Kavya, S., & Sumathi, D. (2025). Staying ahead of phishers: A review of recent advances and emerging methodologies in phishing detection. *Artificial Intelligence Review, 58*(50).

Lappas, D., & Karampelas, P., (2023). Designing an Email Attack by Analysing the Victim's Profile. An Alternative Anti-Phishing Training Method. European Conference on Cyber Warfare and Security. 22. 257-266.

Lappas, D., & Karampelas, P. (2024). Developing Personalized Training Material using AI Tools for Anti-phishing Cyber-security Awareness. Scientific Journal of Safety and Logistics, 2(2).

Lappas, D., Karampelas, P., & Fesakis, G. (2025a). Enhancement of Phishing Email Detection with Bayesian Networks. A Cyber Security Training Module. European Conference on Cyber Warfare and Security, 24(1):328-337.

Lappas, D., Karampelas, P., & Fesakis, G. (2025b). Teaching Bayesian Reasoning as a Pathway Toward Active Thinking and Explainable AI. International Conference on AI Research, 5(1), 218–226.

Masunda, M. (2025). Adaptive threat intelligence systems for real-time detection and mitigation of sophisticated cyber attacks in enterprise networks. *International Journal of Advance Research Publication and Reviews, 02*(05), 470–491.

Mellon, J., & Worrell, C. (2023, November 20). Explainability in cybersecurity data science. *Software Engineering Institute Insights*.

Mutlutürk, M., Wynn, M., & Metin, B. (2024). Phishing and the human factor: Insights from a bibliometric analysis. *Information, 15*(10), 643.

Pawlick, J., Colbert, E., & Zhu, Q. (2019). A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy. *ACM Computing Surveys, 52*(4), 1–28.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Polotskaya, K., Muñoz-Valencia, C.S., Rabasa, A., Quesada-Rico, J.A., Orozco-Beltrán, D., & Barber, X. (2024). Bayesian Networks for the Diagnosis and Prognosis of Diseases: A Scoping Review. *Mach. Learn. Knowl. Extr., 6*, 1243-1262.

Prebot, B., Du, Y., & Gonzalez, C. (2023). Learning about simulated adversaries from human defenders using interactive cyber-defense games. *Journal of Cybersecurity, 9*(1), tyad022.

Push Security. (2025, April 23). Phishing detection is broken: Why most attacks feel like a zero day. *BleepingComputer*. https://www.bleepingcomputer.com/news/security/phishing-detection-is-broken-why-most-attacks-feel-like-a-zero-day/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Rjoub, G., Bentahar, J., Abdel Wahab, O., Mizouni, R., Song, A., Cohen, R., Otrok, H., & Mourad, A. (2023). A survey on explainable artificial intelligence for cybersecurity. *IEEE Transactions on Network and Service Management*.

Shashkov, A., Hemberg, E., Tulla, M., & O'Reilly, U.-M. (2023). Adversarial agent-learning for cybersecurity: A comparison of algorithms. *The Knowledge Engineering Review, 38*, e3.

Sridar, S., et al. (2025). Modeling interdependent cybersecurity threats using Bayesian networks: A case study on in-vehicle infotainment systems. *arXiv preprint*. https://arxiv.org/pdf/2505.09048

Wei, X., & Dong, Y. (2025). A hybrid approach combining Bayesian networks and logistic regression for enhancing risk assessment. *Scientific Reports, 15*, 10291.

Xie, M., Hu, J., & Slay, J. (2010). Evaluating host-based anomaly detection systems: A Bayesian approach. *Computers & Security, 29*(4), 421–431.