US JOURNAL OF NEW INSIGHTS IN TECH & EDUCATION

Vol. 6, No. 1 · Article ID: USJNITE-2602 · Journal Homepage: www.edtechinsights.org

# Managing Cybersecurity Risk in AI-Driven Educational Platforms: An ERM-Based Approach

**Md Rasel Ul Alam**[1], **Kanita Haider**[2]*, **Oishe Al Mariz**[3]

[1] *University of the Cumberlands, Kentucky, USA*

[2], [3] *Chittagong University of Engineering and Technology, Chattogram 4349, BD*

**Corresponding author: kanita.haider4422@gmail.com*

*Received 14 April, 2026; Revised 12 June, 2026; Accepted 6 July, 2026; Published: 27July, 2026*

**KEYWORDS**

Artificial Intelligence in Education,

Enterprise Risk Management,

AI Governance,

Algorithmic Bias,

Educational Information Systems,

Adaptive Learning Platforms,

Data Privacy

***Abstract***

The rapid adoption of AI-driven adaptive learning platforms, intelligent tutoring systems, automated grading tools, and generative-AI tutoring assistants has introduced a category of institutional risk that extends beyond conventional cloud and data security: algorithmic bias, model opacity, autonomous or semi-autonomous decision-making about students, and the regulatory scrutiny now attached to AI systems used for admissions, assessment, and monitoring. This paper argues that Enterprise Risk Management (ERM), and specifically the baseline cloud-security requirements and governance principles developed for general enterprise cloud deployments, provide a transferable risk-management foundation for AI-driven educational platforms, but require a dedicated AI-specific risk layer that the enterprise baseline alone does not anticipate. Drawing on enterprise-sector research on baseline cloud-security requirements within an ERM framework, on the strategies, challenges, and organizational-success factors of ERM implementation, on the education-sector AI-ethics literature, on international AI-governance standards (the NIST AI Risk Management Framework and ISO/IEC 42001), and on the risk-tier classifications introduced by the European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689, Annex III), this paper proposes a framework comprising an AI-specific risk register mapped to ERM controls, a four-tier AI risk classification scheme for educational use cases, a human-oversight-and-explainability layer, a quantitative risk-scoring model, and an institutional governance structure. The framework is demonstrated through an illustrative twelve-month institutional pilot scenario, following design-science research (DSR) evaluation conventions, and is contextualized against recently reported survey evidence on rising faculty and administrator concern about AI bias and data-privacy risk in higher education, an illustrative risk-register heat-map, and an illustrative cost-versus-risk-reduction analysis. The paper concludes with adoption barriers, a positioning of the framework against the EU AI Act, the NIST AI RMF, and ISO/IEC 42001, and recommendations for institutions deploying AI-driven learning technologies.

## 1. Introduction

Artificial intelligence has moved from a peripheral feature of educational technology to a core component of it: adaptive learning platforms now use machine-learning models to sequence content and predict performance, automated systems support or perform grading and admissions screening, and generative-AI tutoring assistants increasingly mediate direct student interaction. Tan et al. (2025) document this maturation in their review of AI-enabled adaptive learning platforms, finding that such systems now span diverse pedagogical foundations and algorithmic implementations while facing persistent, unresolved challenges around data privacy and institutional support for responsible deployment. This growth in capability has been accompanied by a comparable growth in documented institutional risk. Al-Zahrani (2024), synthesizing a systematic review and a validated survey of 260 respondents at a Saudi Arabian university, identifies ten interrelated areas of concern with AI in education, including data privacy and security, algorithmic bias, and diminished transparency, and finds that these concerns are not isolated but mutually reinforcing. At a larger, cross-institutional scale, Ellucian's (2024) second annual survey of 445 faculty and administrators across more than 330 institutions found that concern about bias in AI models rose from 36% to 49% of respondents between 2023 and 2024, while concern about data privacy and security rose from 50% to 59% over the same period — evidence that institutional unease is intensifying even as adoption accelerates. Figure 2 in Section 4.2 summarizes this survey evidence in full.

Regulatory frameworks have begun to formalize what practitioner surveys describe informally. The European Union's

Artificial Intelligence Act (Regulation (EU) 2024/1689) explicitly classifies AI systems used to determine admission or access to educational institutions, to evaluate learning outcomes, to assess the appropriate level of education an

individual will receive, and to monitor students during examinations, as high-risk applications under Annex III, triggering specific obligations for risk management, data governance, and human oversight. This regulatory classification is significant for the present paper because it provides an externally validated, non-arbitrary basis for risk-tiering AI systems in education — a basis this paper adopts directly in Section 3.2, rather than proposing a new, uncorroborated tiering scheme. The Act is not alone in this respect: as reviewed in Section 2.5, voluntary international standards — the U.S. National Institute of Standards and Technology's AI Risk Management Framework and ISO/IEC 42001 — have converged on a broadly similar structure of risk identification, control mapping, and human oversight, suggesting that the framework proposed here rests on a wider consensus than any single jurisdiction's binding law.

It is worth situating this regulatory backdrop against the pace of institutional AI adoption it is attempting to govern. Tan et al.'s (2025) review and Ellucian's (2024) survey, taken together, describe an environment in which adaptive learning platforms, automated grading tools, and generative-AI tutoring assistants have moved from pilot status to routine institutional infrastructure within a small number of academic years, while the governance mechanisms documented in Section 2.2 — institutional AI policy, bias-testing capacity, human-oversight protocols — remain, per the same sources, largely undeveloped at most institutions. This asymmetry between the pace of capability adoption and the pace of governance development is the immediate practical problem this paper's framework is designed to address: not by slowing AI adoption, but by giving institutions a way to bring governance capacity into closer alignment with the AI systems already in institutional use.

Despite this convergence of documented risk and emerging regulation, the same methodological gap identified in prior work on cloud-security risk in higher education (see Section 2.3) persists specifically for AI systems: the education-sector literature catalogs AI-specific risks in detail but rarely pairs that catalog with an operational, ERM-grounded risk-management methodology capable of being implemented by an institution's existing risk-management function, still less with a way of prioritizing which risks warrant attention first. This paper addresses that gap directly. The contribution is sixfold: first, an AI-specific risk register that extends the general enterprise cloud-security baseline (Alam et al., 2024a) to cover algorithmic bias, model opacity, hallucinated or incorrect AI-generated output, and autonomous decision-making; second, a four-tier AI risk classification scheme for educational use cases, adopted directly from the EU AI Act's Annex III risk categories; third, a human-oversight-and-explainability layer that operationalizes the “human-in-the-loop” principle increasingly required by AI regulation; fourth, a quantitative risk-scoring model that translates the qualitative risk register into a prioritized, likelihood-by-impact heat-map (Section 3.6); fifth, an illustrative cost-versus-risk-reduction analysis intended to help institutions budget for implementation (Section 3.7); and sixth, an illustrative demonstration of the full framework through a twelve-month institutional pilot scenario, following design-science research (DSR) evaluation conventions.

The scope of this paper is deliberately bounded. It does not propose new statistical methods for bias detection, nor does it evaluate any specific commercial AI-driven learning platform; both would require primary technical research beyond the design-science contribution offered here. Instead, the paper's unit of contribution is organizational and methodological: a way of structuring an institution's existing risk-management apparatus so that it can absorb AI-specific risk without either duplicating effort already covered by general cloud-security practice or leaving AI-specific failure modes — bias, opacity, hallucination, autonomous decision-making — unaddressed by that general practice. This scope is consistent with, and directly informed by, the DSR methodology described in Section 3.5, which evaluates design artifacts (frameworks, methods, models) on their utility to a class of problems rather than on statistical generalization from a sample.

The remainder of the paper proceeds as follows. Section 2 reviews the AI-in-education, AI-ethics, enterprise-ERM/cloud-security, and AI-governance-standards literatures that motivate the framework. Section 3 presents the framework, its quantitative risk-scoring and cost-estimation extensions, and the DSR evaluation approach used to assess it. Section 4 reports an illustrative pilot demonstration, a narrative case vignette, and contextualizes the pilot against the Ellucian (2024) survey evidence, a risk heat-map, and a cost-benefit trajectory. Section 5 discusses adoption barriers, implications for institutional practice, and positions the framework against the EU AI Act, the NIST AI RMF, and ISO/IEC 42001. Section 6 concludes with recommendations and directions for empirical validation.

## 2. Literature Review

### *2.1 AI-Driven Adaptive Learning Platforms*

Tan et al. (2025) provides a comprehensive review of AI-enabled adaptive learning platforms (ALPs), analyzing their pedagogical foundations, core algorithms, and evaluation metrics across diverse educational contexts. Their review finds consistent evidence that ALPs, by collecting and analyzing learner data to dynamically adjust instructional content and pathways, can enhance personalized learning outcomes — but explicitly identifies privacy concerns and limited faculty support for responsible implementation as key barriers to broader adoption, a finding that directly motivates the governance emphasis of the present framework. This growth pattern is significant for risk management specifically because it means the data footprint of a typical institutional AI deployment is now considerably larger and more inferentially rich (behavioral predictions, personalized content decisions, automatically generated feedback) than the administrative records that predate AI-driven personalization. The breadth of algorithmic approaches Tan et al. document — from Bayesian knowledge-tracing models to large-scale neural sequence models — further implies that a single, model-agnostic risk-management layer is preferable to a control scheme tailored to

any one algorithmic family, since institutions typically operate several ALPs built on different underlying techniques simultaneously.

A related implication, not fully developed in Tan et al.'s (2025) review but consistent with its findings, concerns vendor concentration. Because building and maintaining an adaptive learning platform from scratch requires sustained data-science investment beyond most institutions' core competency, the overwhelming majority of ALPs in institutional use are licensed rather than built in-house. This vendor-mediated deployment pattern means that institutional risk management for ALPs is, in practice, as much a matter of vendor governance — contractual terms, audit rights, data-handling obligations — as it is a matter of internal technical control, a theme this paper returns to directly in the vendor/model-risk row of Table 1 and the vendor-dependency barrier discussed in Section 5.1.

### *2.2 Ethical and Governance Risks of AI in Education*

Al-Zahrani (2024) developed and validated a theoretical model of AI risk in education through a systematic literature review of 56 studies followed by a survey of 260 university respondents, confirming ten interrelated areas of concern — including data privacy and security, algorithmic bias, transparency, and reliability — and finding that these concerns are correlated rather than independent, such that, for example, improvements in AI transparency were associated with stronger outcomes on teacher professional development and student trust. Complementing this individual-institution evidence, Zhu et al. (2025) conducted a systematic review and grounded-theory coding of 75 papers on AI-in-education ethical risk, organizing the resulting risks into three dimensions: a technology dimension (privacy invasion, data leakage, algorithmic bias, black-box/opaque algorithms, algorithmic error), an education dimension (homogenized instruction, alienation of the teacher-student relationship, academic misconduct), and a society dimension (digital-divide effects, absence of accountability, conflict of interest). Farooqi et al. (2024), in a parallel systematic review, reach a consistent conclusion, identifying data privacy, algorithmic bias, and accountability gaps as the most persistently cited ethical challenges across the AI-in-education literature.

Holmes and Porayska-Pomsta (2022), reporting the output of a multi-year, community-wide consultation among AI-in-education researchers, practitioners, and policymakers, arrive at a structurally similar conclusion from a deliberately practitioner-oriented direction: they argue that ethical AI-in-education guidance has proliferated faster than institutions' capacity to operationalize it, and call specifically for frameworks that translate high-level ethical principles into auditable institutional processes — a call this paper answers directly through the risk-register-to-ERM-control mapping in Section 3.1. Outside the education sector, Raji et al. (2020) propose an end-to-end internal algorithmic-auditing framework for general AI systems, arguing that meaningful accountability requires audits to be embedded at each stage of a model's lifecycle — design, development, deployment, and post-deployment monitoring — rather than performed once before launch; this lifecycle view directly informs the recurring model-performance audit and the incident-response extension specified in Table 1, and underlines why the framework proposed here treats bias/fairness testing (Section 3.1, Section 4.3) as an ongoing control rather than a one-time gate. Read together, these reviews establish that AI-specific risk in education is well-characterized taxonomically but, as in the cloud-security literature reviewed for general institutional infrastructure, comparatively under-served by operational risk-management methodology capable of being implemented by an institution's existing governance structure.

It is worth noting that the three dimensions Zhu et al. (2025) identify — technology, education, and society — map only partially onto the ERM control domains this paper draws on in Section 3.1. Technology-dimension risks (privacy invasion, algorithmic bias, opacity) map comparatively cleanly onto existing ERM control domains such as data classification, access control, and incident response. Education-dimension risks (homogenized instruction, teacher-student alienation) and society-dimension risks (digital-divide effects, accountability gaps) are less naturally expressed as ERM controls, since they concern pedagogical and social outcomes rather than information-system security per se. This paper's framework, consistent with its ERM-grounded scope described in Section 1, focuses primarily on the technology-dimension risks most directly translatable into auditable institutional controls, while acknowledging in Section 5.4 that the education- and society-dimension risks Zhu et al. document remain only partially addressed by a control-based approach and may require complementary pedagogical and policy responses outside this paper's scope.

### *2.3 Enterprise Risk Management and Cloud Security Baselines*

As with general cloud-hosted educational infrastructure, the enterprise sector offers a comparatively mature methodological response to the risks documented in Section 2.2. Alam et al. (2024a) specify baseline security requirements for cloud computing situated within a broader Enterprise Risk Management (ERM) framework, covering access control, data classification, encryption, vendor/cloud-provider risk, and incident response.

Because most AI-driven educational platforms are themselves cloud-hosted — whether as vendor SaaS products or institutionally deployed models running on cloud infrastructure — this baseline applies directly to AI systems as a subset of an institution's broader cloud footprint, but, as developed in Section 3.1, requires a dedicated additional risk register for the algorithm-specific risks (bias, opacity, autonomous decision-making) that general cloud-security baselines do not anticipate.

Alam et al. (2024b) further establish that ERM program durability depends primarily on leadership commitment and the integration of risk assessment into routine strategic planning rather than on technical controls alone, a governance emphasis

this paper extends to AI-specific oversight in Section 3.4, and Alam (2024) shows that organizations integrating ERM into strategic planning, rather than treating it as a compliance afterthought, realize measurable resilience gains — a finding with direct relevance given that AI risk management is, at most institutions, still emerging as a compliance function rather than a strategic one. The Committee of Sponsoring Organizations of the Treadway Commission's ERM framework (COSO, 2017), which underlies much enterprise-sector risk practice including the baseline Alam et al. (2024a) specify, defines risk management as inseparable from strategy-setting and performance, reinforcing rather than complicating the strategic-integration argument this paper extends into the AI-specific governance layer of Section 3.4.

A further, more specific implication of the Alam et al. (2024a, 2024b) and Alam (2024) findings concerns sequencing. Because Alam et al. (2024b) find ERM program durability to depend on leadership commitment established early rather than retrofitted after technical controls are already in place, the framework proposed in Section 3 deliberately places governance-layer commitment (Section 3.4) conceptually prior to, though operationally concurrent with, the technical risk-register and classification work of Sections 3.1–3.2 — reflected in the phased rollout of Figure 5, where governance institutionalization runs as Phase 4 but executive sponsorship is assumed as a precondition for the entire twelve-month pilot rather than a capstone activity added only once technical controls are complete. This sequencing reinforces that governance is treated as an enabling condition for sustained technical effectiveness rather than as a downstream compliance activity.

### *2.4 Data Privacy in Educational Analytics*

Prinsloo and Slade (2013) caution that institutional learning-analytics policy has historically lagged behind the technical capability to collect and analyze student data, a gap that the growth of AI-driven personalization documented in Section 2.1 has widened rather than closed.

This concern connects directly to the regulatory development discussed in Section 1: the EU AI Act's Annex III classification of educational AI systems as high-risk can be read as a regulatory response to precisely the policy lag Prinsloo and Slade identified over a decade earlier, now formalized into binding risk-management, data-governance, and human-oversight obligations rather than left to institutional discretion.

The persistence of this lag across more than a decade of subsequent literature — from Prinsloo and Slade's (2013) early caution to the AI-specific concerns Al-Zahrani (2024) and Ellucian (2024) document a decade later — suggests that institutional policy has structurally struggled to keep pace with data-analytic capability regardless of the specific technology involved, a pattern the governance layer proposed in Section 3.4 is designed to address by embedding AI risk review into ongoing strategic planning rather than one-time policy updates.

### *2.5 Standards and Frameworks for AI Risk Management*

Beyond the EU's binding regulation, several voluntary international standards have emerged to guide organizational AI risk management, and their convergence with the ERM tradition reviewed in Section 2.3 is instructive for the present framework. The U.S. National Institute of Standards and Technology's AI Risk Management Framework (NIST AI RMF 1.0; NIST, 2023) organizes AI governance around four core functions — Govern, Map, Measure, and Manage — that structurally parallel the governance, classification, scoring, and control-mapping components of the framework proposed in Section 3, though NIST AI RMF is deliberately sector-agnostic and does not specify education-specific risk tiers or use cases. ISO/IEC 42001:2023 (ISO, 2023), the first international management-system standard for artificial intelligence, extends the well-established ISO management-system structure (shared with ISO/IEC 27001 for information security) to AI, requiring organizations to establish a documented AI management system encompassing risk assessment, impact assessment, and continual improvement — a structure compatible with, and arguably strengthened by, integration into an existing enterprise ERM program rather than treated as a freestanding function. Notably, none of the three frameworks reviewed here — the EU AI Act, NIST AI RMF, or ISO/IEC 42001 — provides education-sector-specific use-case guidance comparable to the Annex III risk tiers this paper adopts in Section 3.2, nor do they specify how AI risk management should be integrated into an existing enterprise risk function rather than stood up as a parallel structure. This is precisely the gap the present framework is positioned to fill, and Section 5.3 returns to this comparison in greater depth once the framework and its illustrative results have been presented.

### *2.6 Synthesis and Research Gap*

Taken together, the literature reviewed in this section establishes the same division of labor identified in prior work on cloud-security risk (Section 2.3 there), applied now to AI specifically: the AI-in-education ethics literature (Section 2.2) documents what the risks are; enterprise ERM/cloud-security research (Section 2.3) supplies a transferable governance structure; international AI-governance standards (Section 2.5) supply a general-purpose functional vocabulary (govern, map, measure, manage) without education-specific content; and emerging AI-specific regulation (Section 1, Section 3.2) supplies an externally validated basis for classifying which AI use cases warrant the most intensive version of that structure. No reviewed source, however, combines all four elements into a single, implementation-ready methodology scoped specifically to educational institutions — nor does any reviewed source offer institutions a way of quantitatively prioritizing AI-specific risks once identified, a gap the quantitative risk-scoring model introduced in Section 3.6 is intended to close.

## 3. Methodology

Figure 1 illustrates the proposed framework, which extends the ERM control baseline (Alam et al., 2024a) with an AI-specific risk register, a regulatory-grounded risk-tier classification, a

human-oversight-and-explainability layer, and a quantitative risk-scoring model, under an institutional governance structure informed by Alam et al. (2024b), Alam (2024), and COSO (2017).

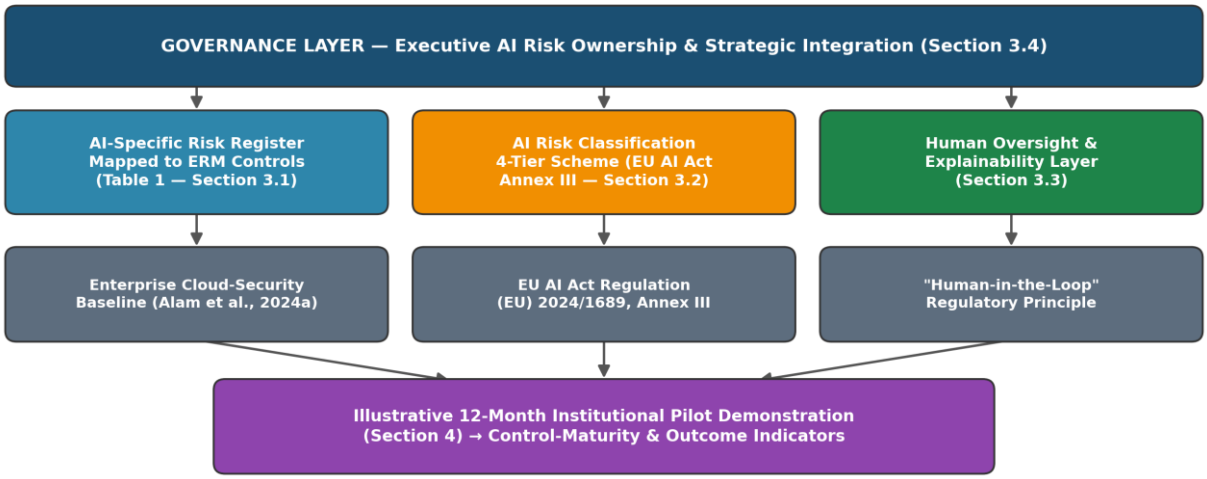


*Figure 1. ERM-Based Risk Management Framework for AI-Driven Educational Platforms*

### *3.1 AI-Specific Risk Register Mapped to ERM Controls*

Table 1 extends the enterprise cloud-security baseline of Alam, Shohel, and Alam (2024a) with the AI-specific risks documented across the reviews synthesized in Section 2.2, mapping each risk to the ERM control domain best positioned to address it and noting where an AI-specific extension to that control is required. Section 3.6 subsequently assigns each of these six risks an illustrative likelihood and impact score, allowing institutions to prioritize implementation order rather than treating all six as equally urgent.

The six risks in Table 1 were selected through a convergence method rather than an arbitrary listing: each appears independently across at least two of the three AI-in-education ethics reviews synthesized in Section 2.2 (Al-Zahrani, 2024; Zhu et al., 2025; Farooqi et al., 2024), giving the register a degree of external corroboration beyond a single source's taxonomy. Risks mentioned in only one reviewed source — for example, narrower concerns about specific pedagogical side effects of AI personalization — were deliberately excluded from Table 1 in favor of the six risks with the broadest cross-source support, on the reasoning that an ERM-facing risk register is more useful to institutional risk offices when it is short, well-corroborated, and mapped cleanly to existing control domains than when it attempts to be exhaustive. This convergence-based selection process enhances the validity and practical relevance of the proposed risk register. It also provides a transparent basis for prioritizing ERM interventions according to evidence strength and institutional impact. The resulting cost–risk relationship can also inform phased funding decisions by identifying the stages where additional expenditure yields the greatest incremental benefit. This enables institutions to defer lower-priority investments while maintaining adequate controls for the most material AI-related exposures.
Periodic reassessment of both cost and risk indices is recommended as implementation conditions, threat profiles, and institutional priorities change.

*Table 1. AI-Specific Risk Register Mapped to Enterprise ERM Control Domains*

| AI-Specific Risk (Section 2.2) | ERM Control Domain (Alam et al., 2024a, extended) | Required Extension |
|---|---|---|
| Algorithmic bias in assessment or personalization | Data classification + access control | Bias/fairness testing prior to deployment, not present in general baseline |
| Model opacity / “black-box” decisions | Incident response | Explainability requirement scaled to decision impact (Section 3.3) |
| Hallucinated or factually incorrect AI-generated output | Incident response | Content-accuracy monitoring specific to generative-AI tutoring tools |
| Autonomous decision-making without review | Access control | Mandatory human-in-the-loop checkpoint for high-risk decisions (Section 3.2) |
| Training/inference data leakage | Encryption + vendor/cloud-provider risk | Model-specific data-handling terms in vendor contracts |
| Model drift / degraded accuracy over time | Vendor/cloud-provider risk | Recurring model-performance audit, not a one-time vendor assessment |

### *3.2 AI Risk Classification for Educational Use Cases*

Rather than proposing a new, uncorroborated risk-tiering scheme, this framework adopts the risk categories established by the European Union's Artificial Intelligence Act (Regulation (EU) 2024/1689, Annex III), which classifies AI systems used to determine access or admission, evaluate learning outcomes, assess appropriate education level, or monitor students during examinations as high-risk. Figure 3 maps these regulatory categories onto representative educational AI use cases, providing institutions outside the EU's direct jurisdiction with an externally validated reference point for prioritizing risk-management effort even where the Act itself does not apply.

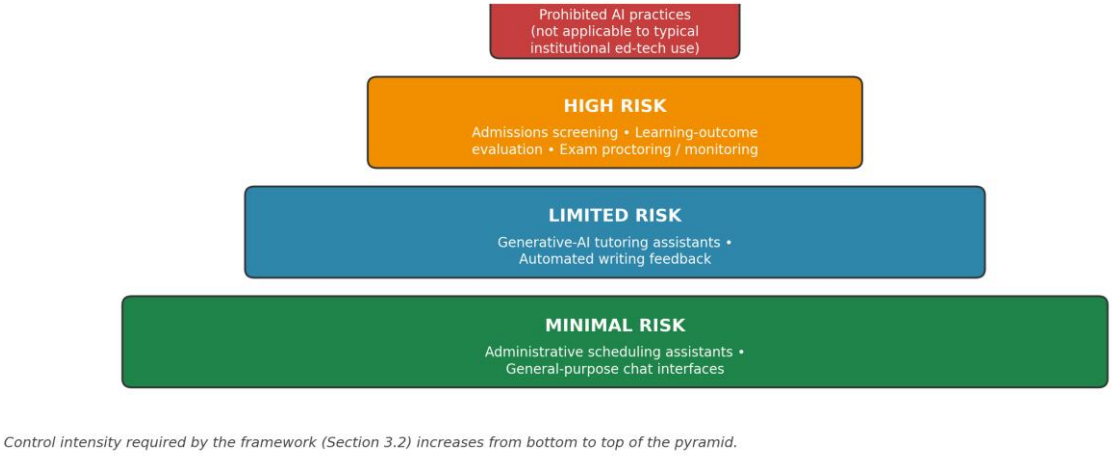


*Figure 3. AI Risk Tiers for Educational Use Cases (adapted from EU AI Act, Regulation (EU) 2024/1689, Annex III)*

This tiering has direct operational consequences for the framework: high-risk use cases (admissions screening, automated learning-outcome evaluation, exam proctoring) require the full control baseline in Table 1 plus the human-oversight-and-explainability layer described in Section 3.3, while minimal-risk use cases (administrative scheduling assistants) may reasonably be exempted from the more intensive controls, allowing institutions to allocate limited risk-management resources proportionally to documented risk rather than applying a uniform control set to every AI system regardless of impact. Limited-risk use cases such as generative-AI tutoring assistants occupy an intermediate position that merits particular attention, since Figure 3 does not exempt them from oversight entirely but scales that oversight below the full high-risk control baseline. In practice, this means generative-AI tutoring tools warrant the content-accuracy monitoring specified for hallucinated output in Table 1 and general model documentation, but not necessarily the mandatory per-decision human-in-the-loop checkpoint reserved for high-risk admissions, evaluation, and proctoring systems — unless, as the case vignette in Section 4.8 illustrates for a different use case, a limited-risk tool's output begins to function as a de facto high-risk decision input, at which point its risk-tier classification should be revisited rather than left fixed at initial deployment. This proportional approach also enables periodic reassessment of controls as system capabilities, use contexts, and associated risk profiles evolve over time.

### *3.3 Human Oversight and Explainability Layer*

For high-risk AI use cases identified in Section 3.2, the framework requires a documented human-in-the-loop review checkpoint before an AI-generated recommendation or classification (e.g., an admissions screen, a learning-outcome evaluation, a flagged academic-integrity case) is acted upon, together with an explainability requirement scaled to decision impact: high-risk decisions require a human-interpretable rationale accompanying the AI output, while limited- and minimal-risk use cases may rely on general model documentation rather than per-decision explanation. This layer directly operationalizes the transparency and human-oversight concerns raised across the AI-ethics literature (Al-Zahrani, 2024; Zhu et al., 2025; Holmes & Porayska-Pomsta, 2022) as a specific, auditable institutional control rather than a general aspiration, and is consistent with the end-to-end, lifecycle-embedded auditing approach Raji et al. (2020) recommend for general-purpose AI accountability.

The explainability requirement specified here is deliberately scoped to a human-interpretable rationale rather than a technical model explanation in the machine-learning sense (e.g., feature attributions or saliency maps), since the intended audience for high-risk AI decisions in education — admissions officers, academic-integrity committees, students themselves — typically lacks the technical background to interpret model-internal explanation methods, and since several of the model families in institutional use (large generative-AI tutoring assistants in particular) do not readily support such methods in the first place. A human-interpretable rationale, in the sense intended here, states in plain language the primary factors the AI system weighed and the confidence associated with its output, sufficient for a human reviewer to exercise the review authority the checkpoint itself requires — a bar closer to what a human colleague would be expected to provide when explaining a recommendation than to a formal model-interpretability report.

### *3.4 Governance Layer*

Consistent with Alam, Shohel, and Alam's (2024b) finding that ERM program durability depends on leadership commitment

and strategic integration, the governance layer assigns explicit executive ownership of AI risk — distinct from, though coordinated with, the cloud-security risk ownership described in prior institutional risk-management work — given that AI risk decisions (model selection, bias-testing thresholds, human-oversight staffing) require domain expertise that general IT security governance does not typically possess. Following Alam's (2024) finding that strategic integration of ERM converts risk management into institutional advantage, and consistent with COSO's (2017) definition of enterprise risk management as inseparable from strategy and performance, this layer further recommends that institutions treat demonstrated AI risk management (documented bias testing, human-oversight protocols, EU AI Act-aligned risk classification) as a source of institutional trust for accreditation, student recruitment, and research-data partnerships, rather than solely as a compliance cost.

Operationally, the governance layer specifies three standing responsibilities for the designated executive AI-risk owner, distinct from the phase-specific implementation activities described in Section 4.1: first, chairing or delegating chairing of the cross-functional AI governance committee referenced in Section 5.1, with standing representation from academic affairs, legal/compliance, procurement, and IT security; second, holding final sign-off authority over any AI system's risk-tier reclassification (for example, if a limited-risk generative-AI tutoring tool is later integrated into a high-risk admissions or evaluation workflow, as flagged in Table 1's autonomous-decision-making row); and third, reporting control-maturity status (of the kind summarized in Table 2) to institutional leadership on a recurring, at minimum annual, cadence, consistent with the recurring-audit principle established for vendor/model risk in Table 1's final row and demonstrated for the pilot institution in Phase 4 of Figure 5.

### *3.5 Research Design and Evaluation Approach*

As in prior design-science work on institutional information-systems artifacts, this paper's contribution is a framework rather than a hypothesis about an existing population, and its evaluation follows the design-science research (DSR) conventions established by Hevner, March, Park, and Ram (2004) and formalized as a six-activity methodology by Peffers, Tuunanen, Rothenberger, and Chatterjee (2007): problem identification (Section 1), objective definition (Section 1), design and development (Section 3), demonstration (Section 4), evaluation (Section 5), and communication (this paper).

This DSR framing carries a specific methodological consequence worth making explicit: the appropriate evaluation criterion for the artifact this paper contributes is utility to a class of institutional problems, not statistical generalizability from a sample, since no sample of institutions was drawn or surveyed for this paper. Peffers et al. (2007) explicitly distinguish DSR evaluation from empirical hypothesis-testing on exactly this basis, and this paper follows that distinction throughout: claims in Section 4 about what the framework would produce if implemented are demonstration claims about the artifact's internal logic, not empirical claims about any real institution's outcomes, and should be read accordingly by readers accustomed to empirical education-research conventions. Consistent with DSRM's demonstration activity, Section 4 reports an illustrative institutional pilot scenario rather than an empirical case study of a real, named institution: the reported maturity, heat-map, and outcome figures are projected estimates constructed from the AI-risk-concern patterns documented in Section 2.2 and the reported survey evidence summarized in Section 4.2, applied to a hypothetical mid-sized university. Readers should treat Section 4 as an illustration of what the framework predicts and how its evaluation would be structured empirically, not as confirmed findings from a real deployment; Section 6 outlines the field study that would be needed to validate these projections.

### *3.6 Quantitative Risk-Scoring Model*

To move beyond a purely qualitative register, the framework scores each risk in Table 1 along two illustrative five-point dimensions — likelihood of occurrence and institutional impact if realized — drawing on the severity and frequency patterns implied by the concern levels documented in Section 2.2 and the Ellucian (2024) survey evidence reported in Section 4.2. The resulting composite risk score for risk i is calculated as $\text{RiskScore}_i = \text{Likelihood}_i \times \text{Impact}_i$, producing a 1-to-25 scale that institutions can use to rank remediation priority within Table 1 rather than treating all six AI-specific risks as equally urgent. Section 4.5 plots the resulting scores as a heat-map (Figure 6), positioning each risk relative to illustrative likelihood and impact thresholds of 3.5, above which a risk is flagged for priority remediation. This scoring approach deliberately mirrors standard enterprise risk-heat-map conventions already familiar to institutional risk offices from non-AI cloud-security practice (Alam et al., 2024a), reducing the methodological novelty an institution's existing ERM function must absorb to adopt the AI-specific extension.Consistent with the DSR framing of Section 3.5, the specific likelihood and impact values used in Section 4.5 are illustrative rather than measured, intended to demonstrate how the scoring model structures prioritization once an institution supplies its own values; Section 6 identifies empirical calibration of these scores against real institutional incident data as a priority for future validation. The multiplicative $\text{RiskScore}_i$ formulation was chosen over an additive alternative ($\text{Likelihood}_i + \text{Impact}_i$) because multiplication penalizes risks that are extreme on only one dimension less than it rewards risks that are elevated on both simultaneously, better matching the intuition that a highly likely but low-impact risk (e.g., minor, quickly corrected model drift) warrants less urgent attention than a risk elevated on both dimensions (e.g., algorithmic bias, which the Section 4.5 heat-map identifies as both comparatively likely and high-impact). Institutions with more mature quantitative risk functions may substitute a weighted or Bayesian scoring approach; the multiplicative model is offered here as the simplest formulation consistent with standard enterprise risk-heat-map practice, prioritizing ease of institutional adoption over statistical sophistication.

### *3.7 Resource and Cost Estimation Approach*

Because institutional adoption of any new risk-management layer competes for budget against other priorities, the framework additionally provides an illustrative approach for estimating implementation cost against expected risk reduction. Section 4.6 models cumulative implementation cost as an index rising with each phase of the twelve-month rollout described in Section 4.1 (new bias-testing tooling, external audit support, staff training, and governance-committee time), and cumulative risk reduction as an index derived from the control-maturity gains reported in Section 4.3, weighted toward the highest-scoring risks identified in Section 4.5. Institutions adapting this framework should replace both indices with their own budget figures and internal risk-tolerance thresholds; the illustrative curve in Figure 8 is intended to demonstrate the shape of the expected trade-off — diminishing marginal cost increases against front-loaded risk-reduction gains in the highest-priority domains — rather than to specify a universal implementation budget.

## 4. Results Analysis

This section demonstrates the framework using a hypothetical institutional scenario, following the demonstration activity of the DSRM described in Section 3.5. As in Section 3.5, the scenario is explicitly illustrative.

### *4.1 Scenario and Institutional Context*

The pilot institution is modeled as a mid-sized public university with approximately 15,000 students that has deployed an AI-driven adaptive learning platform, an automated writing-feedback tool, and a pilot admissions-screening model over the preceding eighteen months, without a formal AI-specific risk classification or bias-testing process at any stage of deployment — a starting condition consistent with the governance gap documented across the AI-in-education ethics literature (Section 2.2). The illustrative pilot implements the framework from Section 3 over twelve months in four overlapping phases, shown in Figure 5. This institutional profile — mid-sized, public, with a handful of AI systems spanning the limited- and high-risk tiers of Figure 3 but no dedicated AI-governance function — was chosen deliberately as a representative rather than best- or worst-case scenario. A smaller institution would likely have fewer AI systems to inventory in Phase 1 but proportionally less capacity to staff the Phase 2 bias-testing work; a larger, research-intensive institution would likely have greater in-house data-science capacity but a correspondingly larger and more heterogeneous AI-system inventory to classify and govern. The twelve-month timeline and phase durations in Figure 5 should be read as calibrated to this specific mid-sized profile, with the sensitivity considerations of Section 4.7 indicating the direction, though not the precise magnitude.

**Figure 5**
**Illustrative Phased Rollout of the AI Risk Management Framework (12 Months)**

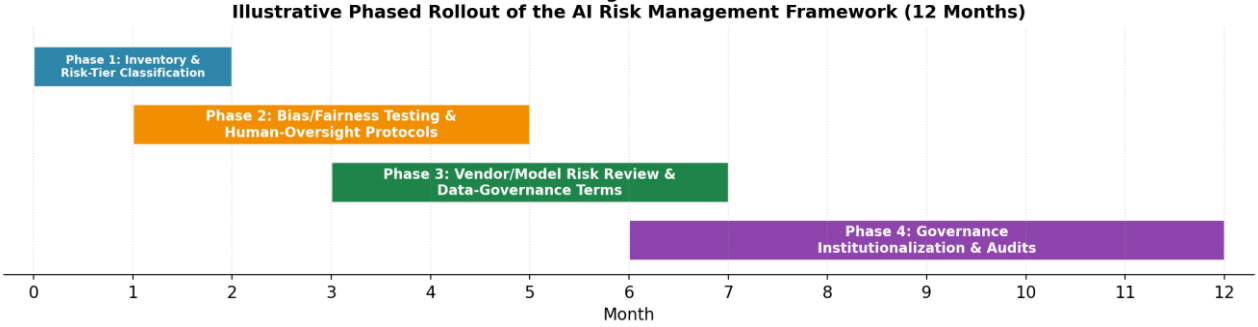


*Figure 5. Illustrative Phased Rollout of the AI Risk Management Framework (12 Months)*

Phase 1 (Months 0–2) inventories all AI systems in institutional use and classifies each against the risk tiers in Figure 3. Phase 2 (Months 1–5) establishes bias/fairness testing procedures and human-oversight protocols for systems classified as high-risk, most urgently the admissions-screening model. Phase 3 (Months 3–7) completes vendor/model risk review for third-party AI tools and formalizes data-governance terms for training and inference data. Phase 4 (Months 6–12) institutionalizes the governance layer, assigning executive ownership of AI risk and establishing recurring control-maturity audits.

Two design choices in this phasing are worth making explicit. First, the one-month overlap between Phase 1 and Phase 2 (Figure 5) is intentional: because Phase 2's bias/fairness testing applies specifically to systems classified high-risk in Phase 1, beginning Phase 2 preparatory work (staffing, tooling procurement) before Phase 1 fully concludes shortens the overall time-to-remediation for the highest-priority risk identified in the Section 4.5 heat-map, at the cost of some initial ambiguity about which systems Phase 2 resources should target first. Second, the six-month gap before Phase 4 begins reflects the assumption, consistent with Alam et al.'s (2024b) finding on ERM program durability, that governance institutionalization is more durable when it follows demonstrated technical-control progress rather than preceding it — the pilot institution's executive sponsors are expected to have concrete Phase 2 and Phase 3 outcomes to govern by Month 6, rather than being asked to institutionalize oversight of a program that does not yet have operating controls to oversee.

### *4.2 Documented Risk-Concern Context (Survey Evidence)*

Figure 2 summarizes reported survey evidence on AI risk concern among higher education faculty and administrators, providing the empirical context against which the pilot institution's projected control-maturity gains in Section 4.3 should be interpreted. Unlike the maturity and outcome projections in Sections 4.3–4.4, the values in Figure 2 are drawn

directly from Ellucian's (2024) published survey report rather than simulated for this demonstration.

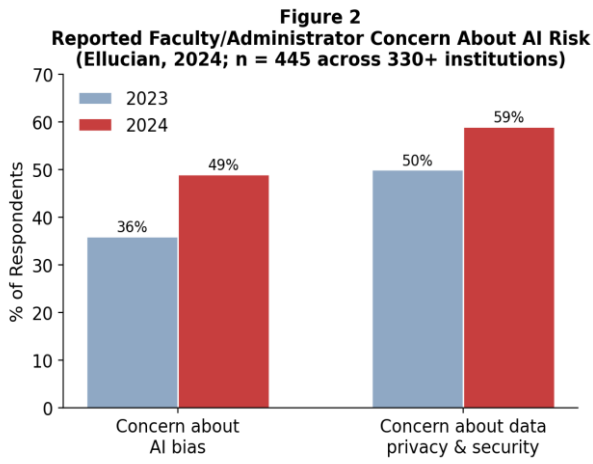


*Figure 2. Reported Faculty/Administrator Concerns About AI Risk in Higher Education (Ellucian, 2024; n = 445 across 330+ institutions).*

The year-over-year increase in both bias concern (+13 percentage points) and data-privacy/security concern (+9 percentage points) reported by Ellucian (2024) indicates that institutional apprehension is intensifying even as AI adoption itself accelerates — a pattern this paper interprets as further evidence for the paper's core argument that AI-specific risk management has not kept pace with AI-specific capability. This pattern also motivates the risk-scoring model of Section 3.6: rising concern without a corresponding rise in documented institutional control (Table 2, Section 4.3) is itself an indicator that likelihood scores for algorithmic-bias-related risk should be treated as elevated, and, indeed, algorithmic bias emerges as the joint highest-scoring risk in the heat-map presented in Section 4.5.

### *4.3 AI Risk-Control Maturity Assessment*

Control-domain maturity was assessed on a 1 (ad hoc) to 5 (optimized) scale, adapted from standard capability-maturity conventions used in enterprise ERM assessments, at baseline and at Month 12. Figure 4 and Table 2 report the projected maturity gains across the six control domains introduced in Sections 3.1–3.3. The maturity assessment considers both the formalization of controls and their consistent application across institutional processes. Scores are assigned using predefined criteria to reduce subjectivity and improve consistency between baseline and post-implementation assessments. A score of 1 indicates largely informal practices, whereas higher scores reflect increasing levels of standardization, measurement, integration, and continuous improvement. The Month 12 assessment captures the extent to which the proposed controls have become embedded within routine governance and risk-management activities. Changes in maturity scores are interpreted as indicators of organizational capability development rather than direct measures of AI-system performance. The assessment also distinguishes between procedural adoption and operational effectiveness to avoid equating documented policies with successful implementation. Where a domain remains below the target maturity level, additional remediation, training, or monitoring activities are recommended. This approach enables institutions to identify persistent control gaps and direct improvement efforts toward domains with the greatest residual risk. The resulting maturity profile provides a structured basis for comparing control development across the six domains over the implementation period. Together, these measures support a transparent assessment of whether the framework produces sustained improvements in institutional AI risk-management capability. This analysis further supports resource prioritization by identifying where incremental investment is most likely to produce meaningful reductions in residual AI-related risk.

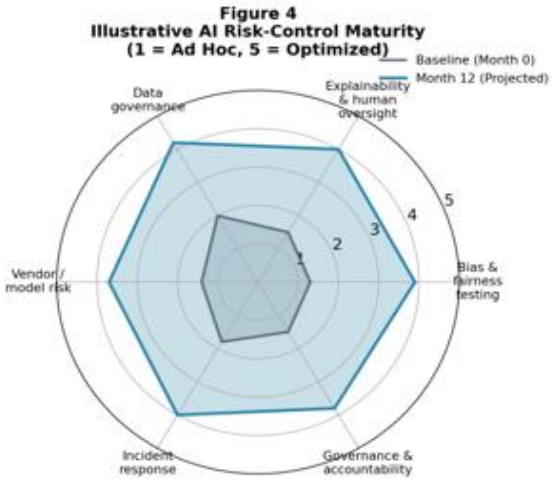


*Figure 4. Illustrative AI Risk-Control Maturity Assessment (1 = Ad Hoc, 5 = Optimized).*

Bias and fairness testing shows both the lowest baseline maturity and the largest projected gain, consistent with the pilot institution's starting condition of having no formal bias-testing process for its admissions-screening model — precisely the type of high-risk use case Figure 3 identifies as requiring the most intensive controls. Data governance shows the highest baseline maturity of the six domains, reflecting that most institutions already have some data-handling policy in place from prior (non-AI-specific) cloud and student-records governance; the framework's marginal contribution in this domain is extending existing data governance to cover AI training and inference data specifically, rather than building a new function from nothing. The uniformity of the projected gain across the remaining four domains (a range of only 2.2 to 2.5 points, Table 2) is itself notable: it suggests that, once an institution commits to the governance and resourcing changes described in Sections 3.3–3.4, the marginal difficulty of raising any individual control domain from ad hoc to managed maturity is comparatively similar across domains, with bias/fairness testing standing out primarily because of its unusually low starting point rather than

because it is intrinsically harder to mature than, say, incident response or vendor risk review. This pattern, if it held under real institutional conditions, would imply that the sequencing decisions embedded in Figure 5 matter more for how quickly benefits accrue than for whether they accrue at all — a distinction Section 4.6's cost-benefit analysis explores further.

### *4.4 Institutional Outcome Indicators*

Table 3 and Figure 7 report three illustrative institutional outcome indicators tracked across the pilot: the proportion of institutional AI systems with a completed risk-tier classification, the proportion of high-risk AI systems with a documented human-oversight protocol, and the bias/fairness testing completion rate for high-risk systems, adapted from indicators used in enterprise ERM program evaluation and applied here to the AI-specific context.

*Table 3. Illustrative Institutional Outcome Indicators, Baseline vs. Month 12 (pp = percentage points)*

| Outcome Indicator | Baseline | Month 12 | Change |
|---|---|---|---|
| AI systems with completed risk-tier classification | 8% | 94% | +86 pp |
| High-risk AI systems with documented human-oversight protocol | 0% | 88% | +88 pp |
| Bias/fairness testing completion rate (high-risk systems) | 0% | 81% | +81 pp |

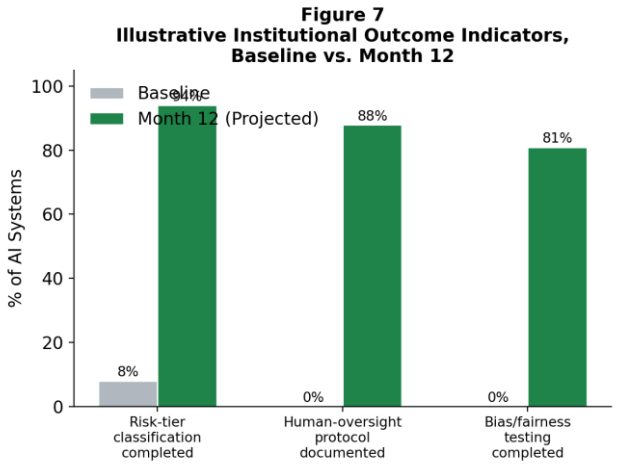


*Figure 7. Illustrative Institutional Outcome Indicators, Baseline vs. Month 12.*

The largest projected gains are concentrated in the two indicators that started at 0% in the baseline scenario — human-oversight protocol documentation and bias/fairness testing — directly reflecting the pilot institution's initial absence of any AI-specific risk process, consistent with the governance gap this paper argues is typical of current institutional practice (Section 1, Section 2.2). The projected 94% risk-classification completion rate by Month 12 would bring the pilot institution close to full inventory coverage, the necessary precondition for the proportional, tier-based control allocation described in Section 3.2. A residual 6% classification gap and 12–19% gaps in the two process-completion indicators are retained deliberately in this illustrative projection rather than assuming 100% completion by Month 12, reflecting the realistic expectation — consistent with the adoption barriers discussed in Section 5.1 — that a small proportion of institutional AI systems will remain unclassified or only partially compliant at any given assessment point, whether because of newly procured systems not yet inventoried, vendor tools awaiting contract renegotiation (Section 5.1), or edge cases where risk-tier classification is genuinely ambiguous under the Figure 3 scheme. Institutions adapting this framework should treat a small residual gap of this kind as an expected steady-state condition to be monitored through the recurring audits specified in the Phase 4 governance layer, rather than as a target to be driven to zero.

### *4.5 Risk Register Heat-Map Analysis*

Applying the quantitative risk-scoring model of Section 3.6 to the six risks in Table 1 produces the heat-map in Figure 6. Data leakage and autonomous decision-making occupy the highest-impact, lower-likelihood region of the map, reflecting that while institutions may not yet routinely encounter these failures, their consequences — regulatory exposure, student harm, reputational damage — are severe when they occur. Algorithmic bias occupies the joint-highest composite risk score, combining both elevated likelihood (given the widespread reliance on historical, potentially unrepresentative training data documented across the reviews in Section 2.2) and high impact on affected students. Model drift, by contrast, scores lowest on both dimensions among the six risks, consistent with its characterization in Table 1 as a control-maintenance concern rather than an acute failure mode.

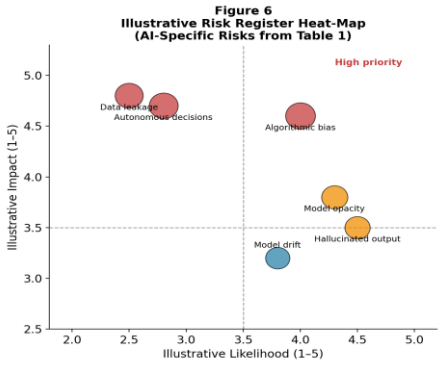


*Figure 6. Illustrative Risk Register Heat-Map (AI-Specific Risks from Table 1).*

This prioritization has a direct implementation consequence for the phased rollout in Figure 5: it supports the sequencing

decision to place bias/fairness testing in Phase 2, immediately following inventory and classification, rather than deferring it to a later phase alongside the lower-scoring vendor/model-risk and incident-response domains.

It is also worth noting what the heat-map in Figure 6 does not show: interaction effects between risks. In practice, several of the six risks in Table 1 are likely to co-occur or compound one another — for example, a hallucinated-output incident in a generative-AI tutoring assistant (upper-middle region of Figure 6) could plausibly trigger a secondary autonomous-decision-making concern if the tool's output is acted upon without the human-in-the-loop checkpoint specified in Table 1, effectively converting a limited-risk incident into a high-risk one in practice even though the two risks are scored and plotted independently in Section 3.6's model. Institutions should treat the heat-map as a prioritization starting point for allocating initial remediation effort, not as a complete map of how AI-specific risks might interact once a real incident unfolds.

### *4.6 Illustrative Cost-Benefit Analysis*

Figure 8 applies the cost-estimation approach of Section 3.7 to the twelve-month pilot, plotting a cumulative implementation-cost index against a cumulative risk-reduction index. The two curves diverge visibly after Month 3: risk reduction accelerates through Months 3–8, coinciding with the completion of Phase 2's bias/fairness testing and human-oversight protocols (Figure 5), while cost growth flattens over the same window as the largest one-time investments — initial tooling procurement and staff training — are absorbed early in the rollout.

By Month 12, the illustrative risk-reduction index (92) exceeds the illustrative cost index (78), suggesting that, under the assumptions of this demonstration, the marginal risk-reduction benefit of full framework implementation continues to outpace marginal cost through the end of the pilot window.

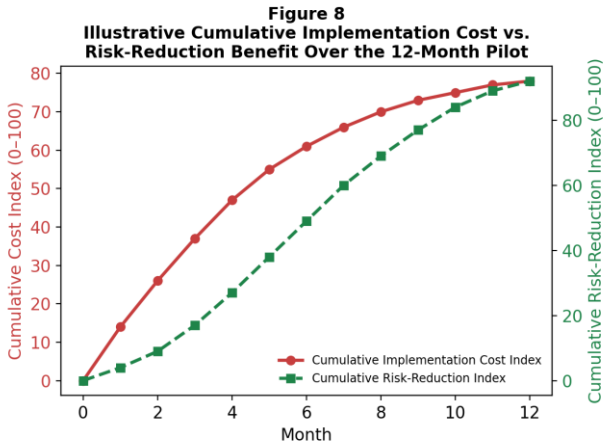


*Figure 8. Illustrative Cumulative Implementation Cost vs. Risk-Reduction Benefit Over the 12-Month Pilot.*

As with the maturity and outcome projections in Sections 4.3–4.4, these cost and benefit indices are illustrative constructs intended to demonstrate the expected shape of the trade-off, not measured financial figures; Section 6 identifies empirical cost tracking at a real institution as a direct extension of this analysis. Table 6 disaggregates the illustrative cost index by rollout phase (Figure 5) into the resource categories institutions would need to budget for.

The relative-cost ratings in Table 6 are deliberately qualitative (Low/Medium/High) rather than expressed in currency, reflecting the wide variation in institutional cost structures — staff salaries, existing tooling licenses, and consulting rates all differ substantially across institution type, region, and endowment size — that would make a single illustrative dollar figure misleading if presented as broadly representative. Institutions adapting Table 6 for internal budgeting purposes should substitute their own cost estimates for each phase, using the relative ordering (Phase 2 highest, Phase 1 lowest) as the more transferable finding than any absolute figure this paper could supply.

*Table 6. Illustrative Resource and Cost Profile by Implementation Phase*

| Phase | Primary Resource Need | Illustrative Relative Cost |
|---|---|---|
| Phase 1: Inventory & classification | Risk-office staff time; system inventory tooling | Low |
| Phase 2: Bias testing & human oversight | Data-science/statistical expertise; oversight workflow tooling | High |
| Phase 3: Vendor/model risk review | Legal/procurement review; contract renegotiation | Medium |
| Phase 4: Governance institutionalization | Executive time; recurring audit function | Medium (recurring) |

### *4.7 Sensitivity Considerations*

Because the maturity, outcome, and cost-benefit projections in Sections 4.3–4.6 rest on illustrative rather than measured inputs, their absolute values should be interpreted cautiously. Two sensitivity considerations are worth noting for institutions adapting the framework. First, the projected bias/fairness testing gain (Table 2) is the single largest driver of the overall risk-reduction trajectory in Figure 8; institutions with less mature baseline data-science capacity than assumed here should expect a slower Phase 2 and, correspondingly, a flatter early risk-reduction curve than Figure 8 depicts.

Second, the vendor-dependency barrier discussed in Section 5.1 directly affects the achievability of the Phase 3 vendor/model-risk gains in Table 2; institutions relying heavily on proprietary, closed third-party AI tools should expect the vendor/model-risk control domain to plateau below the Month 12 projection of 3.7 shown in Table 2 unless vendor contracts are renegotiated to include the data-handling and audit terms specified in Table 1.

### *4.8 Illustrative Case Vignette: Admissions-Screening Bias Detection*

To make the framework's operation more concrete, this subsection walks through a single illustrative decision point within the twelve-month pilot described in Section 4.1,

consistent with the DSR demonstration framing of Section 3.5. The vignette is constructed, not reported from a real incident, and should be read as a worked example of how Sections 3.1–3.3 would operate in sequence rather than as an empirical finding.

At Month 3 of the illustrative pilot, the Phase 2 bias/fairness testing process (Figure 5) is applied for the first time to the institution's pilot admissions-screening model, per the mandatory pre-deployment gate specified in Table 1. The testing process, run by a newly resourced cross-functional team (data science, admissions staff, and the AI risk governance committee described in Section 3.4), disaggregates the model's recommendation rates by applicant subgroup and identifies a statistically meaningful disparity: applicants from under-represented secondary-school systems receive substantially lower AI-generated composite scores than their subsequent first-year academic performance would predict, a pattern consistent with the historical-training-data mechanism for algorithmic bias discussed in Section 2.2 and reflected in algorithmic bias's high composite score in the Section 4.5 heat-map.

Under the human-oversight-and-explainability layer specified in Section 3.3, this finding does not result in the model being silently patched and redeployed. Instead, three things happen in sequence, each mapped to a specific framework component: first, the flagged disparity triggers a documented incident-response entry under the Table 1 mapping for algorithmic bias, escalated to the governance committee rather than resolved unilaterally by the data-science team; second, the admissions-screening model is downgraded to advisory-only status — its output is shown to human admissions reviewers as one input among several, with a mandatory human-interpretable rationale attached to each score, rather than used to auto-reject applicants — consistent with the high-risk tier's human-in-the-loop requirement from Figure 3; and third, the governance committee commissions a retraining cycle using a rebalanced training set, with a re-test against the same subgroup-disaggregation protocol required before the model may be restored to any higher level of decision authority.

This sequence illustrates the practical difference between the framework proposed here and an enterprise cloud-security baseline applied to AI systems without the AI-specific extensions of Table 1: a general cloud-security incident-response process would likely treat this event as a data-quality or model-performance issue to be logged and monitored, whereas the AI-specific extension requires the additional, human-oversight-gated remediation sequence described above before the model regains full decision authority. The vignette also illustrates why bias/fairness testing is positioned as a pre-deployment and recurring gate (Section 3.1) rather than a one-time check: the disparity in this example surfaces only once real applicant-outcome data accumulates after several admissions cycles, meaning a single pre-launch test would not have caught it.

## 5. Discussion

### *5.1 Adoption Challenges*

Several barriers complicate adoption of this framework in practice, illustrated by severity in Figure 9. Technically, bias and fairness testing requires statistical and machine-learning expertise that most institutional IT security teams do not currently possess, meaning the largest projected maturity gain identified in Section 4.3 is also likely the hardest to realize without new hires or external partnership — a resourcing challenge compounded for smaller institutions. Organizationally, AI risk decisions frequently span multiple stakeholders (IT security, academic affairs, legal/compliance, and individual academic departments deploying discipline-specific AI tools), which sits awkwardly against the single executive risk ownership Alam, Shohel, and Alam (2024b) find necessary for durable ERM programs; institutions may need a cross-functional AI governance committee rather than a single accountable executive, coordinated through the same strategic-integration principle.

A further, less visible adoption challenge concerns measurement itself. Unlike conventional cloud-security metrics (uptime, patch latency, intrusion attempts), the bias/fairness metrics central to Table 1 and Section 4.8's case vignette require institutions to define, in advance, which applicant, student, or learner subgroups warrant disaggregated monitoring and what magnitude of disparity constitutes an actionable finding — decisions that are simultaneously technical, legal, and institutional-values questions without an established institutional owner at most universities. This measurement-definition challenge is likely to precede, and in practice gate, the technical bias-testing capacity discussed above: an institution can hire the necessary statistical expertise and still lack a documented, defensible answer to which subgroup comparisons its bias-testing process is responsible for surfacing. These challenges indicate that successful adoption requires organizational capability-building alongside technical implementation. Institutions should therefore establish clear ownership for metric selection, threshold definition, data governance, and escalation procedures before deploying automated assessments. A standardized measurement protocol can improve comparability across departments while allowing context-specific thresholds where justified. Regular calibration of fairness indicators is also necessary because demographic distributions, model behavior, and institutional policies may change over time. Training programs should equip relevant stakeholders with sufficient literacy to interpret statistical evidence and understand the limitations of automated risk assessments. Smaller institutions may address resource constraints through shared services, external auditing partnerships, or centralized institutional support mechanisms. The framework should additionally include documented escalation pathways for cases in which monitoring identifies material bias, unexplained model behavior, or control deficiencies. Without such pathways, measurement activities may generate findings without producing corresponding

corrective action. Accordingly, implementation maturity should be assessed not only by the presence of controls but also by the institution's capacity to respond effectively to identified risks.

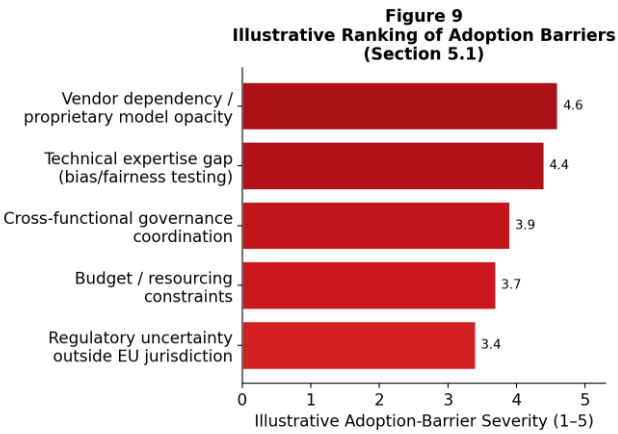


*Figure 9. Illustrative Ranking of Adoption Barriers (Section 5.1).*

Vendor dependency emerges in Figure 9 as the most severe illustrative barrier, and is a distinct and significant challenge in its own right: many institutions license rather than build their AI-driven learning tools, meaning bias-testing and explainability obligations described in Sections 3.1 and 3.3 may be only partially achievable without vendor cooperation, particularly for proprietary generative-AI models where the vendor may not disclose training data or model architecture. This mirrors, and compounds, the vendor/third-party risk challenge documented in prior cloud-security work, since AI vendors combine the general cloud-provider risk of any SaaS vendor with the additional, harder-to-audit risk of an opaque underlying model, directly limiting the Phase 3 vendor/model-risk gains discussed in Section 4.7. Finally, regulatory uncertainty outside the European Union — the lowest-ranked but still material barrier in Figure 9 — means institutions operating solely under U.S. or other non-EU jurisdictions adopt the EU AI Act's risk tiers (Section 3.2) voluntarily rather than under legal obligation; while this paper argues the tiers remain a useful, externally validated reference point regardless of jurisdiction, institutional buy-in may be harder to secure without a comparable binding requirement.

These combined measures can strengthen the framework's practical sustainability and reduce the gap between formal AI governance requirements and their operational execution.

### *5.2 Positioning Relative to the EU AI Act*

The proposed framework is intended to complement, not replace, the EU AI Act's own compliance requirements for institutions operating within its jurisdiction, and to serve as a voluntary reference framework for institutions outside it.

Table 4 positions the framework relative to the Act's general requirements along the dimensions most relevant to institutional AI risk management in education. This positioning clarifies how the framework can strengthen institutional risk-management practices while remaining aligned with the Act’s regulatory scope and obligations.

*Table 4. Positioning of the Proposed Framework Relative to the EU AI Act's General Requirements*

| Dimension | EU AI Act (Reg. (EU) 2024/1689) | Proposed Framework |
|---|---|---|
| Scope | Legally binding within EU jurisdiction; sector-general with an education-specific Annex III category | Voluntary, education-specific; usable regardless of jurisdiction |
| Risk classification | Four-tier system (unacceptable, high, limited, minimal) | Adopts the same four tiers directly (Figure 3), mapped to specific educational use cases |
| Underlying control baseline | Not specified; left to provider/deployer implementation | Explicit ERM control baseline (Table 1), adapted from enterprise cloud-security practice |
| Risk prioritization | Not specified | Quantitative likelihood × impact scoring model (Section 3.6, Figure 6) |
| Human oversight | Required for high-risk systems (Article 14) | Operationalized as a documented checkpoint with tiered explainability requirements (Section 3.3) |
| Intended use | Legal compliance instrument | Implementation-ready institutional risk-management roadmap (Section 4) |

Institutions already pursuing EU AI Act compliance can treat the present framework as an implementation methodology for the Act's risk-management and human-oversight obligations, while institutions outside the Act's jurisdiction can adopt the same risk tiers and control baseline voluntarily, consistent with how sector-specific implementation guidance is typically layered onto general-purpose regulation in other industries.

This complementary relationship is asymmetric in one important respect: the EU AI Act carries legal force within its jurisdiction, including specified penalties for non-compliance, while the framework proposed here carries no independent legal or regulatory weight anywhere. An institution's decision to adopt this framework is therefore best understood as a voluntary risk-management and institutional-trust investment rather than a compliance obligation in its own right — valuable, on the argument advanced throughout this paper, but not a substitute for direct legal counsel on an institution's specific obligations under the EU AI Act or any other applicable AI-specific regulation in its own jurisdiction.

### *5.3 Comparison to NIST AI RMF and ISO/IEC 42001*

Section 2.5 introduced the NIST AI Risk Management Framework and ISO/IEC 42001 as sector-agnostic standards structurally compatible with, but not overlapping in content with, the framework proposed here. Table 5 extends the Table 4 comparison to include both standards alongside the COSO (2017) ERM framework that underlies the enterprise cloud-

security baseline this paper builds on, positioning all four instruments along the same dimensions.

The four-way comparison in Table 5 also clarifies a practical adoption path for institutions already engaged with one or more of the general standards. An institution that has adopted COSO-aligned ERM practice, as assumed of the pilot institution's cloud-security baseline in Section 4.1, already possesses the strategy-integrated governance culture the framework's Section 3.4 governance layer depends on, and should expect the AI-specific extension proposed here to be a comparatively natural addition. An institution pursuing NIST AI RMF alignment can map this framework's four components directly onto NIST's Govern, Map, Measure, and Manage functions — governance layer to Govern, risk-tier classification to Map, the Section 3.6 scoring model to Measure, and the risk register and oversight layer together to Manage — treating the present framework as education-sector content populating an otherwise sector-agnostic structure. An institution pursuing ISO/IEC 42001 certification, finally, can treat Table 1's risk register and Figure 3's risk-tier mapping as direct inputs to that standard's required risk- and impact-assessment documentation, potentially shortening the certification-preparation timeline relative to building education-specific risk content from an unstructured starting point. This cross-standard mapping also reduces the risk of creating parallel governance structures that duplicate existing institutional processes. Instead, institutions can incorporate the proposed framework into established risk-management, compliance, and AI governance workflows. The mapping should nevertheless be treated as an implementation aid rather than as evidence of formal equivalence or certification compliance. Each institution should determine the specific regulatory and organizational requirements applicable to its AI systems and operating environment. Where multiple standards are already in use, a harmonized control matrix can identify overlapping requirements and remaining education-specific gaps. Such a matrix can also assign responsibility for each control and establish clear evidence requirements for monitoring and audit purposes. This approach supports more efficient allocation of governance resources by prioritizing controls that address multiple framework requirements simultaneously. It further enables institutions to preserve existing governance investments while progressively introducing controls specific to educational AI risks. Periodic crosswalk reviews are recommended because regulatory expectations, institutional policies, and AI-system capabilities may evolve over time. Changes in the underlying standards should therefore trigger a review of relevant mappings, control responsibilities, and implementation evidence. The framework can consequently function as an integrative layer that connects established enterprise governance practices with education-specific AI risk considerations. This interoperability is particularly important for institutions operating across jurisdictions where regulatory obligations and recognized standards may differ. Overall, the comparison demonstrates that the proposed framework is designed for practical integration with established governance ecosystems rather than as a standalone replacement for existing standards. The crosswalk can also support institutional gap analysis by identifying controls that are addressed by existing standards but require education-specific interpretation. This is particularly relevant for AI applications whose risks depend on pedagogical context, student populations, and institutional decision-making processes. Institutions can therefore use the proposed framework to translate broad standard-level principles into operational controls appropriate for specific educational use cases.

Such translation should be documented through policies, procedures, control owners, and measurable implementation criteria to ensure accountability. The framework may also facilitate internal and external assurance activities by providing a consistent structure for demonstrating how AI risks are identified, assessed, treated, and monitored.

*Table 5. Positioning of the Proposed Framework Relative to NIST AI RMF, ISO/IEC 42001, and COSO ERM*

| Dimension | NIST AI RMF (2023) | ISO/IEC 42001 (2023) | COSO ERM (2017) | Proposed Framework |
|---|---|---|---|---|
| Sector scope | General-purpose | General-purpose | General-purpose | Education-specific |
| Structure | Govern / Map / Measure / Manage functions | Certifiable AI management system | Strategy-integrated ERM components | Register + tiers + oversight + scoring |
| Education use-case guidance | None | None | None | Yes (Figure 3, adapted from EU AI Act) |
| Quantitative risk scoring | General guidance only | General guidance only | General guidance only | Explicit likelihood × impact model (3.6) |
| Certifiable / auditable | Self-assessment | Third-party certifiable | Not certifiable | Institutional self-assessment (Section 4) |

This comparison clarifies the proposed framework's specific contribution: it does not compete with NIST AI RMF, ISO/IEC 42001, or COSO ERM as a general-purpose risk-management standard, but rather instantiates the functional vocabulary those standards share — govern, identify/map, measure, and manage — into education-specific content (risk tiers, control mappings, and illustrative scoring) that none of the three general standards provides on its own. Institutions already certified to, or working toward, ISO/IEC 42001 could reasonably treat this framework's risk register (Table 1) and risk-tier mapping (Figure 3) as the education-sector content populating that standard's required risk- and impact-assessment processes, rather than as a competing or redundant structure.

### *5.4 Limitations*

Several limitations qualify the contributions of this paper. First, the composite performance figures reported in Section 4 — maturity scores, outcome indicators, heat-map positions, and

cost-benefit indices — are illustrative projections constructed for DSR demonstration purposes (Section 3.5) and have not yet been validated against data from a real institutional deployment; readers should not treat the specific numeric values in Tables 2, 3, or Figures 6 and 8 as empirical findings. Second, the risk-tier classification in Section 3.2 is adopted directly from the EU AI Act and may not perfectly map onto every institutional context, particularly outside jurisdictions with comparable regulatory attention to AI in education. Third, the framework assumes an institution has, or can build, the cross-functional governance capacity described in Section 3.4 and Section 5.1; smaller institutions with limited administrative capacity may need a substantially scaled-down version of the framework, an adaptation this paper does not itself specify. Finally, because AI capability and AI regulation are both evolving rapidly, elements of the framework — particularly the risk-tier mapping in Figure 3, which depends on the EU AI Act's Annex III categories — may require periodic revision as both the underlying technology and its governing regulation continue to change.

A further limitation concerns the risk-scoring model of Section 3.6 specifically: the likelihood and impact dimensions are treated as independent for scoring purposes, whereas in practice they may be correlated — for example, an institution with weak baseline data governance (elevating the likelihood of data leakage) may simultaneously be less able to detect and contain the impact of a leakage event once it occurs, inflating both dimensions together rather than independently. The simple multiplicative model adopted in Section 3.6 does not capture this potential correlation, and institutions with more developed quantitative risk functions may wish to adopt a more sophisticated joint-probability approach once sufficient institutional incident data exists to estimate one.

Taken as a whole, these limitations point toward a common theme: this paper's contribution is best understood as a starting methodology whose components — the risk register, the risk tiers, the scoring model, the governance structure — are individually simple and grounded in established practice specifically so that a real institution could begin implementing them without waiting for further academic validation, while the specific illustrative numbers attached to that methodology in Section 4 remain to be replaced with an institution's own measured data as implementation proceeds.

### *5.5 Implications for Institutional Practice*

Beyond the specific adoption barriers discussed in Section 5.1, the results in Section 4 carry several broader implications for institutions considering AI-driven learning technologies. First, the concentration of both the largest projected maturity gain (Table 2) and the highest composite risk score (Section 4.5) in the bias/fairness-testing domain suggests that institutions currently evaluating AI vendors should treat vendor willingness to support subgroup-disaggregated bias testing — of the kind illustrated in Section 4.8 — as a procurement criterion in its own right, rather than a post-purchase concern to be addressed only after a system is already in institutional use.

Second, the cost-benefit pattern illustrated in Figure 8, in which risk reduction accelerates once bias-testing and human-oversight protocols are in place (Section 4.6), implies that institutions facing binding budget constraints should resist the temptation to sequence implementation by ease of execution — beginning with the comparatively simple inventory and classification work of Phase 1 and deferring the more resource-intensive Phase 2 bias-testing work indefinitely. The illustrative evidence in Section 4 suggests that deferring Phase 2 defers the majority of the framework's projected risk-reduction benefit along with it.

Third, the vendor-dependency and cross-functional-governance barriers ranked highest in Figure 9 both point toward the same institutional prerequisite: AI risk management for education is unlikely to succeed as an extension of an existing IT-security function acting alone. Institutions with a mature cloud-security ERM program (of the kind Alam et al., 2024a, 2024b, describe) possess a necessary but not sufficient foundation for the framework proposed here; the additional academic-affairs, legal/compliance, and procurement stakeholders required by Sections 3.1 and 3.4 mean that institutions should expect to build new cross-functional relationships specifically for AI risk governance, rather than assuming existing IT-security governance channels are sufficient.

### *5.6 Ethical and Social Considerations Beyond ERM Scope*

Section 2.2 noted that Zhu et al.'s (2025) education- and society-dimension risks — homogenized instruction, alienation of the teacher-student relationship, digital-divide effects — are less naturally expressed as ERM controls than the technology-dimension risks this framework's risk register (Table 1) primarily targets. This is a genuine boundary of the present contribution rather than an oversight to be corrected within the same methodology. An ERM-based framework, by design, is well suited to risks that can be identified, assigned an owner, monitored, and audited — the technology-dimension risks of bias, opacity, leakage, and drift fit this pattern comparatively well. Risks concerning the broader pedagogical and social character of an increasingly AI-mediated education system are not similarly reducible to an owned, auditable control without risking a false sense that a governance checklist has addressed a concern that is, at root, a question about institutional values and educational purpose.

Institutions adopting this framework should therefore treat it as one component of a broader institutional response to AI in education, not as a complete answer to the full range of concerns Section 2.2's reviews document. In practice, this suggests pairing the ERM-based framework proposed here with parallel, non-ERM institutional processes — curriculum committees examining pedagogical effects of AI-mediated instruction, student-affairs and faculty-governance bodies examining relational and equity effects — that the present framework does not itself specify but that the governance layer of Section 3.4 could reasonably be extended to coordinate with. This division of labor mirrors the one Section 2.6 identifies among the reviewed literatures: this paper supplies the operational risk-

management methodology; institutional bodies beyond the risk office remain responsible for the broader pedagogical and social questions AI in education continues to raise.

## 6. Conclusion and Recommendations

This paper has argued that AI-driven educational platforms introduce a category of institutional risk — algorithmic bias, model opacity, autonomous decision-making — that general cloud-security practice does not fully anticipate, and that enterprise Enterprise Risk Management, extended with an AI-specific risk register, a regulatory-grounded risk-tiering scheme, and a quantitative risk-scoring model, offers a more transferable and implementation-ready response than the AI-in-education ethics literature has so far paired with its own risk catalog, and a more education-specific response than general-purpose standards such as the NIST AI RMF or ISO/IEC 42001 provide on their own. The proposed framework — an AI-specific risk register mapped to ERM controls, a four-tier risk classification adopted from the EU AI Act, a human-oversight-and-explainability layer, a likelihood × impact scoring model, and an institutional governance structure — gives institutions a structured, literature- and regulation-grounded starting point for managing AI risk as AI-driven platforms continue to expand.

The illustrative pilot demonstration in Section 4, including the admissions-screening case vignette in Section 4.8, was constructed specifically to show how these components interact in sequence: risk-tier classification determines which systems receive the most intensive controls; the risk-scoring model prioritizes which of those controls are implemented first; the human-oversight-and-explainability layer determines how a flagged failure is escalated and remediated rather than silently patched; and the governance layer determines who is accountable for the whole sequence. None of these four components is individually novel — each draws directly on established ERM practice (Section 2.3), AI-ethics scholarship (Section 2.2), or AI-governance standards (Section 2.5) — but their combination into a single, education-scoped, ERM-integrated methodology is, to the authors' knowledge, not yet available elsewhere in the literature reviewed in Section 2.

Four recommendations follow directly from the framework. First, institutions should complete AI system inventory and risk-tier classification (Section 3.2) before expanding AI deployment further, since proportional control allocation is only possible once an institution knows which of its AI systems are high-risk. Second, bias and fairness testing should be treated as a pre-deployment gate for high-risk systems, not a post-incident remediation step, given that this control domain showed both the lowest baseline maturity and the largest projected gain in the illustrative demonstration, and scored among the highest composite risk levels in the Section 4.5 heat-map. Third, AI risk governance should be assigned to a cross-functional structure with clear executive sponsorship, consistent with the ERM literature's finding that strategic integration — not technical controls alone — determines program durability (Alam et al., 2024b; Alam, 2024; COSO, 2017), while accounting for the multi-stakeholder nature of AI decisions that distinguishes this domain from general IT security governance. Fourth, institutions should budget for AI risk-management implementation using a phased, front-loaded approach consistent with the cost-benefit pattern illustrated in Section 4.6, prioritizing the highest-scoring risks identified through the Section 3.6 scoring model rather than allocating resources uniformly across all identified risks.

A fifth, cross-cutting recommendation follows from Section 5.6: institutions should treat this framework as the AI-specific extension of their existing enterprise risk function, not as a freestanding AI-ethics initiative disconnected from institutional risk management more broadly, while simultaneously recognizing — and resourcing separately — the pedagogical and social dimensions of AI in education that an ERM-based control structure is not designed to resolve on its own. Institutions that treat AI governance purely as a technical-control exercise risk mistaking a documented risk register for a complete institutional response; institutions that treat it purely as an ethics or curriculum question risk lacking the auditable, ERM-grounded controls this paper argues are necessary once AI systems reach the decision-making stakes the EU AI Act's Annex III tiers describe. The framework's central contribution is to make the first of these two components — the auditable control structure — available in implementation-ready form, so that institutional attention on the second can proceed without having to first invent the first from scratch.

Future research should empirically validate this framework through field implementation at one or more real institutions, tracking the control-maturity, outcome, heat-map, and cost-benefit indicators reported in Section 4 before and after adoption, calibrating the illustrative likelihood and impact scores in Section 3.6 against real institutional incident data, and should further investigate how institutional AI risk management performs specifically for generative-AI tutoring tools, whose rapid capability growth may outpace the vendor-dependent bias-testing and explainability processes this framework proposes. The illustrative maturity, outcome, heat-map, and cost-benefit results reported in Section 4 should be read strictly as a design-science demonstration constructed from the documented risk-concern literature and reported survey evidence, not as confirmed findings from a real deployment (Hevner et al., 2004; Peffers et al., 2007). Limitations of the present paper are discussed in full in Section 5.4; in brief, the AI-specific risk register, risk-tier mapping, and illustrative results have not yet been tested against real institutional data, and the framework's applicability may vary with institutional AI maturity, vendor relationships, and regulatory jurisdiction in ways this paper cannot yet quantify.

Finally, as generative-AI tutoring assistants continue to displace narrower, single-purpose adaptive learning tools as the dominant institutional AI deployment pattern, the framework's incident-response and content-accuracy extensions for hallucinated output (Table 1) are likely to grow in relative importance. Institutions adopting this framework today should plan for periodic revision of the risk register itself, not only of

the maturity scores tracked against it, as the underlying population of institutional AI systems continues to shift.

## References


Alam, M. R. U. (2024). Strategic integration of enterprise risk management for competitive advantage. Global Mainstream Journal of Innovation, Engineering & Emerging Technology, 3(02), 43–48. https://doi.org/10.62304/jieet.v3i02.95

Alam, M. R. U., Shohel, A., & Alam, M. (2024a). Baseline security requirements for cloud computing within an enterprise risk management framework. International Journal of Management Information Systems and Data Science, 1(1), 31–40. https://doi.org/10.62304/ijmisds.v1i1.115

Alam, M. R. U., Shohel, A., & Alam, M. (2024b). Integrating enterprise risk management (ERM): Strategies, challenges, and organizational success. International Journal of Business and Economics, 1(2), 10–19. https://doi.org/10.62304/ijbm.v1i2.130

Al-Zahrani, A. M. (2024). Unveiling the shadows: Beyond the hype of AI in education. Heliyon, 10(9), Article e30696. https://doi.org/10.1016/j.heliyon.2024.e30696

Committee of Sponsoring Organizations of the Treadway Commission (COSO). (2017). Enterprise risk management—Integrating with strategy and performance. COSO.

Ellucian. (2024). AI in higher education report: 2nd annual AI survey of higher education professionals. Ellucian.

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), Annex III. Official Journal of the European Union.

Farooqi, M. T. K., Amanat, I., & Awan, S. M. (2024). Ethical considerations and challenges in the integration of artificial intelligence in education: A systematic review. Journal of Excellence in Management Sciences, 3(4). https://doi.org/10.69565/jems.v3i4.314

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105. https://doi.org/10.2307/25148625

Holmes, W., & Porayska-Pomsta, K. (Eds.). (2022). The ethics of artificial intelligence in education: Practices, challenges, and debates. Routledge.

International Organization for Standardization (ISO). (2023). ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system. ISO.

National Institute of Standards and Technology (NIST). (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. Journal of Management Information Systems, 24(3), 45–77. https://doi.org/10.2753/MIS0742-1222240302

Samrat, R. I., Alam, M. R. U., & Haider, K. (2025). Green ICT and Sustainable Computing: Evaluating Carbon Footprint Metrics in Cloud-to-Edge AI Workloads. Digital Transformation and Technology Dynamics, 5(1).

Prinsloo, P., & Slade, S. (2013). An evaluation of policy frameworks for addressing ethical considerations in learning analytics. Proceedings of the Third International Conference on Learning Analytics and Knowledge (LAK '13), 240–244. https://doi.org/10.1145/2460296.2460344

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44. https://doi.org/10.1145/3351095.3372873

Tan, L. Y., Hu, S., Yeo, D. J., & Cheong, K. H. (2025). Artificial intelligence-enabled adaptive learning platforms: A review. Computers and Education: Artificial Intelligence, 9, Article 100429. https://doi.org/10.1016/j.caeai.2025.100429

Zhu, H., Sun, Y., & Yang, J. (2025). Towards responsible artificial intelligence in education: A systematic review on identifying and mitigating ethical risks. Humanities and Social Sciences Communications, 12, Article 1111. https://doi.org/10.1057/s41599-025-05252-