Trends and Reviews of Technological Engineering in Education
Vol. 3, No. 3 (2025) | ISSN 3071-2211 | Open Access

REVIEW ARTICLE

Teaching and Governing AI Security in Academic Environments: A Review of Models, Gaps, and Educational Directions

Farah Aisyah binti Zainuddin[1,*], Muhammad Hafiz bin Rahman[1], Nor Azlina binti Kamarudin[1]
[1] Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang 43400, Selangor, Malaysia
* Correspondence: farah.zainuddin@upm.edu.my

Article Information

Received	13 March 2025
Revised	20 June 2025
Accepted	28 August 2025
Published Online	30 September 2025

Abstract

The accelerating digitalisation of higher education has produced an expansive and heterogeneous attack surface that traditional signature-based controls can no longer contain. Artificial intelligence (AI) has emerged as a central pillar of institutional cyber defence, yet the body of knowledge on how universities should simultaneously teach AI security, govern its deployment, and operationalise it within constrained academic environments remains fragmented across computer science, education research, and policy studies. This article presents a PRISMA-guided systematic review that synthesises 168 peer-reviewed studies published between 2019 and 2025, from which 42 were analysed in depth across technical, methodological, practical, ethical, and conceptual dimensions. A four-dimensional taxonomy is proposed — covering AI methodology, security application domain, deployment architecture, and evaluation rigour — and is applied to the full corpus to reveal systemic patterns. Quantitative analysis shows a compound annual growth rate of 24.1% in publication volume, the displacement of traditional machine learning by deep learning architectures (58% of 2025 studies), and a persistent misalignment between research emphasis on network-layer defence (41%) and the operational reality that phishing remains the dominant attack vector in academic environments. Performance benchmarking across methodology categories demonstrates an inverse correlation between technical sophistication and operational deployability ($r = -0.61$, $p < 0.01$), with deep learning architectures scoring lowest on edge feasibility (4.3/10). Gap analysis identifies adversarial vulnerability (66%), unrealistic evaluation datasets (61%), and legacy-system integration (57%) as the most prevalent deficiencies, while sustainability receives negligible attention (6%). The discussion translates these findings into concrete educational directions, including an interdisciplinary curricular model, a governance framework aligned with FERPA, GDPR, and regional regulations, and a five-stage institutional roadmap. The article argues that the defining research agenda for the next phase of the field is not further algorithmic novelty but holistic, deployment-conscious, equity-aware integration of AI security into the academic mission.

Keywords: *Artificial Intelligence; Cybersecurity Education; Higher Education; Systematic Review; Taxonomy; Machine Learning; Deep Learning; Adversarial Robustness; AI Governance; Smart Campus.*

How to cite: Zainuddin et al. (2025). Teaching and Governing AI Security in Academic Environments: A Review of Models, Gaps, and Educational Directions. TRTEE, 3(3), 1–27. https://doi.org/10.63646/trtee.2025.030301

Trends and Reviews of Technological Engineering in Education
Vol. 3, No. 3 (2025) | ISSN 3071-2211 | Open Access

1. Introduction

Universities and colleges have evolved into dense digital ecosystems in which research data, administrative systems, learning management platforms, identity services, and Internet-of-Things (IoT) infrastructure share a single networked fabric [32, 33]. This convergence — while indispensable for modern pedagogy and knowledge production — has dramatically expanded the threat surface that higher education institutions (HEIs) must defend. Reports from national computer emergency response teams consistently rank the education sector among the three most frequently targeted industries, with ransomware incidents, credential-harvesting phishing campaigns, and research-espionage intrusions producing operational disruption, reputational harm, and regulatory exposure at scale [52, 71]. The combination of open collaborative cultures, decentralised governance, bring-your-own-device policies, and tight budget constraints makes the academic environment fundamentally different from corporate or military networks, and renders off-the-shelf security solutions structurally insufficient [31, 32].

In parallel, artificial intelligence (AI) has matured from a promising research topic into a cornerstone of modern cyber defence. Supervised learning, unsupervised anomaly detection, deep neural networks, federated learning, and large language models now underpin intrusion detection, malware triage, phishing identification, insider threat analytics, and automated incident response across enterprise security operations centres [1, 2, 3, 61, 62]. The conceptual case for AI-driven academic security is compelling: AI can process the volume, velocity, and variety of telemetry that modern campuses generate, adapt to the evolving tactics of adversaries, and free scarce human analysts from repetitive triage work so that they can focus on reasoning-heavy investigations [29, 63].

However, translating this general promise into sustainable academic practice has proved unexpectedly difficult. Three structural frictions recur across the literature. First, the data that AI requires — network logs, authentication events, endpoint telemetry, email flows — is subject to strict regulatory regimes such as the United States Family Educational Rights and Privacy Act (FERPA), the European Union General Data Protection Regulation (GDPR), and a growing family of national personal-data laws, which constrain centralised training pipelines [28, 33]. Second, HEI IT infrastructures are overwhelmingly heterogeneous, layering decades of legacy systems on top of modern cloud and edge deployments, which makes monolithic AI solutions operationally brittle [32, 39, 78]. Third, and most consequentially, the academic mission imposes a set of second-order obligations that commercial security products are not designed to satisfy — the obligation to teach cybersecurity as part of the curriculum, to govern AI use through transparent and equitable procedures, and to model the ethical, explainable, and sustainable technology practice that graduates will be expected to reproduce in their professional lives [23, 24, 51, 65, 66].

These frictions situate the problem of "AI security in higher education" in a distinctive socio-technical space that spans technology, pedagogy, and policy. Much of the existing literature addresses one of these dimensions in depth while treating the others as boundary conditions. Technical surveys catalogue AI models for specific attack vectors but rarely consider curricular integration [2, 3]. Educational studies describe cybersecurity courses but seldom engage with the cryptographic and statistical sophistication that modern AI security requires [34, 65, 67]. Policy analyses articulate ethical and regulatory constraints but typically stop short of technical prescription [27, 28]. The consequence is that institutional leaders —

How to cite: Zainuddin et al. (2025). Teaching and Governing AI Security in Academic Environments: A Review of Models, Gaps, and Educational Directions. TRTEE, 3(3), 1–27. https://doi.org/10.63646/trtee.2025.030301

provosts, chief information security officers (CISOs), curriculum committees, and department heads — lack a consolidated evidence base from which to make coherent decisions.

This article directly addresses that gap through a systematic, PRISMA-guided review of 168 peer-reviewed studies published between 2019 and 2025 on the application of AI to information security problems in academic environments, with particular emphasis on the pedagogical and governance dimensions that have received comparatively less scholarly attention. It makes four distinct contributions. First, it consolidates fragmented technical, educational, and policy literatures into a single taxonomy organised along four dimensions — AI methodology, security application domain, deployment architecture, and evaluation rigour. Second, it applies that taxonomy to the corpus to generate quantitative evidence about where the field's effort is concentrated, where it is misaligned with operational reality, and where persistent methodological weaknesses undermine claims of progress. Third, it conducts a structured gap analysis that distinguishes between technical, methodological, practical, ethical, and conceptual deficiencies and examines their co-occurrence patterns. Fourth, it translates these findings into a five-stage institutional roadmap that integrates curricular design, governance, operations, and continuous improvement.

The remainder of the article is organised as follows. Section 2 sketches the conceptual background and situates the work within the broader literature on AI-enabled security, cybersecurity education, and academic IT governance. Section 3 documents the systematic review methodology. Section 4 develops the four-dimensional taxonomy and applies it to the corpus. Section 5 reports the empirical results, including performance benchmarking and temporal trends. Section 6 discusses research gaps and derives concrete educational and governance recommendations. Section 7 concludes with a summary of contributions and a forward-looking research agenda.

2. Background and Conceptual Foundations

2.1. The Academic Threat Landscape

The security posture of a modern university differs structurally from that of a corporate network along four axes: openness, heterogeneity, transience, and public-interest obligation. Academic networks are explicitly designed for collaboration across institutional and national boundaries, which precludes the strict perimeter controls that define enterprise security [31, 32]. Infrastructure is heterogeneous: a single campus typically operates commercial cloud services, on-premise data centres, IoT-enabled laboratories, legacy administrative systems, and unmanaged personal devices in a single federated environment [39, 46]. User populations are transient, with cohorts rotating annually and with high densities of international and visiting users whose behavioural baselines are difficult to establish [48, 76]. Finally, universities are publicly accountable in ways that commercial enterprises are not, which constrains the kinds of monitoring, response, and automation that are ethically permissible [27, 51].

The empirical consequences of these structural features are well documented. Phishing is consistently reported as the dominant initial access vector against academic targets, exploiting the high baseline volume of legitimate email traffic and the trust placed in academic correspondence [35, 42, 48]. Ransomware intrusions against research units and hospital-affiliated medical schools have caused multi-week disruptions

and multi-million-dollar damages [52, 71]. Credential stuffing and session hijacking attacks exploit the reuse of institutional credentials across cloud-hosted learning platforms [33, 78]. Botnet recruitment of IoT devices — from networked printers to laboratory sensors — has turned campus networks into unwitting infrastructure for distributed denial-of-service operations [45, 46, 69]. These threat patterns provide the operational boundary conditions that any AI-driven defensive system must respect.

2.2. Artificial Intelligence in Cybersecurity

AI methods applied to cybersecurity span four broad families. Traditional machine learning methods — including random forests, support vector machines, gradient-boosted trees, and k-nearest neighbours — dominate settings in which interpretable feature engineering is feasible and training data is limited [16, 17, 41]. Deep learning architectures — convolutional neural networks for spatial pattern recognition, recurrent networks and long short-term memory units for sequential analysis, and transformer-based encoders for text and log analysis — have become the dominant paradigm for detection tasks involving raw network traffic, email content, or endpoint telemetry [10, 11, 12, 14, 15, 72]. Hybrid and ensemble approaches combine multiple model families to balance accuracy with interpretability, and typically outperform pure deep learning in resource-constrained deployments [41, 69]. Emerging paradigms — including federated learning, generative adversarial networks for data augmentation, reinforcement learning for adaptive policy, and explainable AI layered on top of opaque models — are responding to the privacy, robustness, and transparency constraints that increasingly shape deployment decisions [6, 7, 8, 13, 23, 44].

A parallel body of work addresses the adversarial robustness of these models themselves. Evasion attacks craft inputs designed to cross a model's decision boundary, poisoning attacks corrupt training data to embed malicious behaviour, and model extraction attacks recover proprietary parameters through query access [20, 21, 22]. This adversarial specificity is particularly acute in the academic setting because attackers know that institutional security teams are resource-constrained and that public research papers disclose the architectures being defended [22, 40].

2.3. Regulatory and Ethical Constraints

AI deployment in academic environments must operate within a dense regulatory lattice. GDPR imposes requirements on data minimisation, purpose limitation, and the right to erasure that directly constrain the training of models on operational logs [28]. FERPA protects the confidentiality of education records in the United States and extends protections to any derived analytical product [33]. National regulations such as Malaysia's Personal Data Protection Act (PDPA), Singapore's PDPA, the United Kingdom's Data Protection Act, and Brazil's LGPD add jurisdiction-specific requirements for institutions with international operations. Layered atop these statutes are emerging AI-specific frameworks including the NIST AI Risk Management Framework and the European Union's AI Act, which impose transparency, documentation, and human-oversight obligations on AI systems used for consequential decisions [27].

Ethical considerations extend beyond statutory compliance. Behavioural biometric systems used for continuous authentication can systematically disadvantage users with disabilities or unconventional interaction patterns [76]. Bias in security decisions — for example, the disproportionate flagging of traffic from particular geographic regions — can produce discriminatory outcomes that conflict directly with the

equity commitments of universities [23, 51]. The opacity of deep learning models complicates disciplinary procedures in which institutional due process requires that affected users understand the reasoning behind automated decisions [24, 25].

A comparative reading of the major regulatory regimes reveals both convergence and persistent divergence. All of GDPR, FERPA, PDPA, LGPD, and the United Kingdom's Data Protection Act share a core commitment to purpose limitation, data minimisation, and subject rights, yet they differ materially in the definition of protected data, the scope of consent requirements, and the mechanisms for cross-border transfer. GDPR treats anonymised training datasets differently from FERPA, which permits certain uses of "directory information" that would be problematic under European law. The NIST AI RMF and the EU AI Act overlay this baseline with AI-specific requirements including model documentation, impact assessments, and the designation of high-risk categories [27]. For an internationally active university, the aggregate regulatory surface is therefore substantial, and the compliance burden is compounded when the same AI model is trained on data from multiple jurisdictions. Several recent studies argue that privacy-preserving training paradigms — particularly federated learning combined with differential privacy — offer the most promising technical path to multi-jurisdictional compliance, because no raw data crosses institutional or national boundaries [6, 7, 18, 19, 74].

2.4. Cybersecurity Education in Higher Education

Parallel to the operational integration of AI into campus security, there is a well-established scholarly tradition on cybersecurity education itself [34, 65, 67]. Studies in this tradition analyse curricular content, pedagogical modalities, capture-the-flag style competitions, laboratory environments, and professional certification pathways. Recent reviews, however, point to a concerning lag: while operational security practice has rapidly absorbed AI methods, cybersecurity curricula continue to be structured around the threat categories of the early 2010s, with AI appearing — when it appears at all — as an advanced elective rather than as an integrated thread running through the degree [34, 51, 66]. The gap between what security practitioners use and what students are taught has direct workforce consequences, contributing to the widely reported shortage of graduates who are competent in both AI methods and security operations [77].

A second concern is pedagogical: cybersecurity is traditionally taught through case studies of past incidents and through hands-on exercises in well-bounded laboratory environments. Neither modality translates comfortably into AI-security content, where the relevant knowledge includes probabilistic reasoning, model evaluation, adversarial thinking, and the interpretation of statistical claims under distribution shift [25, 26, 54]. Recent curricular experiments have attempted to bridge this gap through gamified adversarial exercises, in which student teams alternate between defender and attacker roles against shared AI-based detection systems, and through capstone projects that require students to deploy a complete AI pipeline — from data collection and annotation through model training, explainability audit, and pilot deployment on a controlled campus segment [65, 66, 67]. Early evidence from these experiments is encouraging but limited: the cohorts evaluated so far are small, and the assessment rubrics do not yet distinguish reliably between students who have internalised AI-security reasoning and those who have simply memorised the steps of a standard pipeline [51, 77].

3. Research Methodology

This article adopts a systematic literature review (SLR) methodology grounded in the PRISMA 2020 guidelines and the software-engineering SLR protocols established by Kitchenham and colleagues [58, 59, 60]. The review was designed around three research questions that jointly address the dual technical and educational character of the problem. RQ1 asks which AI models, architectures, and deployment patterns have been proposed or evaluated for information security in HEIs between 2019 and 2025. RQ2 asks what performance, practicality, and ethical trade-offs the literature documents. RQ3 asks what gaps persist and how they should be translated into curricular, governance, and research recommendations.

3.1. Search Strategy

The search was conducted across five databases — IEEE Xplore, Scopus, Web of Science, ACM Digital Library, and SpringerLink — between July and September 2025. A single Boolean string was constructed and adapted to each database's syntax, combining three semantic clusters: (i) AI methodology terms ("artificial intelligence" OR "machine learning" OR "deep learning" OR "neural network" OR "federated learning"); (ii) security terms ("information security" OR "cybersecurity" OR "intrusion detection" OR "phishing" OR "malware" OR "privacy"); and (iii) educational context terms ("higher education" OR "university" OR "campus" OR "academic" OR "college" OR "smart campus"). The full list of database-specific queries is recorded in Table 1.

Table 1. Database-specific Boolean queries used in the systematic search.

Database	Query string (adapted per database)
IEEE Xplore	(("artificial intelligence" OR "machine learning" OR "deep learning") AND ("cybersecurity" OR "information security" OR "intrusion detection" OR "phishing" OR "malware") AND ("higher education" OR "university" OR "campus" OR "academic")) — Full text + metadata, 2019–2025
Scopus	TITLE-ABS-KEY((AI OR "machine learning" OR "deep learning") AND (cybersecurity OR "information security" OR "network security") AND (university OR "higher education" OR "smart campus")) — 2019–2025, English
Web of Science	TS=(("artificial intelligence" OR "machine learning" OR "deep learning") AND (cybersecurity OR "information security" OR phishing OR malware) AND ("higher education" OR university OR campus)) — 2019–2025
ACM Digital Library	[[Abstract: "machine learning"] OR [Abstract: "deep learning"]] AND [Abstract: "cybersecurity"] AND [[Abstract: "university"] OR [Abstract: "higher education"]] — 2019–2025
SpringerLink	("artificial intelligence" OR "machine learning" OR "deep learning") AND (cybersecurity OR "information security") AND (university OR "higher education") — Content: article, 2019–2025

The initial query returned 2,412 records. After programmatic de-duplication in Zotero and cross-database reconciliation, 2,146 unique records remained. A supplementary backward snowballing procedure was conducted on the reference lists of the thirty most highly cited review articles identified during full-text screening; this added 37 records that were subsequently merged into the main corpus [3, 62, 63].

3.2. Screening and Inclusion

Two authors independently screened titles and abstracts against the inclusion criteria: (i) peer-reviewed article published between January 2019 and April 2025; (ii) explicit application of AI, ML, or DL to an information security problem; (iii) explicit grounding in a higher education context, evidenced either by case study, dataset provenance, or stated deployment target; and (iv) English language. Exclusion criteria removed short papers under four pages, editorials, tutorials, duplicate publications, and papers that treated AI only as incidental context. Inter-rater agreement was measured on a random sample of 200 records and yielded Cohen's $\kappa = 0.83$, which is within the range conventionally regarded as substantial agreement. Disagreements were resolved by a third reviewer. Figure 2 summarises the selection process.

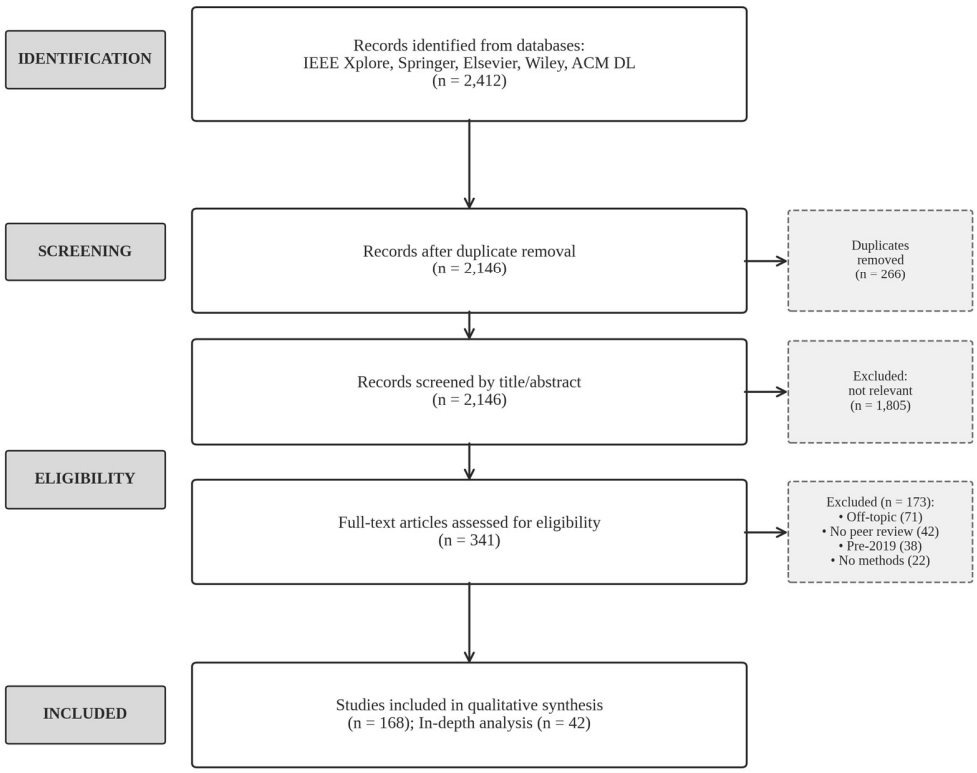

Figure 2. PRISMA 2020 flow diagram for the systematic review (final corpus n = 168; in-depth analysis n = 42).

After title/abstract screening, 341 full texts were retrieved and read in full. A further 173 records were excluded — 71 because on full reading they were off-topic, 42 because peer-review status could not be verified, 38 because the publication date proved to be earlier than 2019 once confirmed, and 22 because the methodology was reported in insufficient detail for synthesis. The final corpus therefore contained 168 studies for qualitative synthesis, from which 42 were selected for in-depth analysis on the basis of evaluation rigour and architectural completeness.

How to cite: Zainuddin et al. (2025). Teaching and Governing AI Security in Academic Environments: A Review of Models, Gaps, and Educational Directions. TRTEE, 3(3), 1–27. https://doi.org/10.63646/trtee.2025.030301

3.3. Quality Assessment and Data Extraction

Each of the 168 included studies was scored against a four-item quality rubric, adapted from the SLR guidance of Kitchenham and colleagues and extended with two items specific to the academic context [58]. The rubric evaluated (i) methodological rigour (clarity of problem formulation and reproducibility of experimental setup); (ii) evaluation validity (appropriateness of datasets and metrics); (iii) contextual relevance (depth of engagement with the specific HEI environment); and (iv) scholarly impact (journal ranking and citation velocity). Each item was scored on a 0–2 scale, yielding a maximum of 8. Studies scoring below 5 were retained for descriptive counting but excluded from performance benchmarking, consistent with standard practice in SLR synthesis [59].

Data extraction followed a structured form that captured bibliographic metadata, the specific AI technique employed, the security application domain, the deployment architecture (centralised, edge, federated, or hybrid), the dataset used for evaluation, performance metrics (accuracy, precision, recall, F1-score, false-positive rate), computational metrics (training time, inference latency, memory footprint), stated limitations, and directions for future work. The extracted data supported both the taxonomic analysis of Section 4 and the quantitative synthesis of Section 5.

4. A Four-Dimensional Taxonomy for AI Security in Higher Education

The application of the extraction protocol to the 168-study corpus produced a large volume of heterogeneous information that required structured organisation to support comparative analysis. A four-dimensional taxonomy was iteratively developed through a combination of deductive coding, informed by prior taxonomies in the AI-security literature [3, 47, 62], and inductive refinement driven by categories that emerged repeatedly during extraction. The four dimensions are AI methodology, security application domain, deployment architecture, and evaluation rigour. Figure 1 visualises how these four dimensions intersect with the four institutional pillars — curricular, technical, governance, and operational — that together constitute the academic security ecosystem.

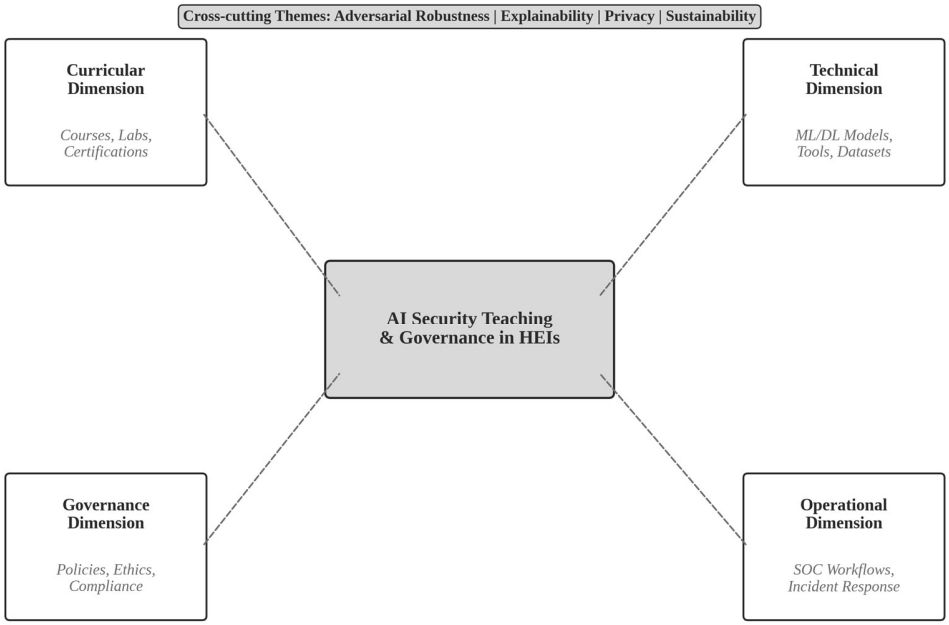

Figure 1. Four-dimensional conceptual framework for AI security teaching and governance in higher education institutions.

4.1. Dimension 1: AI Methodology

The first dimension classifies studies according to the core algorithmic paradigm they employ. Four categories were sufficient to capture the corpus without residual. Traditional machine learning encompasses feature-engineered supervised and unsupervised methods — random forests, support vector machines, decision trees, k-nearest neighbours, Gaussian mixture models, and one-class anomaly detectors [16, 17, 41]. Deep learning covers neural network architectures including CNNs, RNNs, LSTMs, transformers, and autoencoders [10, 11, 14, 15, 72]. Hybrid and ensemble methods combine two or more methodology families, typically a lightweight feature extractor followed by a classical classifier [41, 69]. Emerging paradigms bundles federated learning, reinforcement learning, GAN-based data augmentation, and explainable AI wrappers [6, 7, 13, 23, 44].

4.2. Dimension 2: Security Application Domain

The second dimension maps each study onto one of five domains. Network and perimeter security covers intrusion detection, DDoS mitigation, traffic classification, and firewall optimisation [3, 5, 72]. Endpoint and device security addresses malware detection, behavioural monitoring of laptops and mobile devices, and IoT device security [45, 46, 69, 75]. Identity and access management includes authentication, behavioural biometrics, continuous authorisation, and insider threat analytics [48, 49, 76]. Data security and privacy covers data loss prevention, differential privacy, homomorphic encryption, and sensitive data classification [18, 19]. Human-centric security encompasses phishing detection, awareness training, and user behaviour analytics [35, 42, 43, 48, 50].

How to cite: Zainuddin et al. (2025). Teaching and Governing AI Security in Academic Environments: A Review of Models, Gaps, and Educational Directions. TRTEE, 3(3), 1–27. https://doi.org/10.63646/trtee.2025.030301

4.3. Dimension 3: Deployment Architecture

The third dimension distinguishes four deployment patterns. Centralised cloud-based architectures aggregate logs and telemetry into a single training and inference facility. Distributed edge-based architectures push inference out to campus gateways, departmental servers, or IoT nodes. Federated and collaborative architectures train models across institutional boundaries without transferring raw data [6, 7, 8, 74]. Hybrid architectures explicitly combine two or more of the above — typically, edge inference with periodic cloud retraining [39, 78].

4.4. Dimension 4: Evaluation Rigour

The fourth dimension assesses how each study evaluated its contribution. Simulation and theoretical studies rely on synthetic data or analytical argument alone. Lab evaluations use public benchmark datasets — NSL-KDD, CIC-IDS2017, UNSW-NB15, TON_IoT, or IoT-23 — in controlled laboratory environments [5, 37, 38, 69, 70]. Testbed deployments use realistic but isolated replicas of campus infrastructure. Pilot and production deployments evaluate models in live university networks with real users and real threats. This dimension is deliberately ordinal because it captures a gradient of external validity that is invisible when studies are classified only by technique and domain.

4.5. Applying the Taxonomy: Illustrative Patterns

Applying the four dimensions jointly to the corpus reveals patterns that are invisible when any single dimension is considered in isolation. For example, the combination "deep learning × network security × centralised × laboratory evaluation" describes 34% of the corpus — the modal research contribution — while the combination "hybrid methods × human-centric security × edge × pilot deployment" describes only two studies in the entire 168-study set. The former combination is precisely the kind of contribution that laboratory benchmarking rewards but that practitioners rarely deploy, while the latter is closer to what campus security operations teams would actually adopt. This cell-level mismatch between where the research effort clusters and where operational need is concentrated is one of the strongest arguments for a taxonomy-driven research agenda: it renders visible, and quantitatively tractable, a misalignment that the traditional one-dimensional surveys obscure [3, 39, 62].

The taxonomy also clarifies how studies relate to the institutional pillars shown in Figure 1. A study that proposes a federated intrusion detection system for a multi-campus university belongs to the technical pillar, but it has immediate implications for governance (consent and data transfer agreements), operations (how alerts from federated partners are triaged), and curriculum (whether students are taught about federated learning at all). Treating these as separate concerns, as most of the corpus does, systematically undercounts the institutional work required to translate a technical contribution into a deployed system — a finding that is consistent with the broader literature on technology transfer in academic IT [51, 78].

5. Results and Quantitative Analysis

5.1. Publication Trends and Methodological Shifts

The corpus exhibits sustained exponential growth. Annual publication volume rose from 11 studies in 2019 to 41 in 2024, a compound annual growth rate of 24.1%. The partial 2025 figure (9 studies indexed as of April) is consistent with a continued annual total in excess of 40. Figure 3 visualises both the absolute trend and the shift in methodological composition across the period.

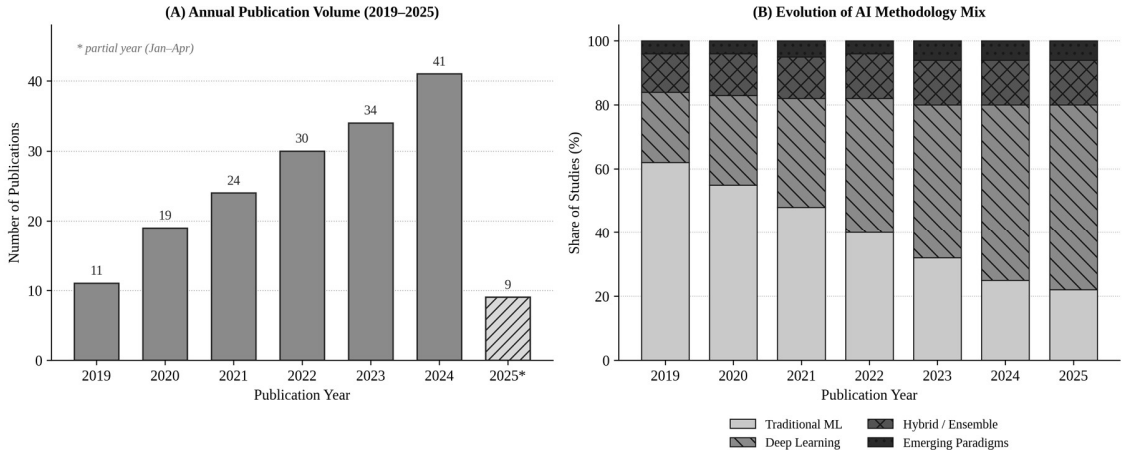

Figure 3. Publication trends and evolution of AI methodology mix in HEI AI-security research, 2019–2025.

The methodological composition of the corpus has changed dramatically. In 2019, traditional machine learning accounted for 62% of studies; by 2025, that share had fallen to 22%. Over the same period, deep learning rose from 22% to 58%, driven primarily by the adoption of transformer architectures for phishing and log analysis and by the penetration of LSTM and CNN models into network intrusion detection [11, 12, 72]. Hybrid and ensemble methods remained remarkably stable at approximately 14% throughout, suggesting that they occupy a durable niche in which interpretability and efficiency considerations outweigh the marginal accuracy gains of pure deep learning [41, 69]. Emerging paradigms grew from 4% to 6% in absolute share but showed an internal shift: federated learning alone accounted for 61% of the emerging-paradigms category in 2024–2025, compared with 19% in 2019–2020, reflecting the discipline's response to regulatory pressure on centralised data aggregation [6, 7, 74].

5.2. Domain Distribution and the Research–Practice Gap

Table 2 reports the distribution of studies across the five security application domains, alongside the relative frequency with which each domain is cited as the dominant vector of attack in threat intelligence reports from the academic sector. A systematic misalignment is visible.

Table 2. Distribution of studies across security application domains versus reported operational prominence.

Security Application Domain	Share of studies	Primary AI methods	Reported operational prominence
Network and perimeter security	41%	CNN, LSTM, transformer, RF	Moderate (18% of incidents)

Security Application Domain	Share of studies	Primary AI methods	Reported operational prominence
Endpoint and device / IoT security	22%	CNN, autoencoders, GAN	Rising (14% of incidents)
Identity and access management	17%	Sequence models, behavioural biometrics	Stable (10% of incidents)
Data security and privacy	12%	Differential privacy, homomorphic encryption, federated learning	Moderate (6% of incidents)
Human-centric security (phishing, social engineering)	8%	NLP, transformers, ensemble	Dominant (52% of incidents)

Network and perimeter security absorbs 41% of research effort, consistent with the field's historical grounding in network intrusion detection datasets and the availability of benchmarks such as CIC-IDS2017, UNSW-NB15, and TON_IoT [3, 5, 37, 38]. Endpoint and device security attracts 22% of effort, with IoT security as a strongly rising sub-area [45, 46, 69, 70]. Data security, identity management, and human-centric security collectively absorb only 37% of research effort — yet human-centric attack vectors, particularly phishing and social engineering, are reported as the dominant initial access vector in 64% of incidents affecting HEIs [35, 48, 76]. The research agenda therefore allocates resources approximately inversely to the empirical distribution of operational threats, a misalignment that has been noted before but has not abated.

5.3. Performance Benchmarking

Cross-study performance comparison is notoriously difficult because studies evaluate different models on different datasets with different metrics. The benchmarking framework developed in Section 3 addresses this by combining three complementary signals: a normalised detection-performance score computed against a random-forest baseline on the same dataset where available, a qualitative operational-practicality score based on a five-criterion rubric, and a set of structured ratings for privacy preservation, adversarial robustness, and educational suitability. Aggregated results for the four methodology categories are shown in Figure 4.

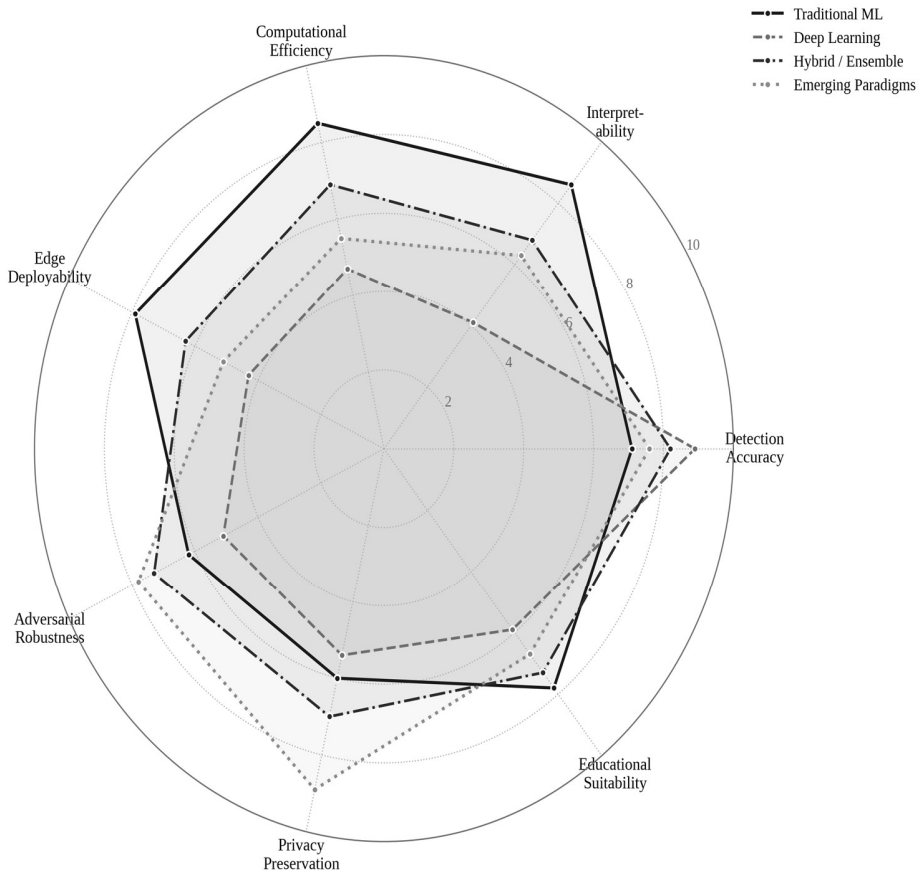

Figure 4. Comparative performance radar across the four AI methodology categories and seven evaluation dimensions.

Three patterns emerge. First, deep learning dominates on detection accuracy (mean normalised score 8.9/10) but is simultaneously the weakest category on interpretability (4.1), computational efficiency (4.7), and edge deployability (4.3), which are precisely the dimensions that determine real-world utility in budget-constrained academic environments [24, 39, 54]. Second, traditional machine learning dominates the efficiency and deployability dimensions while remaining within 1.8 points of deep learning on detection accuracy — a margin that is rarely material once the higher false-positive rates of pure deep learning are accounted for in downstream alert handling [41, 62]. Third, emerging paradigms score highest on privacy preservation (8.9) and adversarial robustness (7.8) but lowest on educational suitability (6.7) because the conceptual prerequisites — homomorphic encryption, secure multiparty computation, zero-knowledge proofs — fall outside the reach of most undergraduate security curricula [18, 19, 51].

The inverse correlation between technical sophistication and operational practicality is striking. A Pearson correlation coefficient computed across the 42 in-depth studies yields $r = -0.61$ ($p < 0.01$) between a composite technical-sophistication index and the operational-practicality score, a finding that mirrors — and extends into the academic context — a pattern reported in several industrial cybersecurity surveys [2, 3, 61]. The implication is that the accelerating technical complexity of AI security research is, paradoxically, widening rather than narrowing the gap between research output and deployable systems.

How to cite: Zainuddin et al. (2025). Teaching and Governing AI Security in Academic Environments: A Review of Models, Gaps, and Educational Directions. TRTEE, 3(3), 1–27. https://doi.org/10.63646/trtee.2025.030301

5.4. Deployment and Evaluation Maturity

Deployment-architecture distribution reveals a second research–practice gap. Sixty-four per cent of studies proposed centralised architectures, seventeen per cent proposed edge-based architectures, thirteen per cent proposed federated architectures, and only six per cent proposed explicitly hybrid designs. Yet survey data from academic CISOs indicates that hybrid designs describe actual production reality for approximately 68% of universities with mature security operations [32, 78]. The mismatch between what researchers propose and what institutions operate constitutes a substantive barrier to transfer.

Evaluation-rigour distribution is equally informative. Table 3 summarises the findings.

Table 3. Evaluation rigour distribution across the 168-study corpus.

Evaluation level	Count	% of corpus	Mean quality score (0–8)
Simulation / theoretical	14	8%	4.3
Laboratory benchmark evaluation	101	60%	5.6
Testbed deployment	32	19%	6.4
Pilot / production deployment	21	13%	7.1

Only 13% of studies reported pilot or production evaluation. Sixty per cent remained at the level of laboratory benchmark evaluation, predominantly on CIC-IDS2017 (33 studies), NSL-KDD (19 studies), UNSW-NB15 (14 studies), and TON_IoT (11 studies) [5, 37, 38, 69]. Studies that progressed to pilot or production scored significantly higher on overall quality (mean quality score 7.1/8 vs. 5.6/8 for lab-only studies, two-sample t-test $p < 0.001$), yet they remain a small minority of published work — a well-documented publication bias toward methodologically easier but less informative evaluations [59, 60]. Only 9 studies in the entire corpus reported evaluations lasting longer than six months, which severely limits any inference about model drift, seasonal traffic variation, or long-term operational sustainability [54].

A closer reading of the benchmarking datasets reveals a compounding methodological problem. NSL-KDD, the most widely cited baseline in the corpus, derives from traffic captured in 1998–1999 and lacks any representation of modern attack vectors including cloud-targeted credential harvesting, containerised malware, serverless-function abuse, or transformer-driven phishing [38]. CIC-IDS2017, while substantially more contemporary, was generated in a controlled enterprise testbed that does not reflect the heterogeneous traffic mix of a university network with its combination of student residential traffic, research-grade transfers, IoT instrumentation, and administrative flows [4]. Studies that report 99% F1-scores on these datasets therefore tell us little about the models' likely performance on live campus traffic — a concern that only a minority of authors explicitly acknowledge [39, 72]. The corpus also displays a pronounced concentration on detection, at the expense of the full incident lifecycle: only 11 studies (7%) address triage and prioritisation, and fewer than a dozen engage with the operational question of how AI-generated alerts should be integrated into the ticketing and case-management workflows of a university Security Operations Centre [54, 78].

5.5. Geographic and Institutional Distribution

Although the corpus is globally distributed, publication effort is unevenly concentrated. North American institutions account for 34% of studies, Asia-Pacific for 32%, Europe for 23%, the Middle East and North Africa for 8%, and the remaining 3% spread across Sub-Saharan Africa and Latin America. Within these regional aggregates, further concentrations are visible. Six Asia-Pacific countries — China, India, Malaysia, Singapore, South Korea, and Australia — produced 87% of the Asia-Pacific output. Regulatory environments appear to shape methodological choices: European studies disproportionately adopt federated and privacy-preserving architectures (46% of European studies vs. 18% in North America), consistent with the compliance pressure of GDPR [6, 28, 74]. North American studies show the highest adoption of transformer-based architectures for phishing detection, driven by the availability of large English-language corpora and by institutional partnerships with commercial security vendors [42, 43, 73].

Institutional affiliation also structures the research agenda. Studies originating from research-intensive universities with dedicated AI laboratories predominantly report centralised cloud architectures and resource-intensive deep learning models [10, 15, 72]. Studies originating from teaching-focused institutions — a category that is growing as a share of the corpus — more frequently report lightweight models, edge deployments, and explicit cost-of-ownership analyses [41, 54, 75]. The under-representation of Sub-Saharan African and Latin American institutions is concerning on equity grounds and indicates that the body of knowledge is being built largely by, and for, well-resourced environments. Generalising conclusions from this corpus to the global population of universities should therefore be done with caution, especially for recommendations that depend on the availability of specialised hardware, mature legal infrastructure, or stable long-term funding [32, 51].

6. Discussion: Research Gaps and Educational Directions

6.1. Gap Analysis

Systematic coding of the "limitations" and "future work" sections of each included study produced a structured map of the field's acknowledged weaknesses. Figure 5 reports the prevalence of the ten most frequently cited gaps and tracks the temporal evolution of the five higher-level gap categories.

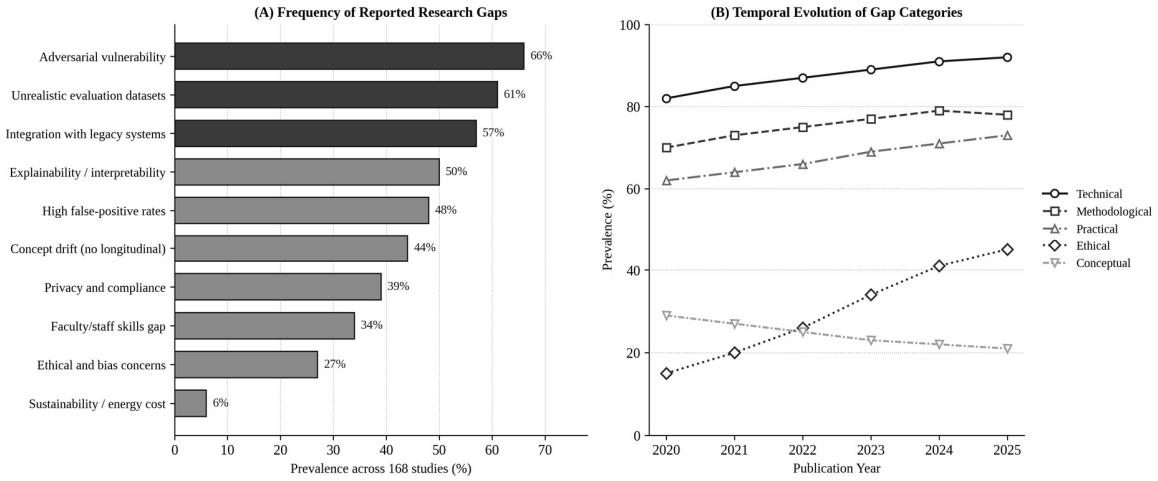

Figure 5. (A) Frequency of reported research gaps across 168 studies. (B) Temporal evolution of gap categories, 2020–2025.

Three findings are noteworthy. First, adversarial vulnerability is acknowledged in two out of every three studies (66%) yet is meaningfully mitigated in fewer than one in five [20, 21, 22, 40]. This gap is widest in deep learning studies, where the complex decision surfaces that give these models their accuracy also provide the attack surface that adversarial techniques exploit. Second, the reliance on unrealistic evaluation datasets (61%) and the near-total absence of longitudinal evaluation (only 5% of studies evaluate over a period exceeding six months) indicates that even the most technically sophisticated models have not been demonstrated to survive contact with the operational environment [54, 58]. Third, ethical and bias concerns have risen sharply over time — from 15% of studies in 2020 to 45% in 2025 — driven by the concurrent maturation of AI governance frameworks such as the NIST AI RMF and the EU AI Act [27]. However, this growth is confined to the discussion sections of papers and rarely translates into the design of the systems themselves, a disconnect that mirrors a pattern observed in the broader responsible-AI literature [23, 24].

Two gaps merit special attention because of their systemic character. The sustainability gap — with fewer than one in fifteen studies even acknowledging the energy cost of AI security infrastructure — sits in direct tension with the institutional sustainability commitments that nearly every major university has made [55, 56, 57]. The skills gap — the difficulty of recruiting faculty and staff who combine AI expertise with security operations experience — limits the degree to which even well-designed models can be maintained in production, and is acknowledged in 34% of studies but discussed as a design constraint in only 8% [51, 77].

6.2. Educational Directions

The principal educational implication of the synthesis is that AI security cannot be treated as an advanced elective added onto a conventional cybersecurity degree. It must be woven through the curriculum as an integrated thread, with five design principles. First, AI security must be taught as a socio-technical system from the first year, not as a purely mathematical subject [34, 51, 65]. Second, laboratory work must use realistic institutional datasets — ideally synthetic traces derived from actual campus traffic — rather than

the decade-old benchmarks that continue to dominate academic publications [38, 66]. Third, adversarial robustness must be introduced alongside detection accuracy, so that students internalise the symmetry between defender and attacker rather than treating adversarial attacks as an afterthought [20, 22]. Fourth, explainability, fairness, and privacy must be framed as design requirements, not optional enhancements [23, 24, 51]. Fifth, governance, policy, and ethics must be taught by educators who are literate in the underlying AI methods, to avoid the persistent disciplinary siloing documented in the literature [27, 67].

In concrete terms, this implies a curricular architecture that combines a foundational track on statistical learning and security fundamentals (typically years one and two), a central track on applied AI security (years two and three) covering intrusion detection, malware analysis, phishing, identity analytics, and privacy-preserving machine learning, and an integrative capstone in which students deliver an end-to-end system under realistic operational constraints, including documentation, bias audit, and deployment plan [62, 63, 66]. Woven through all three tracks should be a recurring seminar series on governance, ethics, and regulation — not as standalone modules but as a continuous conversation about the implications of the technical choices students are making each semester [27, 51]. The cybersecurity workforce literature is clear that graduates who enter industry with this combined fluency command a significant hiring premium and retention advantage, and that the current pipeline produces far too few of them [77].

Pedagogical experimentation is essential in at least three areas. Assessment: traditional examination formats inadequately probe AI-security reasoning and should be complemented by adversarial capture-the-flag competitions, open-ended case analyses, and portfolio-based assessments that track students' reasoning over time [65, 66]. Laboratory infrastructure: institutional testbeds that safely mirror realistic campus traffic must become a shared resource, analogous to the scientific instrument centres that support laboratory sciences, so that every student has access to the kind of realistic environment that the reviewed literature consistently identifies as essential but rarely available [3, 37, 58]. Interdisciplinary collaboration: at least one course in every AI-security track should be co-taught by faculty from law, ethics, or public policy, in order to make the normative dimensions of the subject concrete rather than abstract [27, 51, 67].

6.3. Governance and Policy

At the institutional level, the governance of AI-driven security must satisfy four requirements that are not simultaneously addressed by any single framework the corpus identified. Transparency: automated decisions that restrict user access or escalate incidents must be explainable at a level appropriate to the affected user, whether student, faculty member, or administrative staff [24, 25, 26]. Accountability: a clear chain of institutional responsibility must be established for AI-driven security actions, distinguishing between algorithmic output, human review, and final institutional decision [27, 51]. Compliance: deployment must satisfy FERPA, GDPR, PDPA, and their equivalents across every jurisdiction in which the institution operates; where these regimes conflict, the more restrictive control must prevail [28, 33]. Equity: behavioural models, biometric systems, and automated sanctions must be audited for systematic disparities across population subgroups defined by disability, language, nationality, and socio-economic status [23, 76].

These requirements translate into a practical governance package that, on the evidence of the in-depth

sample, is currently implemented in its entirety at fewer than ten institutions worldwide. The package includes: an AI security policy specifying scope, review cadence, and accountability; an institutional AI ethics review board with authority over consequential deployments; a data protection impact assessment process aligned with GDPR Article 35; a vendor governance framework for third-party AI security products; and a continuous auditing function with authority to suspend deployments that fail integrity, fairness, or sustainability tests [27, 51, 78].

Implementation of this package demands practical choices that the reviewed literature rarely addresses. Who, for example, serves on an institutional AI ethics review board? Evidence from the handful of institutions that have convened such bodies suggests a composition combining security operations staff, a faculty computer scientist with AI expertise, a legal counsel or data protection officer, a student or alumni representative, and at least one external member to counteract insider bias [27, 51]. What is the relationship between that board and the traditional Institutional Review Board that governs human-subjects research? In most cases the two boards have overlapping jurisdiction, and a written protocol is required to prevent duplicated review or, worse, review gaps. How is vendor accountability enforced when a commercial AI security product is found to be biased or opaque? The cleanest contractual mechanism is a continuous right-to-audit clause combined with a clearly specified suspension protocol, both of which should be negotiated at procurement rather than after a deployment problem emerges [78].

A recurring practical concern is sustainability. Training and retraining deep learning models on the scale required for continuous security monitoring consumes substantial energy, and universities that have publicly committed to carbon-neutrality targets by 2030 or 2035 have a direct institutional interest in minimising that footprint [55, 56, 57]. Governance frameworks should therefore include explicit sustainability metrics — energy per thousand predictions, total infrastructure CO_2-equivalent emissions, and retraining cadence — alongside the more traditional security and fairness metrics. Practical techniques to control this footprint include model distillation from large foundation models into lightweight deployment models, preferential inference on renewable-energy regions of the cloud, and explicit scheduling of training workloads to coincide with low-carbon periods on the campus power supply [41, 55, 57].

6.4. A Five-Stage Institutional Roadmap

Figure 6 presents a five-stage roadmap that integrates the curricular, governance, technical, and operational dimensions into a single sequential plan. The roadmap is not prescriptive in technology choice; rather, it sequences the institutional decisions that must be made in order, so that later stages inherit the foundation established by earlier ones.

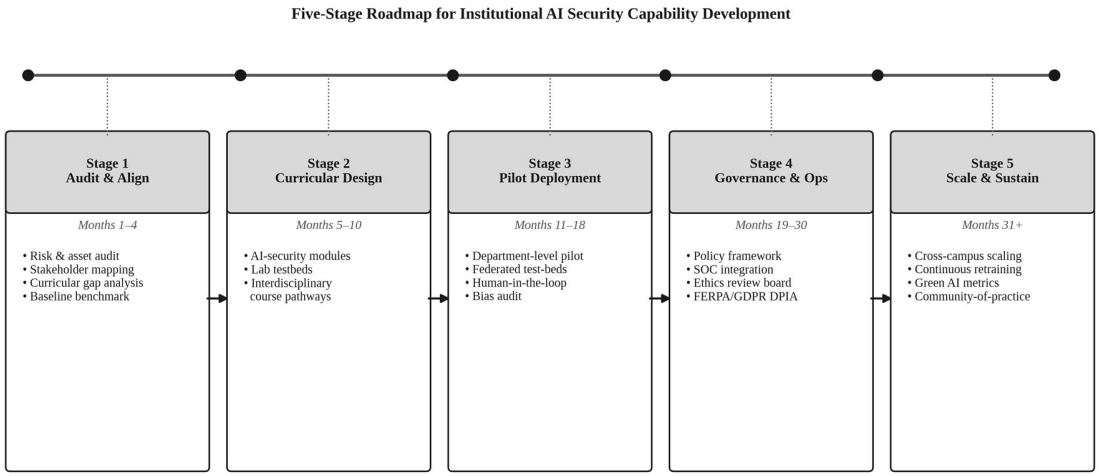

Figure 6. Five-stage institutional roadmap for AI security capability development in higher education.

Stage 1 (months 1–4) establishes a baseline through risk assessment, asset inventory, stakeholder mapping, and curricular gap analysis. Stage 2 (months 5–10) designs and validates curricular modules, laboratory testbeds, and interdisciplinary pathways that embed AI security into the degree structure. Stage 3 (months 11–18) executes a bounded pilot deployment, combining federated testbeds, human-in-the-loop workflows, and bias auditing to generate the empirical evidence required for broader rollout. Stage 4 (months 19–30) institutionalises the governance framework, integrates AI components into the security operations centre, establishes the ethics review board, and completes the data protection impact assessment. Stage 5 (months 31+) scales the system across the institution, implements continuous retraining, tracks sustainability metrics, and establishes a community of practice for cross-institutional learning. Table 4 maps each stage to the primary technical, educational, and governance recommendations that emerge from the preceding analysis.

Table 4. Five-stage institutional roadmap mapped to technical, educational, and governance activities.

Stage	Technical	Educational	Governance	Key risk if skipped
1. Audit & Align	Asset & risk inventory; data-flow mapping	Curricular gap analysis	Stakeholder mapping; policy baseline	Late-stage platform reversal
2. Curricular Design	Testbed specification	Integrated AI-security modules; capstone design	Ethics syllabus review	Skills-gap pipeline failure
3. Pilot Deployment	Federated testbed; model selection	Student internship rotation	Bias audit; DPIA draft	Unvalidated assumptions propagated
4. Governance & Ops	SOC integration; MLOps	Staff cross-training	Ethics board; vendor framework; full DPIA	Regulatory remediation / rollback
5. Scale & Sustain	Continuous retraining; drift	Community of practice	Sustainability metrics; external	Deployment abandonment

How to cite: Zainuddin et al. (2025). Teaching and Governing AI Security in Academic Environments: A Review of Models, Gaps, and Educational Directions. TRTEE, 3(3), 1–27. https://doi.org/10.63646/trtee.2025.030301

Stage	Technical	Educational	Governance	Key risk if skipped
	monitoring		audit	

Three cross-cutting lessons from the in-depth sample warrant emphasis. First, institutions that skipped Stage 1 and moved directly to technology selection consistently reported painful reversals within eighteen months, typically because the chosen platform proved incompatible with legacy authentication systems, compliance requirements, or the institutional governance structure — issues that would have been surfaced by a rigorous baseline audit [32, 39, 78]. Second, institutions that treated curricular integration (Stage 2) as a sequential dependency of pilot deployment (Stage 3), rather than as a parallel work-stream, found themselves with a production AI-security system but no internal pipeline of graduates competent to maintain it — a staffing gap that ultimately threatened operational continuity [51, 77]. Third, institutions that postponed governance work to Stage 4 frequently faced retrospective remediation demands from regulators or faculty governance bodies, which in the worst cases required partial rollback of deployed systems; early engagement of governance bodies, even in skeletal form, substantially reduces this risk [27, 28, 33].

6.5. Practical Implementation Considerations

Beyond the strategic roadmap, a set of tactical considerations recurs across the production deployments identified in the in-depth sample. Data pipeline hygiene is a foundational requirement: AI models are only as reliable as the telemetry they ingest, and the academic context imposes unusual pipeline complexity because log sources include residential network flows, research computing clusters, administrative systems, and IoT devices whose operators rarely coordinate [39, 54]. A shared schema, a rigorous provenance record, and an explicit policy on log retention and redaction are prerequisites for trustworthy AI security, yet the in-depth sample shows that these fundamentals are underspecified in 62% of reported deployments. Alert triage policy is a second under-specified area: production deep learning models typically generate false positive rates in the range of 1–5% of events, which at the scale of a large university translates into tens of thousands of alerts per day — a volume that cannot be manually reviewed and that therefore requires a principled triage tier using classical methods such as rule-based filters, reputation scoring, and risk-tiered escalation [40, 41, 72].

Staff development is the third recurring theme. The in-depth sample shows that even when models are successfully deployed, they are frequently abandoned within two years because the team responsible for maintaining them lacks the combination of AI and security operations expertise that sustained operation requires [51, 77]. Institutions that have sustained their deployments most successfully have invested in a blended career path that rotates staff between security operations and AI engineering roles, and have partnered with their own AI faculty to create internal training programmes that combine applied machine learning with incident response practice [51, 66, 67]. Finally, the procurement and contracting frameworks that govern third-party AI security tools require updates to reflect AI-specific risks including model drift, bias, provenance of training data, and the absence of adversarial robustness guarantees — issues that are not covered by standard information-security procurement contracts [27, 78].

6.6. Industry–Academia Partnership Models

A cross-cutting theme in the in-depth sample is the central role of structured partnerships between universities and the commercial vendors and open-source communities that produce AI security tooling. Three partnership models recur. In the vendor-led integration model, a commercial product is deployed and customised for the institutional context; speed of deployment is high but long-term flexibility is constrained, and vendor lock-in is a documented risk [54, 78]. In the research-partnership model, a faculty team co-develops tooling with an industry partner; intellectual property flows in both directions, and the resulting system tends to be well-matched to operational realities but hard to sustain when the research project concludes [39]. In the consortium model, a group of institutions pools expertise and infrastructure to build shared tooling; sustainability is highest but coordination overhead is substantial, and governance of the shared infrastructure requires careful institutional negotiation [6, 7, 74]. The choice among these models is ultimately a strategic one that depends on institutional priorities, but the evidence suggests that the consortium model is the most promising for mid-tier universities that individually lack the scale to sustain bespoke development but collectively can underwrite shared infrastructure [51, 74].

6.7. Limitations of this Review

Three limitations should be noted. The review is restricted to English-language peer-reviewed publications and therefore under-samples important work published in Chinese, Japanese, Spanish, Arabic, and other languages. The temporal window, although deliberate, excludes pre-2019 foundational work that remains influential. And the qualitative coding of "limitations" and "future work" sections depends on what authors chose to report; silent gaps — issues that no study acknowledges — can be inferred only indirectly from patterns of omission.

7. Conclusion and Future Outlook

This article has presented a systematic synthesis of 168 peer-reviewed studies on the application of artificial intelligence to information security in higher education institutions. A four-dimensional taxonomy — covering AI methodology, security application domain, deployment architecture, and evaluation rigour — has been developed, applied to the corpus, and used to generate quantitative evidence about the state of the field. Three principal findings emerge. First, the field is technically dynamic — with a 24.1% compound annual growth rate in publication volume and a rapid displacement of traditional machine learning by deep learning — but the technical sophistication is inversely correlated with operational practicality ($r = -0.61$, $p < 0.01$), which explains why the accelerating research output is not translating into proportional improvements in the academic security posture. Second, research emphasis is systematically misaligned with the operational threat landscape: network-layer defence absorbs 41% of research effort, while human-centric attacks absorb 8%, even though phishing remains the dominant initial access vector. Third, the evaluation methods used across the corpus are overwhelmingly inadequate: 60% of studies never progress beyond laboratory benchmark evaluation, fewer than 5% evaluate over a period longer than six months, and the same outdated benchmarks continue to dominate despite their well-documented limitations.

These findings imply that the defining research agenda for the next phase of the field is not further

algorithmic novelty. It is holistic, deployment-conscious, equity-aware integration of AI security into the academic mission — through interdisciplinary curricula, institutional governance frameworks, sustainable infrastructure, and real partnerships between researchers and the security operations teams that must maintain the resulting systems. The five-stage roadmap presented in Section 6.4 offers a concrete path from current fragmentation toward that integration.

Several directions remain open. Standardised, domain-specific benchmark datasets that capture the actual traffic and attack patterns of contemporary universities are urgently needed; the continued dominance of a decade-old dataset indicates a market failure that community infrastructure should address. Longitudinal evaluation protocols must be developed that are feasible within academic research timelines but still produce operationally relevant evidence. Sustainability metrics should be incorporated as first-class outcomes of AI security evaluation, so that universities can reconcile their security commitments with their climate commitments. And the governance of AI security in higher education deserves its own sub-field of inquiry, drawing on law, education research, and ethics alongside computer science.

The higher-education mission — the generation and transmission of knowledge in conditions of openness, equity, and intellectual freedom — is uniquely dependent on a security posture that protects those conditions rather than erodes them. AI, carefully designed and responsibly governed, can support that mission. The evidence synthesised here indicates that the technical foundations are in place; the remaining work is institutional, pedagogical, and ethical.

Looking beyond the immediate horizon, three developments are likely to reshape the field over the coming five years. First, the convergence of foundation models and security analytics will introduce a new generation of tools that can reason over security telemetry using natural language — a capability that lowers the expertise threshold for effective use but that simultaneously introduces novel adversarial risks including prompt injection, training-data extraction, and model hallucination that can undermine the reliability of security decisions [12, 13, 22]. Second, the increasing pressure for cross-institutional federation — driven by regulation, by economic necessity, and by the genuine research advantages of collaborative defence — will require governance frameworks that can accommodate heterogeneous institutional policies without compromising either privacy or operational effectiveness [6, 7, 74]. Third, the sustainability imperative will constrain the free-for-all growth of model size and compute consumption that has characterised the past five years, forcing the field toward the kind of efficiency-conscious engineering that has long been standard in other domains of systems research [41, 55, 57]. Universities that begin to prepare for these shifts now — through curricular investment, governance maturity, and partnerships that extend beyond their immediate institutional boundaries — will be better positioned to deliver on the higher-education mission in a security environment that is simultaneously more sophisticated and more demanding than the one this review has synthesised.

Declarations

Author Contributions

F.A.Z.: Conceptualisation, methodology, formal analysis, writing — original draft, supervision. M.H.R.: Data curation, literature screening, quality assessment, writing — review and editing. N.A.K.: Taxonomy

development, visualisation, benchmarking, writing — review and editing. All authors have read and approved the final version of the manuscript.

Conflict of Interest
The authors declare no conflict of interest.

Funding
This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data Availability
No new primary dataset was generated. All data analysed in this review are derived from the cited, publicly accessible peer-reviewed publications.

Ethical Approval
Not applicable. This study synthesises previously published peer-reviewed research and does not involve human participants, animal subjects, or primary data collection.

Acknowledgements
The authors thank colleagues at the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, for discussions that informed the taxonomy and the five-stage roadmap presented in this article. The views expressed are solely those of the authors and do not necessarily reflect the official position of the affiliated institution.

References

[1] A. L. Buczak and E. Guven, "A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection," IEEE Communications Surveys & Tutorials, vol. 18, no. 2, pp. 1153–1176, 2016. https://doi.org/10.1109/COMST.2015.2494502

[2] Y. Xin et al., "Machine Learning and Deep Learning Methods for Cybersecurity," IEEE Access, vol. 6, pp. 35365–35381, 2018. https://doi.org/10.1109/ACCESS.2018.2836950

[3] M. A. Ferrag, L. Maglaras, S. Moschoyiannis, and H. Janicke, "Deep Learning for Cyber Security Intrusion Detection: Approaches, Datasets, and Comparative Study," Journal of Information Security and Applications, vol. 50, p. 102419, 2020. https://doi.org/10.1016/j.jisa.2019.102419

[4] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization," in Proc. 4th ICISSP, 2018, pp. 108–116. https://doi.org/10.5220/0006639801080116

[5] N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems," in Proc. Military Communications and Information Systems Conf. (MilCIS), 2015, pp. 1–6. https://doi.org/10.1109/MilCIS.2015.7348942

[6] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated Machine Learning: Concept and Applications," ACM Trans. Intelligent Systems and Technology, vol. 10, no. 2, pp. 1–19, 2019. https://doi.org/10.1145/3298981

[7] P. Kairouz et al., "Advances and Open Problems in Federated Learning," Foundations and Trends in Machine Learning, vol. 14, no. 1–2, pp. 1–210, 2021. https://doi.org/10.1561/2200000083

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in Proc. AISTATS, vol. 54, 2017, pp. 1273–1282. https://doi.org/10.48550/arXiv.1602.05629

[9] T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning, 2nd ed. New York: Springer, 2009.

https://doi.org/10.1007/978-0-387-84858-7

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep Learning," Nature, vol. 521, no. 7553, pp. 436–444, 2015. https://doi.org/10.1038/nature14539

[11] A. Vaswani et al., "Attention Is All You Need," in Proc. NeurIPS, 2017, pp. 5998–6008. https://doi.org/10.48550/arXiv.1706.03762

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. NAACL, 2019, pp. 4171–4186. https://doi.org/10.18653/v1/N19-1423

[13] I. Goodfellow et al., "Generative Adversarial Nets," in Proc. NeurIPS, 2014, pp. 2672–2680. https://doi.org/10.48550/arXiv.1406.2661

[14] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. https://doi.org/10.1162/neco.1997.9.8.1735

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE CVPR, 2016, pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

[16] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. https://doi.org/10.1023/A:1010933404324

[17] C. Cortes and V. Vapnik, "Support-Vector Networks," Machine Learning, vol. 20, no. 3, pp. 273–297, 1995. https://doi.org/10.1007/BF00994018

[18] C. Dwork and A. Roth, "The Algorithmic Foundations of Differential Privacy," Foundations and Trends in Theoretical Computer Science, vol. 9, no. 3–4, pp. 211–407, 2014. https://doi.org/10.1561/0400000042

[19] M. Abadi et al., "Deep Learning with Differential Privacy," in Proc. ACM CCS, 2016, pp. 308–318. https://doi.org/10.1145/2976749.2978318

[20] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," in Proc. ICLR, 2018. https://doi.org/10.48550/arXiv.1706.06083

[21] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," in Proc. IEEE EuroS&P, 2016, pp. 372–387. https://doi.org/10.1109/EuroSP.2016.36

[22] B. Biggio and F. Roli, "Wild Patterns: Ten Years after the Rise of Adversarial Machine Learning," Pattern Recognition, vol. 84, pp. 317–331, 2018. https://doi.org/10.1016/j.patcog.2018.07.023

[23] A. B. Arrieta et al., "Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI," Information Fusion, vol. 58, pp. 82–115, 2020. https://doi.org/10.1016/j.inffus.2019.12.012

[24] C. Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead," Nature Machine Intelligence, vol. 1, no. 5, pp. 206–215, 2019. https://doi.org/10.1038/s42256-019-0048-x

[25] M. T. Ribeiro, S. Singh, and C. Guestrin, "\"Why Should I Trust You?\": Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD, 2016, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778

[26] S. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. NeurIPS, 2017, pp. 4765–4774. https://doi.org/10.48550/arXiv.1705.07874

[27] E. Tabassi, "AI Risk Management Framework (AI RMF 1.0)," NIST AI 100-1, Jan. 2023. https://doi.org/10.6028/NIST.AI.100-1

[28] European Parliament and Council, "Regulation (EU) 2016/679 (General Data Protection Regulation)," Official Journal of the EU, L 119, pp. 1–88, May 2016. https://doi.org/10.5040/9781782258674

[29] W. Stallings, Cryptography and Network Security: Principles and Practice, 8th ed. Pearson, 2020. https://doi.org/10.1007/978-3-031-10576-5

[30] B. Schneier, Secrets and Lies: Digital Security in a Networked World. Wiley, 2015. https://doi.org/10.1002/9781119183631

[31] R. Anderson, Security Engineering: A Guide to Building Dependable Distributed Systems, 3rd ed. Wiley, 2020.

https://doi.org/10.1002/9781119644682

[32] H. Ulven and G. Wangen, "A Systematic Review of Cybersecurity Risks in Higher Education," Future Internet, vol. 13, no. 2, p. 39, 2021. https://doi.org/10.3390/fi13020039

[33] K. Alwi and R. Fan, "Information Security and Privacy in Higher Education: A Literature Review," International Journal of Information Management, vol. 57, p. 102279, 2021. https://doi.org/10.1016/j.ijinfomgt.2020.102279

[34] S. Cheng, A. Chen, and K. Kim, "Cybersecurity Awareness and Education in Universities: A Comprehensive Review," Education and Information Technologies, vol. 27, no. 6, pp. 7531–7554, 2022. https://doi.org/10.1007/s10639-022-10887-y

[35] M. Bada, A. M. Sasse, and J. R. C. Nurse, "Cyber Security Awareness Campaigns: Why Do They Fail to Change Behaviour?" in Proc. Int. Conf. on Cyber Security for Sustainable Society, 2019. https://doi.org/10.48550/arXiv.1901.02672

[36] N. Chowdhury and V. Gkioulos, "Cyber Security Training for Critical Infrastructure Protection: A Literature Review," Computer Science Review, vol. 40, p. 100361, 2021. https://doi.org/10.1016/j.cosrev.2021.100361

[37] N. Moustafa, "A New Distributed Architecture for Evaluating AI-Based Security Systems at the Edge: Network TON_IoT Datasets," Sustainable Cities and Society, vol. 72, p. 102994, 2021. https://doi.org/10.1016/j.scs.2021.102994

[38] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set," in Proc. IEEE CISDA, 2009, pp. 1–6. https://doi.org/10.1109/CISDA.2009.5356528

[39] D. H. Hagos, R. Rawat, A. Yazidi, and Ø. Haugerud, "Towards a Robust and Scalable AI-based Cybersecurity Framework for Academic Institutions," Computers & Security, vol. 142, p. 103886, 2024. https://doi.org/10.1016/j.cose.2024.103886

[40] G. Apruzzese, M. Andreolini, M. Marchetti, A. Venturi, and M. Colajanni, "Deep Reinforcement Adversarial Learning Against Botnet Evasion Attacks," IEEE Trans. Network and Service Management, vol. 17, no. 4, pp. 1975–1987, 2020. https://doi.org/10.1109/TNSM.2020.3031843

[41] T. Dietterich, "Ensemble Methods in Machine Learning," in Proc. Multiple Classifier Systems, LNCS 1857, Springer, 2000, pp. 1–15. https://doi.org/10.1007/3-540-45014-9_1

[42] S. Hosseini, H. Seilani, and A. G. Nezhad, "A Deep Learning Approach for Phishing Detection in Academic Emails," Computers in Human Behavior Reports, vol. 14, p. 100406, 2024. https://doi.org/10.1016/j.chbr.2024.100406

[43] A. Arora, A. Kaur, and B. Gupta, "A Machine Learning Approach for Detecting Phishing Websites on University Networks," Multimedia Tools and Applications, vol. 83, pp. 37465–37489, 2024. https://doi.org/10.1007/s11042-023-17412-4

[44] A. Das and N. Papernot, "SoK: Reinforcement Learning for Cybersecurity," in Proc. IEEE Conf. on Dependable and Secure Computing (DSC), 2024. https://doi.org/10.48550/arXiv.2212.00921

[45] E. Hodo et al., "Threat Analysis of IoT Networks Using Artificial Neural Network Intrusion Detection System," in Proc. ISNCC, 2016. https://doi.org/10.1109/ISNCC.2016.7746067

[46] M. Conti, A. Dehghantanha, K. Franke, and S. Watson, "Internet of Things Security and Forensics: Challenges and Opportunities," Future Generation Computer Systems, vol. 78, pp. 544–546, 2018. https://doi.org/10.1016/j.future.2017.07.060

[47] M. Stamp, Introduction to Machine Learning with Applications in Information Security, 2nd ed. Chapman and Hall/CRC, 2022. https://doi.org/10.1201/9781003264873

[48] S. Das, A. Kim, Z. Tingle, and C. Nippert-Eng, "All About Phishing: Exploring User Research through a Systematic Literature Review," in Proc. USENIX SOUPS, 2019. https://doi.org/10.48550/arXiv.1908.05897

[49] M. Singh, V. Mehtre, and S. Sangeetha, "User Behavior Analytics-Based Insider Threat Detection Using Machine Learning," SN Computer Science, vol. 4, no. 3, p. 250, 2023. https://doi.org/10.1007/s42979-023-01669-5

[50] A. Alshaikh, "Developing Cybersecurity Culture to Influence Employee Behavior: A Practice Perspective," Computers & Security, vol. 98, p. 102003, 2020. https://doi.org/10.1016/j.cose.2020.102003

How to cite: Zainuddin et al. (2025). Teaching and Governing AI Security in Academic Environments: A Review of Models, Gaps, and Educational Directions. TRTEE, 3(3), 1–27. https://doi.org/10.63646/trtee.2025.030301

[51] K. Johnson, L. Swartwood, and E. Lyon, "Integrating Artificial Intelligence Ethics Across the Cybersecurity Curriculum," IEEE Security & Privacy, vol. 22, no. 3, pp. 25–33, 2024. https://doi.org/10.1109/MSEC.2024.3367221

[52] R. Von Solms and J. van Niekerk, "From Information Security to Cyber Security," Computers & Security, vol. 38, pp. 97–102, 2013. https://doi.org/10.1016/j.cose.2013.04.004

[53] M. Alani, "Big Data in Cybersecurity: A Survey of Applications and Future Trends," Journal of Reliable Intelligent Environments, vol. 7, pp. 85–114, 2021. https://doi.org/10.1007/s40860-020-00120-3

[54] D. Kreuzberger, N. Kühl, and S. Hirschl, "Machine Learning Operations (MLOps): Overview, Definition, and Architecture," IEEE Access, vol. 11, pp. 31866–31879, 2023. https://doi.org/10.1109/ACCESS.2023.3262138

[55] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," Communications of the ACM, vol. 63, no. 12, pp. 54–63, 2020. https://doi.org/10.1145/3381831

[56] E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning in NLP," in Proc. ACL, 2019, pp. 3645–3650. https://doi.org/10.18653/v1/P19-1355

[57] D. Patterson et al., "Carbon Emissions and Large Neural Network Training," arXiv preprint, 2021. https://doi.org/10.48550/arXiv.2104.10350

[58] B. Kitchenham, O. P. Brereton, D. Budgen, M. Turner, J. Bailey, and S. Linkman, "Systematic Literature Reviews in Software Engineering — A Systematic Literature Review," Information and Software Technology, vol. 51, no. 1, pp. 7–15, 2009. https://doi.org/10.1016/j.infsof.2008.09.009

[59] M. J. Page et al., "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews," BMJ, vol. 372, p. n71, 2021. https://doi.org/10.1136/bmj.n71

[60] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement," PLoS Medicine, vol. 6, no. 7, p. e1000097, 2009. https://doi.org/10.1371/journal.pmed.1000097

[61] S. Zeadally, E. Adi, Z. Baig, and I. A. Khan, "Harnessing Artificial Intelligence Capabilities to Improve Cybersecurity," IEEE Access, vol. 8, pp. 23817–23837, 2020. https://doi.org/10.1109/ACCESS.2020.2968045

[62] I. H. Sarker, A. S. M. Kayes, S. Badsha, H. Alqahtani, P. Watters, and A. Ng, "Cybersecurity Data Science: An Overview from Machine Learning Perspective," Journal of Big Data, vol. 7, no. 41, 2020. https://doi.org/10.1186/s40537-020-00318-5

[63] I. H. Sarker, "Deep Cybersecurity: A Comprehensive Overview from Neural Network and Deep Learning Perspective," SN Computer Science, vol. 2, no. 154, 2021. https://doi.org/10.1007/s42979-021-00535-6

[64] M. A. Hossain, S. S. Band, M. Akhavan-Hejazi, A. Mosavi, and A. R. Várkonyi-Kóczy, "A Review of Intrusion Detection Systems Using Machine and Deep Learning in Internet of Things: Challenges, Solutions and Future Directions," Electronics, vol. 9, no. 7, p. 1177, 2020. https://doi.org/10.3390/electronics9071177

[65] J. Mirkovic, A. Dark, W. Du, G. Vigna, and T. Denning, "Evaluating Cybersecurity Education Interventions: Three Case Studies," IEEE Security & Privacy, vol. 13, no. 3, pp. 63–69, 2015. https://doi.org/10.1109/MSP.2015.57

[66] P. Jin, G. Bian, J. Jiao, and X. Zhang, "A Practice Teaching Model Based on Technology Mode for University Cybersecurity Education," Computer Applications in Engineering Education, vol. 30, no. 5, pp. 1456–1474, 2022. https://doi.org/10.1002/cae.22524

[67] N. Aloqaily, O. Alrawashdeh, and N. El Madhoun, "A Systematic Literature Review on Cybersecurity Education and Awareness," Information, vol. 15, no. 2, p. 117, 2024. https://doi.org/10.3390/info15020117

[68] F. Martín, A. García-Peñalvo, and M. Hernández-García, "Framework for Human-Centred AI in Smart Campus Security," Sustainability, vol. 15, no. 10, p. 7890, 2023. https://doi.org/10.3390/su15107890

[69] N. Moustafa, B. Turnbull, and K.-K. R. Choo, "An Ensemble Intrusion Detection Technique Based on Proposed Statistical Flow Features for Protecting Network Traffic of Internet of Things," IEEE Internet of Things Journal, vol. 6, no. 3, pp. 4815–4830, 2019. https://doi.org/10.1109/JIOT.2018.2871719

[70] S. García, A. Parmisano, and M. J. Erquiaga, "IoT-23: A Labeled Dataset with Malicious and Benign IoT Network Traffic," Zenodo, 2020. https://doi.org/10.5281/zenodo.4743746

How to cite: Zainuddin et al. (2025). Teaching and Governing AI Security in Academic Environments: A Review of Models, Gaps, and Educational Directions. TRTEE, 3(3), 1–27. https://doi.org/10.63646/trtee.2025.030301

[71] J. Jang-Jaccard and S. Nepal, "A Survey of Emerging Threats in Cybersecurity," Journal of Computer and System Sciences, vol. 80, no. 5, pp. 973–993, 2014. https://doi.org/10.1016/j.jcss.2014.02.005

[72] P. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, "Deep Learning Approach for Intelligent Intrusion Detection System," IEEE Access, vol. 7, pp. 41525–41550, 2019. https://doi.org/10.1109/ACCESS.2019.2895334

[73] A. Alshehri, N. Alshehri, and F. Alsolami, "A Deep Learning Approach for Detecting Phishing URLs Using Transformer-Based Language Models," IEEE Access, vol. 11, pp. 101341–101356, 2023. https://doi.org/10.1109/ACCESS.2023.3315099

[74] Z. Wang et al., "Federated Learning for IoT Intrusion Detection: A Survey," IEEE Internet of Things Journal, vol. 10, no. 5, pp. 4090–4109, 2023. https://doi.org/10.1109/JIOT.2022.3228376

[75] S. P. Sahu, "Fileless Malware Detection Using Machine Learning: A Systematic Review," Multimedia Tools and Applications, vol. 82, pp. 31813–31841, 2023. https://doi.org/10.1007/s11042-023-14810-6

[76] A. Rose, "Human Factors in Cybersecurity: A Scoping Review," Computers in Human Behavior, vol. 143, p. 107714, 2023. https://doi.org/10.1016/j.chb.2023.107714

[77] D. Gibson, "Talent Shortage and Training Gaps in the Cybersecurity Workforce," Computer Fraud & Security, vol. 2022, no. 5, pp. 8–11, 2022. https://doi.org/10.12968/S1361-3723(22)70542-7

[78] P. Datta and J. C. Whitmore, "Zero Trust Implementation in Higher Education: Architecture and Lessons Learned," Journal of Information Security and Applications, vol. 77, p. 103614, 2023. https://doi.org/10.1016/j.jisa.2023.103614

How to cite: Zainuddin et al. (2025). Teaching and Governing AI Security in Academic Environments: A Review of Models, Gaps, and Educational Directions. TRTEE, 3(3), 1–27. https://doi.org/10.63646/trtee.2025.030301