Human Systems Engineering and Design (IHSED 2026), Vol. 237, 2026, 402–410
https://doi.org/10.54941/ahfe1008179

An Integrated Governance Model for University AI: Extending ISO/IEC 27001 for Higher Education

Jungbauer Christoph and Geresics-Földi Eszter

Ferdinand Porsche FernFH – University of Applied Sciences, Austria

ABSTRACT

Generative Artificial Intelligence has become part of everyday practice in higher education. It supports learning and teaching, but it also enables new forms of academic misconduct, as students can use large language models to bypass assessment rules and to produce outputs that are difficult to attribute to individual performance. In parallel, universities face AI-related information security risks such as sensitive data disclosure and intellectual-property leakage. ISO/IEC 27001:2022 provides a well-established baseline for governing confidentiality, integrity, and availability, but it does not explicitly address AI-specific risk sources such as training-data dependencies, prompt-based attacks, non-deterministic outputs, model drift, and limited explainability. This paper develops an integrated AI governance model for higher education using Design Science Research (DSR). The artefact extends an ISO/IEC 27001-based ISMS by integrating AI lifecycle risk perspectives from ISO/IEC 23894 and AI management system integration concepts from ISO/IEC 42001, AITRiSM and the NIST AI Risk Management Framework. Because AI systems and misuse patterns evolve rapidly, the model is using an Observe–Orient–Decide–Act (OODA) loop to ensure the possibility of short, evidence-based adaptation. The proposed model is evaluated qualitatively through the conformance to the standards and real-world academic use cases that illustrate AI-enabled misconduct in artefacts.

Keywords: Higher education, AI governance, ISO/IEC 27001, Academic misconduct, OODA loop

INTRODUCTION

Artificial Intelligence (AI), and in particular generative AI, has become part of everyday practice in higher education. It supports personalised learning and enables new pedagogical approaches. At the same time, it introduces a new integrity problem: students can use generative models to bypass assessment rules and to produce outputs that are difficult to attribute to individual performance. Large Language Models (LLMs) are therefore no longer peripheral tools, but increasingly embedded in academic workflows and, indirectly, in learning outcomes and assessment decisions. In parallel, industry analyses report that more than 70% of organisations deploy generative AI for operational tasks, which highlights recurring concerns such as sensitive information disclosure, intellectual-property leakage,

Received March 10, 2026 Revised June 23, 2026 Accepted July 17, 2026 Available online September 25, 2026

and misuse of AI-driven workflows (Enterprise Cybersecurity Solutions, Services & Training, 2025). These patterns are transferable to universities, where learner-generated content, student data, and institutional intellectual property represent sensitive assets.

Current Information Security Management Systems (ISMS), particularly those based on ISO/IEC 27001:2022, provide a well-established framework for governing confidentiality, integrity, and availability (CIA). However, they do not explicitly account for AI-specific challenges, including:

- quality assurance of training data,
- machine-learning threats and vulnerability landscapes,
- model drift, hallucination, and instability of generative outputs,
- opacity and explainability limitations that hinder accountability.

In the context of higher education, these challenges provide a potential vector for cheating, plagiarism, or unauthorised assistance during assessments. Traditional ISMS frameworks insufficiently address these hybrid risks.

The AI Trust, Risk and Security Management (AI TRiSM) framework developed by Habbal et al. (2024) provides a structured triad that offers a foundation for enhancing AI governance in the educational domain. This paper examines the applicability of AI TRiSM to academic environments and proposes an extension to ISO/IEC 27001 that integrates the whole AI lifecycle, as AI exceed the scope of conventional IT risks.

BACKGROUND AND RELATED WORK

In the last few years, generative AI tools have become normal for many students. They use them to write or improve texts, to summarise articles, to generate code, or to writing emails, or for "cleaning" texts for plagiarism detectors. Some universities offer their own AI tools, sometimes based on Retrieval Augmented Generation (RAG), which answer topic specific questions.

On the positive side, this can help students to learn faster and get more feedback. However, there is also a downside. AI tools can make it much easier to cheat or to bypass assessment rules. Typical examples include:

- letting a language model write large parts of an assignment,
- asking an AI system to solve exam questions directly,
- using paraphrasing to avoid plagiarism detection,
- pasting internal slides, exam questions, or feedback into public AI tools,
- trying to gain information from an internal RAG system.

Industry reports on AI data security show similar patterns in companies: people paste sensitive content into external AI tools or misuse internal AI assistants in ways that were not planned by the organisation (Enterprise Cybersecurity Solutions, Services & Training, 2025). In a university, the same behaviour can lead to loss of intellectual property or leakage of student data. Because of this AI is also a risk management and governance topic.

ISO/IEC 23894:2023 is a standard that explains how to manage risks that come from AI systems (A. Standards, 2025). It is based on the general risk management ideas from the ISO 31000 series, but it focuses specifically on AI. The standard states that organisations should look at AI risks in a structured way and over the whole lifecycle of an AI system.

For the context of this paper, several topics from ISO/IEC 23894 are useful:

- AI can change existing risks and can also create new types of risks.
- Risk management should consider all stakeholders, not only IT. In a university, there are also students and lecturers.
- The standard lists typical AI risk sources, such as lack of transparency, quality problems with training data, bias, adversarial attacks, and lifecycle issues.
- Risk sources cam be public Generative AI services or internal AI tools.

ISO/IEC 23894 does not tell universities which tools they should or should not use. But it states that AI risks should be identified, analysed, and treated with a proper process, not just handled informally.

ISO/IEC 42001:2023 defines a complete AI management system (AIMS) (A. Standards, 2025). The structure is similar to other management system standards (like ISO/IEC 27001 for information security). For a university, ISO/IEC 42001 can be used as a structured approach to AI.

The AI policy should clearly state what is allowed and not allowed when using AI in teaching and assessment. This should match the exam regulations and codes of ethics.

Risk assessment should explicitly include academic misconduct and information security: for example, the risk that students use AI in a way that makes grading unreliable or unfair. Monitoring means observing how AI systems are used in real life, including potential misuse by students.

ISO/IEC 42001 also explains how an AI management system can be integrated into other management systems, such as an ISMS according to ISO/IEC 27001 (ÖVE/ÖNORM, 2025). For many universities, this is important, because they already have a security management structure in place. AI does not replace that – it adds a new layer on top.

This paper will therefore build on these three standards. ISO/IEC 23894 provides the risk perspective, and ISO/IEC 42001 provides the management-system perspective. On this basis, concepts of AI TRiSM and ISO/IEC 27001 can help to address AI misuse by students and the related security and governance challenges in teaching and assessment.

Research on information security within AI is mostly focused on technical analyses of security and privacy risks. As an outcome governance frameworks are suggested. But there is no AI-specific risk management framework, that includes that into existing information security management systems.

From a technical perspective Golda et al. provide a comprehensive survey of privacy and security concerns in generative AI, covering threats like training-data leakage, model inversion, membership inference, data poisoning, prompt injection, and the misuse of synthetic media. They state

that generative models must be treated as information assets within formal risk management and governance processes. Their work establishes a general security and privacy risk landscape for generative AI, but it does not address how these risks should be integrated into ISO/IEC 27001-based information security management systems especially not into the specific context of education related topics (A. Golda, 2024).

Yao et al. focus on LLMs security and privacy that explicitly distinguishes three perspectives: "the good" (how LLMs can support security tasks like malware detection), "the bad" (abuse of LLMs for social engineering or phishing), and "the ugly" (vulnerabilities of LLMs themselves (like prompt injection or data leakage). They classify attack types and defence mechanisms across the LLM and provide a detailed threat taxonomy and defence landscape for LLMs, but do not address how these risks should be operationalised within information security management systems (Yao et al., 2024). Zhou et al. focus on ChatGPT. They also classify threats across the LLM lifecycle and argue that LLMs simultaneously act as assets and attack surfaces, which demands a governance processes (2024).

Huang et al. research why and how students cheat with generative AI. They show that students perceive the use of generative AI in higher education as inevitable. Peer norms and personal ethics are the strongest drivers of cheating. The study states the central role of social and moral factors in AI related abuse, but does not address how these risks should be embedded into information security management systems (Huang et al., 2025).

The NIST AI Risk Management Framework (AI RMF) proposes a lifecycle approach to identifying, analysing, and treating AI risks, with explicit attention to the protection of individuals and organisations (AI Risk Management Framework, 2025) whereas AI TRiSM bundles trustworthiness, risk management, and security into an integrated governance concept and provides a conceptual link between AI-specific risk controls and traditional information security responsibilities.

Usually the ISO/IEC 27001 is the base of the ISMS, while AI-specific standards extend the standard. ISO/IEC 23894:2023 positions AI risk management as a specialisation of generic risk management, adding AI-specific risk sources such as data quality, model behaviour, and lifecycle issues. ISO/IEC 42001:2023 formalises this further by defining an AI management system (AIMS) that is explicitly designed to be integrated into ISO/IEC 27001. But these contributions focus on enterprises and do not explicitly address the integrity and confidentiality requirements in the academic field of higher education.

METHODOLOGY

This paper uses the Design Science Research (DSR) methodology to develop an integrated governance model for Artificial Intelligence (AI) in higher education. The developed artefact developed is an extension of an existing Information Security Management System (ISMS) based on ISO/IEC 27001, enhanced with AI-specific governance mechanisms.

DSR enables the systematic combination of established standards and conceptual frameworks into a model that is both theoretically grounded and practically applicable.

Problem Identification and Research Objectives

The research originates from the observation that AI technologies are increasingly embedded within university teaching, learning, and assessment processes. While existing ISMS frameworks address the confidentiality, integrity, and availability of information assets, they insufficiently cover AI-specific risks. These include model behavior, training data quality, lifecycle management, or the misuse of AI systems for academic misconduct. Higher education institutions face domain-specific integrity requirements, such as plagiarism prevention, and intellectual property protection, which are not explicitly addressed by common governance frameworks.

The leading question is:

How can an ISO/IEC 27001-based ISMS be extended to systematically govern AI-related risks in higher education, including academic misconduct and AI-specific security threats?

Knowledge Base and Framework Analysis

As already shown within the related work, the design is grounded in a structured analysis of existing standards and frameworks relevant to information security, AI risk management, and AI governance.

Artifact: Integrated AI Governance Model

Based on the analysis, the core artifact is designed as an integrated AI governance model for higher education. The artifact extends ISO/IEC 27001 by embedding AI-specific governance requirements across the AI lifecycle.

Evaluation Approach

The artifact is evaluated using a conceptual and qualitative evaluation approach. The evaluation is conducted through:

This evaluation does not aim to provide quantitative measurements but to demonstrate that the proposed model provides a systematic and traceable coverage of relevant risks and governance requirements.

The proposed governance model is not empirically validated through case studies or implementations at specific universities. The model can serve as a foundation for future empirical studies and practical implementations.

Artefact Design

In the context of university AI, risks and misuse patterns do not remain stable over time, as both the underlying systems (e.g. model updates) and user behaviour adapt continuously. For this reason, the proposed artefact adopts the Observe–Orient–Decide–Act (OODA) loop as the operational

governance logic, enabling iterative, evidence-based adjustments of policies and controls within the existing ISO/IEC 27001 management structure.

This section specifies the design and structure of the proposed artefact. The artefact is an Integrated AI Governance Model for Higher Education that extends an ISO/IEC 27001:2022 based Information Security Management System (ISMS) with AI-specific governance mechanisms. The focus is on the full AI lifecycle and on the specific demands of universities, where teaching, assessment, and academic integrity introduce new requirements. Since AI systems and AI usage patterns change over time, the model uses an Observe–Orient–Decide–Act (OODA) loop as its operational governance logic to enable short, evidence-based adaptation cycles.

Additional Governance Layers

- *L1. Normative Layer*
- *L2. Management Layer*
- *L3. AI Lifecycle Layer*
- *L4. Technical Layer*

Core Design Concept

A central design decision is to treat AI systems (especially LLM-based systems and RAG assistants) simultaneously as information assets (models, prompts, embeddings, training data, retrieved documents, logs), and attack surfaces (prompt injection, data exfiltration via outputs, adversarial inputs, abuse of retrieval).

Higher-Education Control Extensions

To make the model education-specific, the artefact introduces an explicit governance objective. Ensuring that AI tools cannot be used to bypass assessment rules or create unfair advantages.

Operationally, cheating via AI is addressed through controls such as:

- classification of assessments by AI exposure and enforceability level (e.g. AI-prohibited / AI-limited with disclosure / AI-allowed with attribution / AI-integrated by design),
- explicit rules on permitted AI functions per assessment type (e.g. grammar support vs. content generation vs. code synthesis), including required disclosure and documentation,
- assessment design controls that reduce the value of generic model outputs (e.g. oral defence, process evidence, personalised prompts, staged submissions, in-class components),
- controls for exam material handling and information boundaries (preventing leakage of questions, solutions, feedback, and internal slides into public AI tools),
- monitoring and incident handling procedures for AI-related misconduct, including defined escalation paths aligned with examination regulations and disciplinary processes.

Uses Cases

Context and Assessment Setting

In this use case, a student submits a compensatory written assignment ("Kompensationsleistung") that describes the status of his Master's thesis (actual state, possible hurdles, time management).

Observed Submission Characteristics

The submission summarises thesis progress along typical phases (outline, literature review, methodology, data collection, and ongoing analysis) and also mentions practical challenges such as technical issues with an online survey tool, low response rates, and the need to account for previously unplanned variables during analysis.

At the same time, the overall text is structured very generic and relies on broadly applicable academic phrases that could be applied to almost any thesis without demonstrating that they originate from the own work process.

Specific suspicious findings:

- Similar statements are repeated with minor wording changes ("Zusammenfassend…", "Insgesamt…", "Ein weiterer…"), which increases volume without adding new information.
- Long passages read like general best-practice advice which is typical for generated filler content.
- The text switches terminology (e.g. it refers to a "Dissertation" although the document is a Master's thesis), which is a common artifact of generic text generation.

Misuse Pattern (Cheating via AI)

The relevant misuse pattern is not plagiarism in the classic sense, but AI-enabled misrepresentation of progress: an LLM is likely used to generate a plausible narrative of academic work to satisfy assessment requirements, while the submission itself does not provide evidence that would allow a lecturer to validate the reported progress.

Assets and Impacts

The primary asset impacted is assessment integrity. Grading decisions become less reliable because the artefact cannot be clearly linked to student performance or actual thesis progress. Another impact is operational. The supervision and feedback process becomes less effective because it may be based on fake progress claims.

Governance Layers and Controls Applied

In the proposed governance model, this use case is addressed through multiple layers:

- L1 (Normative layer):
  - Define that AI-generated text is not acceptable (or require explicit disclosure of the usage within a certain range). Require that progress reports must include verifiable elements (e.g. specific artefacts, decisions taken, specific next steps).
- L2 (Management layer):
  - Treat "AI-generated assessments" as an integrity risk in the ISMS/AIMS risk assessment. Define an escalation path that is aligned with exam regulations and ethics.
- L3 (Lifecycle layer):
  - If an AI tool is allowed for writing support, define boundaries (e.g. allowed for language polishing with disclosure, not for content generation that replaces progress evidence).
- L4 (Technical/operational safeguards):
  - Use process-oriented assessment controls rather than "AI detection": staged submissions, version history, draft checkpoints, and short oral validation (e.g. 5–10 minute) triggered by defined indicators.

OODA Loop Operationalisation

This use case fits the OODA logic particularly well:

- Observe: identify indicators (high redundancy, low specificity, missing specific artefacts).
- Orient: interpret the observation in the institutional context (what was required for this compensatory assignment and what AI use is allowed/disallowed).
- Decide: select treatment (request specific artefacts, require a brief oral clarification., apply the formal academic misconduct path only if policy violation is confirmed).
- Act: implement outcome (document decision, refine guidance on permitted AI assistance for future cohorts).

Resulting evidence artefacts are auditable through reviewer notes, oral clarification protocol, decision record, and updated assessment criteria.

Summary

Artificial Intelligence, and especially generative AI, has become part of everyday use in higher education. This can affect teaching and learning positively, but it also introduces a new integrity problem. Students can use AI tools to bypass assessment rules and to produce outputs that are difficult to attribute to individual performance. ISO/IEC 27001:2022 provides a well-established baseline for governing confidentiality, integrity, and availability (CIA). However, it does not explicitly address AI-specific characteristics that become relevant in academic settings, such as training-data dependencies, model drift and hallucinations, ML-specific threats, and limited explainability. As a result, AI misuse in assessment processes constitutes a hybrid risk that is not sufficiently covered by traditional ISMS practice alone.

For this reason, the paper develops an integrated AI governance model for higher education using Design Science Research (DSR). The artefact builds on ISO/IEC 27001 as the baseline governance structure and integrates AI-specific perspectives from ISO/IEC 23894 (risk sources across the AI lifecycle) and ISO/IEC 42001 (AI management system integration). AI TRiSM and the NIST AI RMF provide the conceptual framing for combining trust, risk, and security considerations across the lifecycle.

The artefact is structured into four governance layers (normative policy, management processes, lifecycle controls, and technical safeguards) and treats AI systems simultaneously as information assets and as attack surfaces. The evaluation follows a qualitative approach using standards-conformance and risk-coverage checks, supported by real-world use cases that illustrate AI-assisted academic text production and artefacts.

REFERENCES

A. Golda *u. a.*, "Privacy and security concerns in generative AI: A comprehensive survey", *IEEE Access*, Bd. 12, S. 48126–48144, 2024.

A. Habbal, M. K. Ali, und M. A. Abuzaraida, "Artificial Intelligence Trust, risk and security management (AI trism): Frameworks, applications, challenges and future research directions", *Expert Syst. Appl.*, Bd. 240, S. 122442, 2024.

"AI Risk Management Framework", *NIST*, Juli 2021, Zugegriffen: 25. November 2025. [Online]. Verfügbar unter: https://www.nist.gov/itl/ai-risk-management-framework

A. Standards, "ÖVE/ÖNORM EN ISO/IEC 23894", Austrian Standards. Zugegriffen: 19. November 2025. [Online]. Verfügbar unter: https://www.austrian-standards.at/de/shop/ove-onorm-en-iso-iec-23894-2024-09-15~p4280738

A. Standards, "ISO/IEC 42001:2023", Austrian Standards. Zugegriffen: 19. November 2025. [Online]. Verfügbar unter: https://www.austrian-standards.at/de/shop/iso-iec-42001-2023-2023-12-18~p2897075

D. Huang, N. Hash, J. J. Cummings, and K. Prena, "Academic cheating with generative AI: Exploring a moral extension of the theory of planned behavior", *Comput. Educ. Artif. Intell.*, S. 100424, 2025.

"Enterprise Cybersecurity Solutions, Services & Training | Proofpoint US", Proofpoint. Zugegriffen: 13. November 2025. [Online]. Verfügbar unter: https://www.proofpoint.com/us

"ÖVE/ÖNORM EN ISO/IEC 27001:2023 09 01 | Austrian Standards". Zugegriffen: 19. November 2025. [Online]. Verfügbar unter: https://www.austrian-standards.at/de/shop/ove-onorm-en-iso-iec-27001-2023-09-01~p2907214

Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, und Y. Zhang, "A survey on large language model (llm) security and privacy: The good, the bad, and the ugly", *High-Confid. Comput.*, Bd. 4, Nr. 2, S. 100211, 2024.

W. Zhou *u. a.*, "The security of using large language models: A survey with emphasis on ChatGPT", *IEEECAA J. Autom. Sin.*, 2024.