Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 19 No. 23 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijim.v19i23.58679

PAPER

Responsible AI in Security Pedagogy: A Proactive Ethical Framework for Mobile Learning and Simulation Platforms

Luca Güttner[1] (✉), **Raphael Röttinger**[2]

[1]University of Münster, Münster, Germany

[2]University of Krems, Krems, Austria

l_guet02@uni-muenster.de

ABSTRACT

The proliferation of artificial intelligence (AI)-driven mobile learning systems for corporate security training presents new ethical challenges. This paper addresses the urgent need for a proactive ethical framework to govern these technologies. It systematically analyzes the key ethical dilemmas, including the tension between security monitoring and employee privacy, the risk of discriminatory outcomes from algorithmic bias, the ambiguity of accountability for AI-driven errors, and the lack of transparency inherent in 'black box' systems. This study proposes an actionable framework for the responsible design and deployment of these platforms. Drawing upon foundational principles from the UNESCO Recommendation on the Ethics of AI, the framework translates concepts like fairness, accountability, and human oversight into practical guidance, offering a tool for developers and practitioners to embed ethics into the entire technology lifecycle.

KEYWORDS

artificial intelligence (AI) ethics, mobile learning, security training, ethical framework, algorithmic bias

1 INTRODUCTION: THE PROLIFERATION OF AI IN CORPORATE SECURITY TRAINING

The landscape of corporate security training is undergoing a fundamental transformation, driven by the dual pressures of an increasingly sophisticated threat environment and the pervasive integration of artificial intelligence (AI) into enterprise solutions. Traditional security awareness programs, usually characterized by static, one-size-fits-all modules and infrequent updates on technological developments, are proving insufficient against dynamic adversaries who employ ever-changing tactics. In response, organizations are rapidly adopting AI-driven

Güttner, L., Röttinger, R. (2025). Responsible AI in Security Pedagogy: A Proactive Ethical Framework for Mobile Learning and Simulation Platforms. *International Journal of Interactive Mobile Technologies (iJIM)*, 19(23), pp. 134–148. https://doi.org/10.3991/ijim.v19i23.58679

Article submitted 2025-08-15. Revision uploaded 2025-10-04. Final acceptance 2025-10-11.

pedagogical platforms to deliver smarter, more efficient, and adaptive learning experiences that are personalized, scalable, and cost-effective [1].

This new generation of security training leverages AI to analyze employee roles, behaviors, and knowledge gaps, creating tailored educational pathways that address specific, relevant risks. The delivery of this training is increasingly optimized for the modern workforce through mobile-first platforms that provide microlearning modules designed to fit seamlessly into an employee's day rather than disrupt it. These platforms promise to enhance knowledge retention, improve engagement through gamification, and provide automated progress monitoring to satisfy complex compliance and regulatory standards. From hyper-realistic phishing simulations to adaptive learning paths, AI is revolutionizing how organizations build a security-conscious culture and fortify their human firewall against threats like the aforementioned phishing, Business Email Compromise (BEC), and insider risks [2].

1.1 The emergence of a new ethical frontier

This rapid technological advancement, however, has significantly outpaced the development of corresponding ethical and governance frameworks [1]. The very capabilities that make these AI tools so powerful, namely their ability to monitor behavior, personalize content, and assess risk at an individual level, introduce a new frontier of ethical challenges that are unique to the high-stakes context of security pedagogy. The deployment of these systems raises profound questions about the balance between organizational security and employee privacy, the potential for algorithmic bias to create discriminatory outcomes [3], the diffusion of accountability when automated systems err, and the lack of transparency inherent in complex 'black box' models [4].

Existing governance structures are often inadequate to address these specific challenges due to their high change rate. Generic corporate AI policies may lack the nuance required for educational applications, while traditional ethical frameworks for education technology may not fully account for the surveillance and risk-assessment functions embedded within modern security training platforms. This gap creates a significant risk that organizations, in their pursuit of a more solid security posture, may inadvertently deploy systems that are invasive, unfair, or unaccountable, thereby eroding employee trust and creating new forms of legal and reputational liability [5].

1.2 Methodology and research goal

This paper thus argues for the urgent need to move beyond reactive, ad-hoc ethical considerations and establish a proactive, domain-specific framework for the responsible design, development, and deployment of AI in security pedagogy. The central contribution of this study is the proposal of the Proactive Ethical Framework for Responsible Security Pedagogy (PERSP). This framework is intended as an actionable, lifecycle-based tool designed to guide developers, security professionals, and corporate educators. To ensure its universal applicability and grounding in fundamental human rights, the PERSP framework is explicitly anchored in the core values and principles of the UNESCO Recommendation on the Ethics of AI, the first global standard-setting instrument in this domain [6], [7]. By translating these global

principles into concrete guidelines tailored for mobile learning and simulation platforms, PERSP provides a systematic methodology for embedding ethics into the technological architecture and organizational policies governing these powerful new tools.

1.3 Roadmap of the article

The remainder of this paper is structured to build a comprehensive case for the PERSP framework: Section 2.0 will provide a detailed analysis of the current AI-powered security pedagogy, examining the key technologies and their applications. Section 3.0 then presents a systematic taxonomy of the core ethical dilemmas that arise from these applications, focusing on privacy, bias, accountability, and transparency. Section 4.0 establishes the normative foundation for the proposed framework by analyzing the UNESCO Recommendation and synthesizing its principles with other leading governance models such as the *EDUCAUSE AI Ethical Guidelines*, the *NIST AI Risk Management Framework (AI RMF)*, and Google's *Secure AI Framework (SAIF)*. Section 5.0 formally presents the PERSP framework, detailing its structure and lifecycle-based guidelines. Section 6.0 demonstrates the framework's practical utility by applying it to two illustrative real-world scenarios. Finally, Section 7.0 discusses the challenges to implementation and outlines directions for future research.

2 THE EVOLVING LANDSCAPE OF AI-POWERED SECURITY PEDAGOGY

The shift from passive, compliance-driven education to active, adaptive, and data-informed pedagogy is characterized by three key technological advancements: personalization of learning content, the creation of hyper-realistic simulations, and the use of sophisticated behavioral analytics for continuous monitoring and assessment [8]. Understanding these capabilities is essential for appreciating both the pedagogical promise and the ethical complexity of these systems.

Its capacity to deliver personalization at scale removes the limitations of generic, one-size-fits-all training modules and is perhaps the primary advantage of AI in security pedagogy [9]. AI algorithms analyze a wide array of data points, including an employee's job role, department, past training performance, and even online behaviors to construct customized learning paths that are directly relevant to their specific risk profile. For instance, a member of the finance department might receive a series of phishing simulations centered on invoice and wire transfer fraud and ATM manipulation [10], while an IT administrator is presented with scenarios related to vulnerability management and system hygiene. This targeted approach ensures that training changes from a formal, abstract exercise to a relevant and engaging experience. Based on an individual's performance in quizzes and simulations, the AI can adjust the difficulty and content of subsequent modules in real-time [11]. An employee who struggles to identify basic phishing indicators will be provided with foundational modules and reinforcement exercises, whereas a more advanced user can be challenged with complex, multi-stage social engineering scenarios. This adaptive capability is crucial for pedagogical effectiveness, as it ensures that learners are consistently and appropriately challenged, which enhances knowledge retention and prevents both boredom and frustration. AI's ability to adapt to the user thus has some educational advantages over human educators who have to operate

based on prepared material that can only be customized to a certain degree (if at all) in a given situation [8]. It can do this because it can produce new material on the spot, even if it takes the complex form of a simulation or a game.

AI-powered platforms can generate highly realistic and context-aware phishing emails that mirror the latest tactics used by malicious actors, incorporating details specific to the organization or the individual employee to increase their believability. This capability is being extended through the use of generative AI to create next-generation training threats, such as deepfake voice messages from executives or synthetic video calls, designed to prepare employees for the increasing sophistication of AI-powered attacks they will face in the wild. This is a critical development, as attackers themselves are leveraging generative AI to automate and personalize their campaigns at an unprecedented scale, making traditional, easily spotted phishing emails obsolete.

2.1 Behavioral analytics and automated monitoring

AI systems automate the tracking of training progress by analyzing quiz results, simulation outcomes, and user interactions within the platform. This data provides organizations with a view of their collective security awareness, helps identify persistent knowledge gaps, and generates the detailed reports necessary to meet industry and regulatory compliance standards. However, the application of AI-driven monitoring extends beyond the training module itself. Advanced platforms now incorporate behavioral analytics that track risky online activities on corporate networks, such as clicking on unknown links, reusing passwords, or attempting to access unauthorized data [12]. This functionality represents a significant expansion of the platform's role. Initially, this data is used for pedagogical purposes to further customize training and provide targeted interventions for high-risk employees. Yet, the same technology is increasingly being positioned as a tool for proactive insider threat detection, capable of flagging anomalous behavior that may indicate malicious intent or a compromised account.

This evolution reveals a critical convergence of what were once distinct organizational functions. The technologies that enable realistic training simulations are the very same ones used by attackers to craft sophisticated social engineering campaigns. This creates a dual-use dilemma, where the training platform itself, if not properly secured and governed, could become a tool for malicious actors to hone their skills or launch internal attacks. A security training platform that can generate deepfake voice messages of a CEO for a phishing simulation could, in the wrong hands, be used to generate a deepfake for an actual fraud attempt.

Simultaneously, the progression from simple progress tracking of quiz results to comprehensive behavioral analytics for insider threat detection marks a fundamental shift in purpose [12]. Security training platforms are no longer purely educational; they are becoming integrated components of an organization's active security monitoring and surveillance infrastructure. The data collected to assess a learner's knowledge gaps is now being used to assess their risk level to the organization. This functional convergence creates a concerning ethical tension between the goals of pedagogy and the imperatives of security enforcement, particularly concerning employee privacy, consent, and the potential for a 'chilling effect' on workplace culture, which "impairs individuals in the exercise of their personal liberty to make decisions that are truly their own" [13]. Any responsible framework

must therefore establish transparent boundaries to govern this dual-purpose data collection and use, ensuring that the line between education and surveillance is not crossed without explicit justification and oversight.

3 A TAXONOMY OF ETHICAL DILEMMAS IN AI-DRIVEN SECURITY TRAINING

The effectiveness of AI-powered training is predicated on its ability to collect and analyze vast quantities of granular data about employee behavior. When these systems move beyond quiz scores to monitor keystrokes, the websites employees visit, and even facial expressions through laptop-integrated webcams, or analyze the sentiment of internal communications, they transform the learning environment into a site of continuous surveillance. This creates a 'panoptic' effect, where employees may feel constantly observed, leading to increased stress and a potential erosion of trust in their employer [14]. Surveys indicate that a majority of employees are already concerned about AI-enabled data tracking in the workplace, a fear that these systems can easily exacerbate.

A significant risk associated with this intensive data collection is function creep, the gradual and often unannounced expansion of data usage beyond its originally stated purpose. Data gathered to personalize a phishing simulation or identify a knowledge gap could later be repurposed for employee performance evaluations, productivity scoring, or even disciplinary actions, often without the employee's specific, informed consent for these secondary uses [13]. This practice not only violates user expectations but also runs afoul of foundational data protection principles like purpose limitation, which are enshrined in regulations such as the General Data Protection Regulation (GDPR). Without suitable governance and transparent policies, organizations risk creating a surveillance culture that is both ethically fraught and legally non-compliant.

3.1 Encoded inequity: Algorithmic bias and discrimination

AI training systems, particularly those that assign risk scores or create adaptive learning paths, are thus at risk of becoming powerful vectors for unfairness and discrimination. Algorithmic bias refers to systematic and repeatable errors in an AI system that result in prejudiced outcomes, where specific individuals or groups are unjustly privileged or disadvantaged [3]. This bias does not have to arise from malicious intent but is typically inherited from flaws in the data and design of the system, often reflecting and amplifying existing societal prejudices. Bias can be introduced at multiple stages of the AI lifecycle: The most common source of bias is the data used to train the machine learning model. If historical security incident data shows a correlation (spurious or otherwise) with a particular demographic group (e.g., based on age, tenure, or geographic location), the AI may learn to unfairly associate that group with higher security risk. Similarly, if the data is not representative and under-samples certain groups, the model's performance for those groups may be significantly worse, leading to inequitable assessments.

Bus bias can also be embedded by developers themselves, either through their own unconscious assumptions or through the use of proxy variables. For example, an algorithm might not use a protected attribute like race directly, but it might use

a variable like zip code, which can be highly correlated with race, leading to a discriminatory outcome by proxy. If a biased AI system then flags individuals from a certain group as high-risk more often, and this leads to them receiving more scrutiny or negative attention, this can generate new data that appears to validate the AI's original biased assessment [3].

For example, Amazon's AI recruiting tool was found to be biased against female candidates because it was trained on historical, male-dominated resume data [15]. The same potential for harm exists within security pedagogy; an employee could be unfairly labeled a 'high risk,' denied opportunities, or subjected to undue scrutiny based on a biased algorithmic assessment.

3.2 The accountability gap: Assigning responsibility for AI-driven errors

When an AI-driven training system makes a harmful error, like unfairly penalizing an employee or failing to train them on a critical threat that leads to a breach, determining who is responsible presents a formidable challenge. The complexity and opacity of many advanced AI models, often referred to as 'black boxes,' make it difficult to trace the specific cause of an erroneous output [4]. This creates a so-called accountability gap or a diffusion of responsibility across the many actors in the AI value chain [16]. This value chain includes the original developers of the AI model, the company that integrates the model into the training platform, and the organization that deploys the platform for its employees. A core ethical imperative, therefore, is to ensure that AI systems do not displace ultimate human responsibility and that liability can be determined. This requires the establishment of clear governance structures that define roles and responsibilities, mandate meaningful human-in-the-loop (HITL) oversight for high-stakes decisions, and ensure that a human is always accountable for the system's impact [17].

3.3 The opaque algorithm: Transparency and the 'Black Box' problem

Closely related to accountability is the challenge of transparency. For an AI system to be considered responsible, its operations and decision-making processes must be understandable to stakeholders [18]. It is useful to distinguish between two key concepts: 1) transparency refers to the clear communication of a system's capabilities, limitations, data sources, and purpose. An organization is transparent when it informs employees that an AI is being used to assess them and explains what data is being collected, and 2) explainability (or interpretability) refers to the ability to provide a human-understandable reason for a specific output or decision. While transparency might state that an AI generates a risk score, explainability answers why a particular employee received a high score.

As was already mentioned in the introduction, many powerful AI models, especially those based on deep learning, operate as 'black boxes,' where even their creators cannot fully articulate the reasoning behind a specific output [5], [19]. This makes it impossible to conduct meaningful audits for bias, as the factors driving the model's behavior are either hidden or entirely unknowable. Also, it renders accountability meaningless; if a decision cannot be explained, it cannot be effectively challenged or defended. Consequently, the adoption of Explainable AI (XAI) techniques, which are designed to make models more interpretable, should be

considered a prerequisite for deploying AI in high-stakes pedagogical applications like employee risk assessment [5].

The here-presented dilemmas are not isolated problems but are rather fundamentally interwoven, often creating numerous negative consequences as a result. An opaque, 'black box' algorithm (the transparency dilemma) that makes an unexplainable decision about an employee's risk level creates a vacuum of accountability, as no one can be held responsible for a decision that cannot be understood [5]. This lack of accountability and oversight allows for the possibility that the decision was driven by biased data that unfairly profiles the employee based on a protected demographic characteristic. This interconnectedness demonstrates that any effective ethical framework cannot address these issues individually but rather must adopt a holistic, systemic approach, recognizing that a solution in one area, such as mandating explainability, is a necessary condition for achieving progress in another, such as establishing clear accountability.

4 FOUNDATIONAL PRINCIPLES FOR A PROACTIVE ETHICAL FRAMEWORK

To address the complex, interconnected dilemmas identified in the previous section, an ethical framework is required that must be grounded in established, widely accepted normative principles while also being tailored to the specific context of security pedagogy [20]. This section builds the foundation for the PERSP framework by first conducting a comparative analysis of existing AI governance models, then anchoring the framework in the universal human-rights-based principles of the UNESCO Recommendation, and finally synthesizing these with practical insights from domain-specific guidelines.

4.1 A comparative analysis of existing AI governance frameworks

In AI governance there exist numerous frameworks developed by international bodies, government agencies, academic consortia, and industry leaders [21]. While there is considerable overlap, each framework has a unique focus, target audience, and set of priorities. A comparative analysis of the four most prominent frameworks, the UNESCO Recommendation on the Ethics of AI, the EDUCAUSE AI Ethical Guidelines, the NIST AI RMF, and Google's Secure AI Framework (SAIF), reveals both a consensus on core ethical values and the absence of a single framework that comprehensively addresses the unique intersection of pedagogy, security, and workplace surveillance. UNESCO provides the overarching ethical mandate, EDUCAUSE centers the learner, NIST offers a practical risk management process [22], and SAIF addresses the crucial security of the AI system [7], [23]. A truly effective framework for security pedagogy can be obtained by synthesizing these distinct strengths into a single, coherent model.

The PERSP framework uses the UNESCO Recommendation on the Ethics of AI as its primary normative anchor because, as the first global standard adopted by 193 member countries, it provides a powerful, consensus-based foundation rooted in international human rights law (Allahrakha, 2024). Its emphasis on protecting human dignity and fundamental freedoms makes it uniquely suited for a framework intended to govern technologies that directly impact employees in the workplace.

The framework is guided by the overarching values of Human Rights and Human Dignity and Ensuring Diversity and Inclusiveness. This means that all guidelines are designed to protect the fundamental rights of employees and ensure that AI training systems are equitable and accessible to all [7]. The following principles form the ethical pillars of the framework:

I. **Proportionality and Do No Harm:** AI systems must be necessary for a legitimate security goal, and their potential for harm must be rigorously assessed and mitigated.
II. **Fairness and Non-Discrimination:** Systems must be designed and audited to prevent biased or discriminatory outcomes.
III. **Safety and Security:** The AI systems themselves must be secure and resilient against attack or misuse.
IV. **Right to Privacy and Data Protection:** Employee privacy must be respected throughout the entire AI lifecycle.
V. **Human Oversight and Determination:** Ultimate responsibility for decisions and their consequences must always reside with humans, not algorithms.
VI. **Transparency and Explainability:** The operations and decisions of AI systems must be understandable to those they affect.
VII. **Responsibility and Accountability:** Mechanisms must be in place to successfully assign responsibility for AI system outcomes.

While the UNESCO Recommendation provides the foundational but complex 'why,' other frameworks offer the more practical 'how' that can be adopted for a more encompassing general ethical framework [6]. The PERSP framework combines these global principles with more granular, domain-specific insights to create a more actionable model. From the EDUCAUSE Guidelines, the framework incorporates a strong pedagogical focus. It adopts the principle of beneficence, reframing it for the corporate context: AI in security training must be demonstrably for the good of the employee-learner, enhancing their skills and safety, not merely serving as a tool for surveillance or compliance checking. From the NIST AIRMF, the framework adopts its structured, four-function process as its operational backbone. The PERSP guidelines are organized to align with the NIST functions of Govern (establishing policies and roles), Map (identifying risks and context), Measure (testing and validation), and Manage (mitigation and monitoring). This provides a clear, repeatable process for organizations to follow when implementing the framework. From industry frameworks such as Google's SAIF and IEEE standards, the framework integrates a robust focus on the technical Security and Reliability of the AI platform itself. This directly addresses the insight that the training platform can be a dual-use technology. It mandates that the security of the AI model, its data pipelines, and its deployment infrastructure be treated as a first-order ethical requirement, ensuring the tool designed to teach security does not itself become a security vulnerability.

5 THE PROACTIVE ETHICAL FRAMEWORK FOR RESPONSIBLE SECURITY PEDAGOGY

The PERSP is designed as a practical tool to guide organizations through the ethical complexities of deploying AI in security training. It ensures that ethical considerations are embedded proactively at every stage, from initial concept to eventual

decommissioning, rather than being treated as an afterthought or a compliance checklist. The PERSP framework is organized around four key stages of the AI lifecycle and five core ethical pillars.

AI Lifecycle Stages:

1. **Ethical Design and Inception:** The conceptualization and planning phase, before development begins.
2. **Responsible Development and Validation:** The technical phase of data preparation, model training, and testing.
3. **Transparent and Accountable Deployment:** The operational phase when the system is in use by employees.
4. **Responsible Decommissioning:** The final stage of retiring the system and its associated data.

Core Ethical Pillars:

1. **Fairness & Non-Discrimination:** Ensuring equitable treatment and outcomes for all employees.
2. **Transparency & Explainability:** Providing clarity on how the system works and why it makes specific decisions.
3. **Accountability & Redress:** Establishing clear lines of responsibility and mechanisms for challenging outcomes.
4. **Privacy & Security:** Protecting employee data and securing the AI system itself.
5. **Human Oversight & Autonomy:** Ensuring meaningful human control and respecting employee agency.

The intersection of these stages and pillars generates a series of actionable guidelines that provide concrete direction for practitioners.

5.1 Stage 1: Ethical design and inception

This initial stage is the most critical for proactively embedding ethics. Decisions made here have cascading effects throughout the entire lifecycle. PERSP's **Guideline 1** (Proportionality & Necessity) states that before committing to an AI solution, organizations must conduct and document a necessity assessment. This involves clearly defining the specific security problem to be solved and evaluating whether an AI-driven approach is proportional to the risk. The assessment must weigh the potential pedagogical benefits against the inherent risks to employee rights, considering less invasive alternatives first. **Guideline 2** (Multi-Stakeholder Consultation) demands that the design process must not be confined to security and IT departments. It must actively involve a diverse group of stakeholders, including representatives from HR, legal and compliance departments, ethics officers, and, crucially, a representative sample of employees from various roles and demographic backgrounds. This ensures that diverse perspectives on fairness, privacy, and usability are incorporated from the outset.

Guideline 3 (Mandatory Ethical Impact Assessment – EIA) is drawn from UNESCO's methodology, demanding that organizations must conduct a formal EIA for any proposed AI training system (UNESCO, 2025). This structured process proactively identifies and evaluates potential harms related to bias, discrimination,

surveillance, and psychological impact (e.g., stress from hyper-realistic simulations). The EIA's findings must be documented and used to inform a go/no-go decision or to mandate specific design modifications. Data protection cannot be an add-on, which is why **Guideline 4** (Privacy and Security by Design) requires principles of Privacy by Design to be embedded in the system's core architecture. This includes strict adherence to data minimization (collecting only the data absolutely necessary for the defined pedagogical purpose), purpose limitation (prohibiting the repurposing of data without explicit consent), and implementing robust security protocols like end-to-end encryption from the very beginning.

5.2 Stage 2: Responsible development and validation

During this stage, the focus shifts to the technical implementation of the ethical principles defined during the design phase. For this, organizations must establish rigorous data governance protocols (**Guideline 5**: Data Governance and Bias Mitigation). Training datasets must be carefully curated and audited for historical biases and to ensure they are representative of the entire employee population. Proactive bias mitigation techniques, such as re-weighting data or using fairness-aware machine learning algorithms, should be implemented during model training to actively counteract potential sources of discrimination. In the model selection process, Guideline 6 (Prioritizing Explainability) demands that preference must be given to inherently interpretable models (so-called 'white-box' models) where possible. If the complexity of the task necessitates a black-box model (e.g., a deep neural network), its use is contingent upon the integration of recognized XAI techniques (such as LIME or SHAP) that can provide human-understandable justifications for its outputs [5].

According to **Guideline 7** (Rigorous and Holistic Testing) the validation process must extend beyond measuring simple predictive accuracy. It must include dedicated testing for fairness, evaluating the model's performance across different demographic subgroups to detect any disparate impact. Furthermore, the system must undergo adversarial testing to assess its resilience against manipulation and ensure its security and reliability in a real-world environment.

5.3 Stage 3: Transparent and accountable deployment

Once the system is operational, the focus shifts to ensuring its use is transparent, accountable, and subject to human control. **Guideline 8** (Radical Transparency and Informed Consent) establishes that employees must be clearly and unambiguously informed when they are interacting with an AI system for training and assessment. The terms of use must explain in plain language what data is being collected, how it will be used for pedagogical purposes, and what its limitations are. For particularly invasive methods (e.g., sentiment analysis, deepfake simulations), specific, opt-in consent should be required.

Guideline 9 (Meaningful Human Oversight): AI systems should be deployed as tools to *augment* human educators and security analysts, not to *replace* them. A clear HITL protocol must be established for all high-stakes decisions. For example, an AI-generated risk score should never be the sole basis for any administrative or disciplinary action; it can serve as an input, but the final judgment must be made by a qualified human after an independent review of the evidence.

Guideline 10 (Mechanisms for Redress and Contestability): Organizations must establish clear, accessible, and well-publicized channels for employees to question, appeal, or request a human review of an AI-driven assessment. Employees have a right to understand the basis of a decision that affects them and to challenge its validity if they believe it is inaccurate or unfair.

Guideline 11 (Continuous Monitoring and Auditing): Deployment is not the end of the process. Organizations must implement ongoing monitoring to detect model drift (where performance degrades over time as real-world data changes) and to periodically audit the system for the emergence of new biases or unintended consequences. These audits should be conducted by a multidisciplinary team, and their results should be made available to the governance body.

5.4 Stage 4: Responsible decommissioning

The final stage of the lifecycle is often overlooked but is critical for data protection, which requires **Guideline 12** (Secure Data Disposition). When an AI training system is retired or replaced, there must be a formal process for the secure and permanent deletion of all associated employee data. This process must comply with the organization's data retention policies and all applicable privacy regulations, ensuring that sensitive employee information does not persist indefinitely.

6 APPLYING THE FRAMEWORK: ILLUSTRATIVE EXAMPLE SCENARIOS

To demonstrate its practical utility, this section applies the PERSP framework to one challenging but realistic scenario where organizations are deploying advanced AI in security training. This case illustrates how the framework guides concrete decision-making and mitigates tangible harms.

6.1 Scenario 1: AI-powered phishing simulation with deepfake technology

A multinational corporation, concerned about sophisticated social engineering attacks, plans to deploy a new mobile security training application. To achieve maximum realism, the Chief Information Security Officer (CISO) proposes using a feature that leverages generative AI to create hyper-personalized phishing simulations. This includes using AI-generated text that mimics the communication style of the employee's direct manager and, for high-risk employees, using deepfake audio technology to leave voicemails that sound exactly like a senior executive requesting an urgent action. The goal is to prepare employees for the next generation of AI-powered threats.

Stage 1 (Design): The project would immediately trigger several PERSP guidelines. The Ethical Impact Assessment (Guideline 3) would be the first critical step. This assessment would identify the significant potential for psychological harm, such as extreme stress, anxiety, or a breakdown of trust in leadership, caused by receiving a highly convincing fake message from a trusted superior. The Proportionality & Necessity test (Guideline 1) would force the CISO to justify whether the potential for such severe emotional distress is proportional to the incremental pedagogical benefit over a highly realistic text-based simulation. During the Multi-Stakeholder

Consultation (Guideline 2), HR and employee representatives would likely voice strong objections to the use of deepfakes, citing the potential for it to create a toxic and fearful work environment.

Stage 2 (Development): If the project were to proceed in a modified form (e.g., without deepfakes), Data Governance (Guideline 5) would be crucial. The framework would mandate that the data used for personalization be strictly limited to non-sensitive professional context (e.g., job title, department) and explicitly prohibit the use of personal or demographic data that could introduce bias.

Stage 3 (Deployment): The principle of Radical Transparency and Informed Consent (Guideline 8) would be non-negotiable. The framework would require the company to go beyond a simple notice and obtain explicit, opt-in consent from employees before enrolling them in such an advanced and potentially stressful simulation program. A clear and easily accessible Mechanism for Redress (Guideline 10) would need to be established, allowing employees to immediately report a simulation they find unduly manipulative or harmful and to opt-out of future advanced simulations without penalty.

It is plausible that the PERSP framework would guide the organization toward a more responsible outcome. It would likely lead to the abandonment of the deepfake voice component due to its disproportionate potential for harm identified in the EIA. The project would proceed with AI-generated text-based simulations, but with strict data governance, full transparency, and robust opt-out and redress mechanisms, balancing the need for effective training with the ethical obligation to protect employee well-being.

7 DISCUSSION: CHALLENGES TO IMPLEMENTATION AND FUTURE RESEARCH

While the PERSP framework provides a comprehensive roadmap for the responsible use of AI in security pedagogy, its successful implementation is not without challenges. These hurdles are not primarily technical but are deeply rooted in organizational culture, resource allocation, and the current state of AI literacy. Addressing these challenges is essential for translating ethical principles into sustained practice.

As the study of Sevilla-Bernardo, Cervera, and Robles (2025) has shown, even if the advantages regarding AI are recognized, there is a certain hesitancy among educators to implement it. For principles such as transparency and accountability to be meaningful, all stakeholders, beginning with the developers building the models, to the HR managers creating policies, to the employees interacting with the systems, must possess a foundational understanding of how AI works, its capabilities, and its limitations [24]. Without this shared knowledge base, governance becomes a top-down compliance exercise rather than a shared cultural commitment. Overcoming this requires sustained investment in tailored training and education for all levels of the organization. Implementing a rigorous ethical framework such as PERSP may be perceived by some as a bureaucratic impediment that slows down innovation, increases costs, and hinders agility. A culture that prioritizes rapid deployment over deliberate ethical reflection will naturally resist processes such as mandatory Ethical Impact Assessments or multi-stakeholder consultations. Overcoming this resistance requires strong, visible commitment from senior leadership, who must champion the long-term strategic value of building trust and mitigating risk over short-term efficiency gains.

The technical requirements for responsible AI are non-trivial. Implementing true Explainable AI, conducting thorough bias audits, and deploying advanced privacy-preserving machine learning techniques require specialized expertise and significant computational resources [5]. Many organizations, particularly small and medium-sized enterprises, may lack the in-house talent or financial resources to implement these measures comprehensively, creating a potential gap between ethical aspirations and technical reality. The legal and regulatory landscape governing AI is still in its early stages of development and varies significantly across jurisdictions. This ambiguity can create uncertainty for organizations, making them hesitant to invest in robust governance frameworks for fear that future regulations will render their efforts obsolete or non-compliant. While frameworks such as PERSP are designed to be proactive and align with foundational principles likely to inform future laws, the lack of clear legal safe harbors remains a challenge.

These challenges also point to critical areas for future academic and applied research. Advancing the field of responsible AI in security pedagogy will require concerted effort in the following areas: There is a pressing need for research into the development of standardized, quantifiable metrics for auditing AI training systems. Future work should focus on creating validated scoring rubrics for fairness, transparency, and accountability that would allow for objective, comparable assessments of different platforms. Since the use of AI for continuous monitoring and hyper-realistic simulation is a new phenomenon, there is a critical need for longitudinal psychological studies that examine the long-term effects of these systems on employee morale, stress levels, creativity, and trust in management. Such research would provide the empirical data needed to refine the 'Proportionality and Do No Harm' principle [6], [7].

As training simulations become more sophisticated, they begin to resemble the very attacks they are designed to prevent. This raises novel pedagogical and ethical questions. Future research should explore the concept of adversarial pedagogy, investigating the ethical boundaries of using deceptive, manipulative, and potentially psychologically distressing techniques (such as deepfakes) for educational purposes.

8 CONCLUSION

The integration of AI into corporate security pedagogy marks a pivotal moment, offering unprecedented opportunities to create a more adaptive, engaging, and effective defense against ever-evolving cyber threats. The capabilities of AI to personalize learning, simulate realistic attacks, and analyze behavior hold the promise of transforming security awareness from a passive compliance exercise into a dynamic and integral component of organizational resilience. However, this paper has argued that this technological promise is inextricably linked to a set of profound ethical risks. The very power of these systems introduces significant challenges to employee privacy, algorithmic fairness, accountability, and transparency.

This study has systematically dissected these dilemmas, demonstrating how the convergence of training and surveillance, coupled with the 'black box' nature of many AI models, creates a pressing need for proactive ethical governance. In response, this paper has proposed the PERSP. By anchoring itself in the universal, human-rights-based principles of the UNESCO Recommendation on the Ethics of AI and synthesizing them with the operational rigor of frameworks from NIST, EDUCAUSE, and industry leaders, PERSP offers a comprehensive and actionable solution. Its lifecycle-based structure, with guidelines for each stage from design to

decommissioning, provides a practical roadmap for organizations to navigate this complex terrain.

The illustrative scenario demonstrates that the framework is a practical decision-making tool that can guide organizations toward balancing security imperatives with their ethical obligations. Ultimately, the proactive integration of ethics is not a constraint on innovation but a prerequisite for its sustainable success. By adopting a robust framework like PERSP, organizations can build security training platforms that are not only technologically advanced but also fundamentally trustworthy. This fosters a security culture that empowers employees and respects their rights, which is the true foundation of a resilient and ethical organization in the digital age.

9 REFERENCES

[1] N. Onyebuchi, O. Ayeni, N. Hamad, B. Osawaru, and O. Adewusi, "AI in education: A review of personalized learning and educational technology," *GSC Advanced Research and Reviews*, vol. 18, no. 2, pp. 261–271, 2024. https://doi.org/10.30574/gscarr.2024.18.2.0062

[2] N. S. Al-Musib, F. M. Al-Serhani, M. Humayun, and N. Z. Jhanjhi, "Business email compromise (BEC) attacks," *Materials Today: Proceedings*, vol. 81, pp. 497–503, 2023. https://doi.org/10.1016/j.matpr.2021.03.647

[3] Chapman University, "Bias in AI," chapman.edu, 2025. [Online]. Available: https://www.chapman.edu/ai/bias-in-ai.aspx [Accessed: Nov. 15, 2024].

[4] Z. Zhang, H. Al Hamadi, E. Damiani, C. Y. Yeun, and F. Taher, "Explainable artificial intelligence applications in cyber security: State-of-the-art in research," *IEEE Access*, vol. 10, pp. 93104–93139, 2022. https://doi.org/10.1109/ACCESS.2022.3204051

[5] M. Attoresi, V. Bernardo, X. Lareo, and L. Velasco, "TechDispatch – Explainable artificial intelligence," Publications Office of the European Union, Luxembourg, Tech. Rep. QT-AD-23-002-EN-N, #2/2023, Europa.eu, 2023. [Online]. Available: https://data.europa.eu/doi/10.2804/802043

[6] N. Allahrakha, "UNESCO's AI ethics principles: Challenges and opportunities," *International Journal of Law and Policy*, vol. 2, no. 9, pp. 24–36, 2024. https://doi.org/10.59022/ijlp.225

[7] UNESCO, "Recommendation on the ethics of artificial intelligence," 2025. [Online]. Available: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics [Accessed: Nov. 15, 2024].

[8] G. L. Vousinas, "Exploring the impact of AI on education: Implications and future trends," *Journal for Future Society and Education*, vol. 2, no. 2, pp. 4–12, 2025. https://doi.org/10.3991/jfse.v2i2.53691

[9] H. Luan and C.-C. Tsai, "A review of using machine learning approaches for precision education," *Educational Technology & Society*, vol. 24, no. 1, pp. 250–266, 2021.

[10] R. Röttinger, "Steigende Bedrohungen durch Überfälle und Automatensprengungen," *DSD – Der Sicherheitsdienst*, no. 4, pp. 21–23, 2024. [German].

[11] C. Conati, O. Barral, V. Putnam, and L. Rieger, "Toward personalized XAI: A case study in intelligent tutoring systems," *Artificial Intelligence*, vol. 298, p. 103503, 2021. https://doi.org/10.1016/j.artint.2021.103503

[12] C. Zhong and A. Yayla, "Cognitive impacts of explainable AI in cybersecurity incident response: Challenges and propositions," *Inf Syst Front*, 2025. https://doi.org/10.1007/s10796-025-10609-y

[13] T. Naef, *Data Protection without Data Protectionism: The Right to Protection of Personal Data and Data Transfers in EU Law and International Trade Law, ser. European Yearbook of International Economic Law*, Cham, Switzerland: Springer, 2023. https://doi.org/10.1007/978-3-031-19893-9

[14] D. P. Bhave, L. Teo, and R. S. Dalal, "Privacy at work: A review and a research agenda for a contested terrain," *Journal of Management*, vol. 46, no. 1, pp. 127–164, 2020. https://doi.org/10.1177/0149206319878254

[15] X. Chang, "Gender bias in hiring: An analysis of the impact of amazon's recruiting algorithm," *Advances in Economics, Management and Political Sciences*, vol. 23, no. 1, pp. 134–140, 2023. https://doi.org/10.54254/2754-1169/23/20230367

[16] F. Santoni de Sio and G. Mecacci, "Four responsibility gaps with artificial intelligence: Why they matter and how to address them," *Philosophy & Technology*, vol. 34, no. 4, pp. 1057–1084, 2021. https://doi.org/10.1007/s13347-021-00450-x

[17] M. K. Pasupuleti, "Human-in-the-Loop AI: Enhancing transparency and accountability," *International Journal of Academic and Industrial Research Innovations (IJAIRI)*, vol. 5, no. 5, pp. 574–585, 2025. https://doi.org/10.62311/nesx/rphcr18

[18] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, "Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges," *Philosophy & Technology*, vol. 31, no. 3, pp. 611–627, 2018. https://doi.org/10.1007/s13347-017-0279-x

[19] U. Peters, "Explainable AI lacks regulative reasons: Why AI and human decision-making are not equally opaque," *AI and Ethics*, vol. 3, no. 3, pp. 963–974, 2023. https://doi.org/10.1007/s43681-022-00217-w

[20] M. Loi and M. Christen, "Ethical frameworks for cybersecurity," in *The Ethics of Cybersecurity*, M. Christen, B. Gordijn, and M. Loi, Eds., Cham, Switzerland: Springer, 2020, pp. 73–95. https://doi.org/10.1007/978-3-030-29053-5_4

[21] N. K. Corrêa *et al.*, "Worldwide AI ethics: A review of 200 guidelines and recommendations for AI governance," *Patterns*, vol. 4, no. 10, p. 100857, 2023. https://doi.org/10.1016/j.patter.2023.100857

[22] A. Filani, "How nist framework, IoT and cybersecurity work together," *researchgate.net*, 2025. [Online]. Available: ResearchGate. https://doi.org/10.13140/RG.2.2.23062.66885

[23] Google, "Secure AI Framework von Google – Google Sicherheitscenter," Google.com, 2025. [Online]. Available: https://cloud.google.com/use-cases/secure-ai-framework [Accessed: Nov. 15, 2024]. [German].

[24] J. Sevilla-Bernardo, L. Cervera, and J. Martin-Robles, "Impact and opportunities of generative artificial intelligence in education: A study of academic perceptions," *International Journal of Emerging Technologies in Learning (iJET)*, vol. 20, no. 3, pp. 55–71, 2025. https://doi.org/10.3991/ijet.v20i03.55809

10 AUTHORS

Luca Güttner is an assistant researcher at the department of philosophy at the University of Münster. His main areas of research are philosophy of antiquity, ethics, and political philosophy (E-mail: l_guet02@uni-muenster.de).

Prof. Dr. Raphael Röttinger teaches, conducts research, and provides consulting as a professor and expert in the fields of terrorism and amok incidents, as well as leadership studies for special operational scenarios (TAG/MANV/LebEL). He is also the CEO of an internationally operating corporate group.