ARTIFICIAL INTELLIGENCE AND RISK ENGINEERING: TRANSFORMING EDUCATIONAL SYSTEMS INTO SOCIO-TECHNICAL INFRASTRUCTURES

Codruta-Oana HAMAT, PhD, Prof.,
Babes-Bolyai University, Cluj-Napoca, ROMANIA
Dan NEGOITESCU, PhD, Assoc. Prof.,
Politehnica University, Timisoara, ROMANIA
Cornelia-Victoria ANGHEL-DRUGARIN✉, PhD, Lect.,
Babes-Bolyai University, Cluj-Napoca, ROMANIA
Nicoleta MIREA, PhD, Research Eng., *Politehnica University, Research Centre for Engineering and Management, Timisoara, ROMANIA*
✉**Corresponding author: cornelia.anghel@ubbcluj.ro**

ABSTRACT: This paper studies how artificial intelligence (AI) redefines risks in modern educational systems as socio-technical infrastructures. The analysis begins with the assumption that the cultural and epistemological decomposition of institutions is accelerated by automation, algorithmic decision making, and machine judgement. Based on research in engineering risk management, cybernetics, and AI security, the article modelled the spread of "epistemic errors" as technical failures in feedback-driven systems. Introduce a conceptual bridge between technical fault tolerance and cultural resilience, arguing that the absence of human interpretive control is a high level of system vulnerability. The study proposes a hybrid model of risk engineering in education, which integrates principles of technical reliability with human-centred ethics. The results highlight that the main challenges of AI in education are not efficiency, but the preservation of the integrity of epistemology: the ability to maintain meaning, responsibility, and truth in the algorithmic environment.

KEY WORDS: artificial intelligence; risk engineering; socio-technical systems; education; epistemic integrity

1. INTRODUCTION

In recent years, artificial intelligence (AI) has become a technological force that has not only transformed but also poses a fundamental epistemic challenge (Mirea et al., 2024). In addition to its computational capabilities, AI redefines how institutions understand, classify, and respond to risks. The development of predictive analytics, automated decision-making, and data-driven management systems introduces new forms of uncertainty, not only technical, but also cognitive and moral. This transformation requires a reexamination of the meaning of "risk" in the increasingly digital infrastructure-mediated educational system (Sarfraz & Ivascu, 2021). This paper follows the critical reflection initiated in what are the remaining destructions, which explored the limits of digital rationality and the erosion of human interpretation services. These earlier works questioned the silent violence of optimization, the gradual displacement of meaning and critical thinking by algorithmic efficiency. This philosophical concern is translating into a concrete risk engineering framework applied to educational institutions navigating the pressures of digital transformation.

The problem we have to address is not only how to manage risk, but also how to redefine it. In the digital educational environment, risks are no longer limited to the security, security of data, or reliability of systems. It extends to the

epistemological domain: What kinds of knowledge are valued, what forms of reasoning are delegated to machines, and what ethical vulnerabilities arise when human judgment is replaced or limited by algorithmic logic (Mirea et al., 2021). Through the socio-technical risk propagation model of artificial intelligence, we propose that educational risks be understood as the dynamic interaction between human interpretation, institutional design, and technological interaction. The challenge is not to eliminate uncertainty, but to make uncertainty understandable and to transform risks into learning spaces and not simply control objects (Anghel-Drugarin et al., 2024).

2. STATE OF THE ART

Risk engineering uses engineering, system theory, and cybernetics to address risk as the property of artefacts, information flow systems, and connected systems of human actors. Its main concerns are the identification of failure mode, design of control and feedback loops, redundancy and resilience, and integration of monitoring and audit mechanisms into the technical system. In today's digital world, this tradition has been extended to include cyber threats (confidentiality, integrity, availability) and epistemic threats (distributions in knowledge production), and the term "risk" is operational and interpretative. Consequently, practical risk engineering combines quantitative assessment (fault tree, probability risk analysis) with organisational control (management, human-in-the-loop protocols) to prevent lower-level technical failures from accumulating into higher-level institutional failures (Lin et al., 2021).

2.1 Socio-technical systems and AI

AI systems must be analysed as socio-technical systems: their behaviour and effects are generated from interaction between data, algorithms, interfaces, institutional practices, and social norms (Khalifeh et al., 2025). Recent research (Kudina & van de Poel, 2024) has focused on three interdependent dimensions: 1) Technical architecture models, data pipelines, explainability/tracing mechanisms, and technical safeguards. 2) Organisational integration, workflows, roles (the review of model outputs), training, and audit processes that determine how algorithm outputs are used. 3) Social and cultural context, public perception, professional standards, expectations of responsibility, and regulatory frameworks that shape acceptance and legitimacy. Social and technical lenses change the governance problem from the "fixation of models" to shaping a broader system of interpretation and action on model output: therefore, design choices, monitoring mechanisms, and participatory processes are the first element of safety and responsibility. Recent reviews and modelling work explicitly argue that governance must be multidimensional (technical, organisational, social) and adapt to the ability to evolve AI (Kudina & van de Poel, 2024).

2.2 McKinsey - AI in Romania's public sector: opportunities, challenges, key facts

McKinsey's[1] recent country analysis of Romania's public sector combines potential gains and systemic constraints related to education policy and institutional transformation. *Economic and productivity potential:* McKinsey estimates that the adoption of Gen-AI in Romania's public sector could generate significant productivity gains (according to recent analyses: up to €1 billion per year in public sector productivity) and that a broader projection of GDP increase is related to large-scale adoption. *Enablers identified:* The required enablers include robust data infrastructure, interoperable systems, and concentrated investments in AI talents and training across the public workforce. *Major risks and warnings:* The report highlights deficiencies in infrastructure and data, the limited enforcement capacity of governance frameworks, and the essential role of 'human-to-human' control to maintain interpretive supervision, all factors that increase the

[1] https://www.mckinsey.com/industries/public-sector/our-insights/the-transformative-potential-of-ai-in-romanias-public-sector

vulnerability to epistemic risks when AI is deployed on a scale. *Implications:* Measurement of Romania's readiness and high potential benefits provides a policy window for the expansion of artificial intelligence in public services, but without deliberate social technology controls (data governance, teacher/administrator literacy, explanations, and accountability mechanisms), the educational sector is exposed to the spread of risks ranging from technical shortcomings to institutional distrust and epistemological distortion.

3. CASE STUDY

AI has entered the educational sector as both a promise and a risk. In Romania, this transformation coincides with a wider digital transformation in the public sector, in which automation, data analysis, and generative AI are being studied to improve efficiency and transparency. However, as Chiara Marcati warned in *Computer Weekly* (2024), the integration of AI into decision-making introduces new vulnerabilities: algorithm clarity, unresolved data sources, and the erosion of human interpretive agencies[2]. McKinsey & Company's recent analysis, *The Transformative Potential of Artificial Intelligence in Romania's Public Sector* (2024) reveals that the country is in an intermediate stage of its readiness: progress in infrastructure and digital strategy, but gaps in governance, talent, and ethical supervision. These two perspectives (one qualitative and the other empirical) frame our case studies on the spread of epistemic risk in digitalized educational institutions (Munstermann et al., 2025).

3.2. Methodology

The case study adopts a comparative socio-technical lens that combines: (i) thematic interpretation of the Marcati red flag type (data reliability, traceability, and regulatory displacement); (ii) quantitative indicators and qualitative insights from McKinsey and Company's report on Romania's AI preparedness and digital governance capabilities; (iii) the application of these insights to hypothetical educational institutions undergoing an AI. based transformation (evaluation, allocation of resources, and administrative decision). This triangulation enables us to observe how epistemic risks originate from the technical level and spread through the institutional and cultural layers, a process previously shown in the social and technological risk propagation models of AI.

3.3. Findings

Marcati and McKinsey's perspectives converge on a key point: Technical reliability is not equal to epistemic integrity. In Romania, the AI infrastructure still exists and institutional oversight is uneven, creating operational and interpretational risks. For example, McKinsey estimates that Romania could increase its annual income from the adoption of AI in the public sector by up to €1 billion, but stresses that only 32% of citizens are ready for changes driven by AI. Educational institutions reflect this difference: they are encouraged to digitize rapidly, but lack a cognitive, ethical and management framework to responsibly interpret AI outputs (Table 1).

Table 1. AI in the public sector

Rank	Dimension	Ideal AI Governance (Benchmark)	Romanian Educational Practice (Observed/Estimated)
1	Data Quality & Integrity	Continuous validation, diverse datasets, transparency in preprocessing.	Fragmented datasets; reliance on legacy systems; limited metadata transparency.

[2] https://www.computerweekly.com/news/366632644/Interview-Shaping-the-future-of-AI-in-the-UAE

2	Algorithmic Traceability	Explainable AI systems with human-in-the-loop review protocols.	Opaque decision systems; weak teacher/administrator oversight; rare post-hoc auditing.
3	Human Oversight & Literacy	Institutional AI literacy training and interpretive accountability.	Low AI literacy; teachers rely on automated reports; interpretive disengagement.
4	Governance & Regulation	Ethical boards and compliance units integrated with AI strategy.	Partial alignment with the principles of the EU AI Act principles; limited enforcement capacity.
5	Cultural Legitimacy & Trust	Inclusive communication and participatory digital culture.	Scepticism and fear of automation; perception of inequality and loss of agency.

3.4. Implications

The combined evidence highlights a systemic paradox: Romania's public and educational sectors embrace the transformation driven by AI, but risk engineering remains narrowly technical, focusing on cybersecurity and efficiency rather than on epistemic governance. In order to mitigate these risks, educational institutions must: 1) Integrate epistemic auditing into the adoption of AI tools and ensure that algorithm outputs are interpreted and not accepted; 2) Improve interpretive literacy between teachers and administrators to maintain human supervision in digital assessment processes; 3) Institutionalize traceability as a governance principle, allowing for a clear, unbiased, and contestable explanation of all decisions supported by AI; 4) By bringing bridges between cultural legitimacy and participation in digital education, AI will improve human understanding rather than replace it.

Finally, Marcati's red flags and McKinsey's metrics converge and reveal a single insight: The greatest risk of AI-driven education is not technical failure, but institutionalization of epistemological blindness. In this sense, risk engineering becomes a form of cultural preservation, a means of protecting not only the system, but also the meaning itself.

4. THE AI SOCIO-TECHNICAL RISK PROPAGATION MODEL

4.1. Conceptual Basis

The proposed AI Social Technology Risk Promotion Model is conceptually based on Failure Mode and Effect Analysis (FMEA), a classic engineering method for identifying and attenuating potential failure points in complex systems. In traditional engineering, FMEA maps how errors originate from the component level and spread through interdependent subsystems, causing partial or complete system failures. In the context of artificial intelligence applied to educational systems, "components" include not only technical (data, algorithms, software), but also organizational and cultural: human decision-making processes, institutional management, and interpretation frameworks (Anghel-Drugarin et al., 2024). Thus, the model adapts FMEA to the socio-technical field by assuming that failures in subsystems (e.g. bias data or opaque algorithmic logic) can cause cascading effects that threaten organizational reliability or even epistemic integrity. Risk is therefore redefined as the probability of loss of interpretation control rather than just mechanical malfunction. This conceptual translation enables the application of risk engineering logic in non-physical systems: meaning, trust, responsibility, and educational value become measurable parameters of system safety.

4.2. Model Layers and risk flow

The *AI Socio-Technical Risk Propagation Model* consists of four interacting layers (Table 2). Each layer represents a different dimension of potential failure and propagation:

Table 2. The AI Socio-Technical Risk Propagation Model

Layer	System Component	Failure Mode (Origin)	Propagation Path	Effect (Impact)	Mitigation / Control

L1	**Data Layer**	Incomplete, biased, or non-representative data	Training process → model baseline	Distorted knowledge or exclusion bias	Data validation, ethical data governance
L2	**Algorithmic Layer**	Model opacity, overfitting, lack of explainability	Decision logic → automated output	Epistemic drift (loss of meaning and contextual relevance)	Explainable AI, human interpretability, algorithm audits
L3	**Organizational Layer**	Over-automation, lack of human oversight, managerial dependency	Administrative workflows → policy adoption	De-skilling, erosion of accountability	Hybrid decision systems, transparent AI governance
L4	**Cultural Layer**	Loss of interpretive authority and educational ethos	Dissemination → social perception	Degradation of epistemic integrity, public mistrust	Ethical education, cultural resilience programs

In the digital education environment, risks cannot be managed only through technical safeguards and compliance protocols. It should be redefined as a systemic and epistemological phenomenon. Deficiencies in data, classification, or evaluation are rarely isolated, and they spread upward, forming organisational standards and cultural expectations. Algorithms' bias, for example, not only distorts results; it reconfigures institutional values and often reinforces inequalities under the guise of neutrality. The resulting failure is not only an operational inefficiency, but also a crisis of meaning: distortions in truth, interpretation, and educational objectives that cannot be corrected by technical redundancy alone. This dynamic can be understood by the following risk flow model. The model shows how failures originating from data or algorithmic levels tend to move to higher-order structures, organisational, ethical and cultural, and shows the recursive nature of epistemological risks within socio-technical systems. Failures at the data or algorithm level tend to spread upwards to organisational and cultural levels. Once epistemic failures occur (e.g., automatic classification systems produce inaccurate evaluations without human correction), the result is not only operational inefficiency, but also distortion of truth and meaning, and technical redundancy alone cannot be corrected.

4.3. Application and Discussion

When applied to the education system, this model emphasises that risk propagation is non-linear and feedback-driven. An error in a data set may strengthen the algorithm's bias, which then institutionalises the wrong assumption by policies or practices. In such cases, engineering resilience must include not only technical redundancy, but also epistemic redundancy, the ability of human agents to reinterpret, question, and revalidate algorithmic results. This model provides a diagnostic framework to assess (i) where vulnerabilities arise (technical, procedural, cognitive); (ii) how vulnerabilities cross the boundaries between humans and machines; and (iii) what controls (moral, procedural, cultural) are necessary for confinement (see figure 1). In this sense, the social technology and risk propagation models of AI extend traditional risk engineering to the fields of educational neuroscience, suggesting that the ultimate safety of AI-driven systems depends on maintaining both meaning and functional reliability.

5. RISK MITIGATION. ENGINEERING STRATEGIES

To mitigate the epistemological and systemic risks in intelligently controlled education institutions, an integrated approach requires a

combination of technical, organisational, and epistemological security measures. Risk engineering in this context does not focus on eliminating uncertainty, but rather on creating resilient systems that maintain interpretation, traceability, and human intervention (Amin et al., 2019; Lin et al., 2021; Rosa et al., 2025).

5.1. Technical countermeasures

Technical mitigation begins with improving the stability of the infrastructure and algorithmic accountability. *1) Redundancy and monitoring:* Create parallel data validation pipelines and redundancy mechanisms to detect inconsistent results between data sets and model output. Continuous monitoring of AI performance (as a result of anomaly detection, bias drift analysis, and reliability measurement) ensures early detection of epistemological failure. *2) AI Audit and Traceability:* Ensure that AI audit trails record data origin, model versions, and decision-making rationalities. Regular algorithmic audits (both internal and third-party) must verify compliance with fairness, accuracy, and interpretation standards. *3) Explainability Tools:* Introducing an AI Explainability Method (XAI) to make decisions interpretable to nontechnical stakeholders so that teachers and administrators can question the automated results before taking action. Implementing *explainability tools* (or, more precisely, *epistemic and interpretability frameworks)* ensures that AI systems remain auditable, transparent, and cognitively aligned with human reasoning (Kudina & van de Poel, 2024).

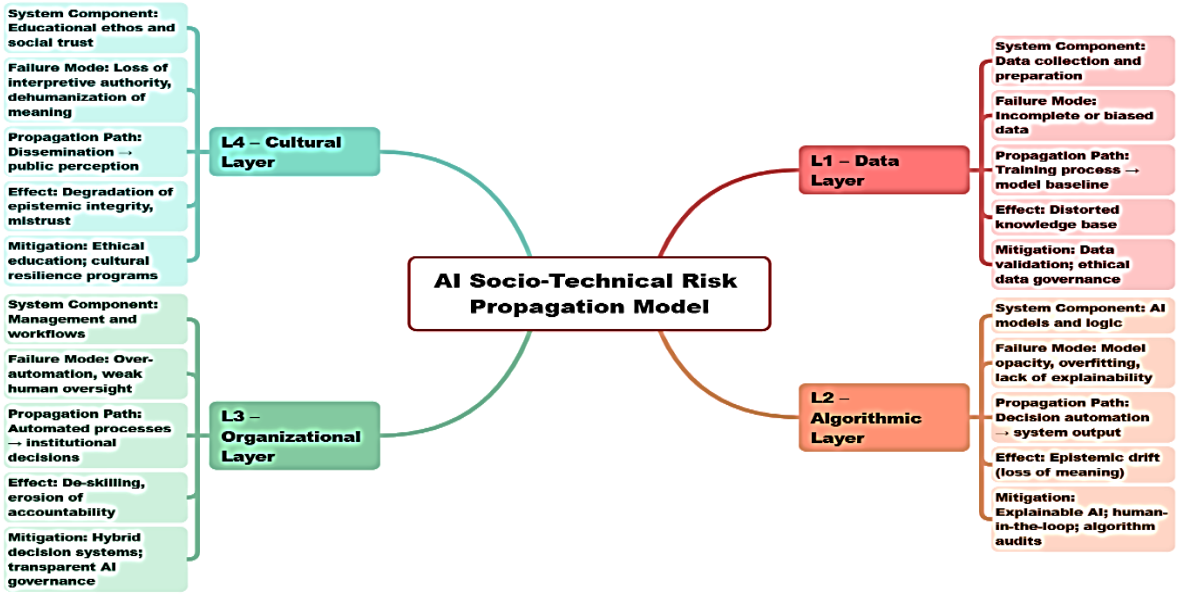

Figure 1. The fourth layers of the AI Socio-Technical Risk Propagation Model

5.2. Institutional countermeasures

Beyond the technical level, institutional resilience depends on governance frameworks to preserve interpretive human supervision. In order to ensure that all AI-assisted decisions, be they evaluative, classification or prediction, include human validation checkpoints, a sustainable educational ecosystem requires human design. Teachers and administrators must maintain the authority to override or contextualise algorithmic results, maintaining the priority of professional judgment. In order to support this, AI governance committees should be established as a multidisciplinary council composed of educators, data scientists, and ethical researchers, tasked with supervising algorithm deployments and defining acceptable institutional risk thresholds. Similarly important is capacity-building and knowledge of AI: educators need to understand how data-driven models work and move from passive users to informed interpreters of AI systems. Finally, feedback and accountability channels must be implemented to enable a transparent contestation of algorithmic decisions and thus strengthen institutional trust, epistemological safety, and collective responsibility.

5.3. Epistemic safety in educational infrastructure

Finally, risk reduction must extend to epistemic safety – the protection of meaning, context, and

interpretation accuracy in digital learning ecosystems. *Hybrid Knowledge Validation:* Combine algorithmic assessment with peer review, qualitative feedback, and contextual judgment to preserve the human dimension of assessment. *Ethics Data Governance:* Design educational data sets with explicit attention to representation, context integrity, and consent to ensure that the "knowledge base" of the AI system reflects the diversity of learners. *Cognitive resilience:* Promote metacognitive skills and digital reflexivity between teachers and students, cultivate the awareness of how algorithms mediate knowledge and shape educational experiences. In synthesis, risk engineering in the learning environment enabled by artificial intelligence must shift from compliance and control to epistemological resilience: it must continuously align technology with institutional significance, cultural trust, and human ability to interpret uncertainty.

CONCLUSION

The evolution of AI from conversational assistance to complex agency and multimodal systems has shown that education is no longer a purely cognitive or institutional process, but a risk-orientated socio-technical infrastructure. AI models become an integral part of research, analysis, and creative production, and their evaluation, selection, and ethical design capabilities become a new form of educational literacy. What used to be the field of education now extends to system design and data governance. Consequently, the transformation of education through AI is not only technical, but also infrastructural, requiring schools and universities to develop both meaning-orientated engineers and code-orientated engineers.

REFERENCES

[1] Amin, M. T., Khan, F., & Amyotte, P. (2019). A bibliometric review of process safety and risk analysis. *Process Safety and Environmental Protection*, *126*, 366–381.

[2] Anghel-Drugarin, C.-V., Hamat, C.-O., & Mirea, N. (2024). TEACHERS' PERCEPTIONS OF ONLINE EDUCATION FROM ROMANIA VERSUS EUROPE DURING 2020-2024: A CASE STUDY. *Annals of "Constantin Brancusi" University of Targu-Jiu*, *2*. https://fcl.eun.org/about-icwg1

[3] ANGHEL-DRUGARIN, C.-V., MIREA, N., BANADUC, G., LUNGU, I., & LANDA, B. L. (2024). NAVIGATING CRISES: TECHNOLOGY, COMMUNICATION, AND THEIR ROLES IN THE COVID-19 PANDEMIC AND TEACHER STRIKES. *ACTA TECHNICA NAPOCENSIS - Series: APPLIED MATHEMATICS, MECHANICS, and ENGINEERING*, *67*(3S). https://atna-mam.utcluj.ro/index.php/Acta/article/view/2565

[4] Khalifeh, A., Qasim, D., Elqirem, I., Al-Ababneh, H., Dajani, D., & Alshamayleh, H. (2025). *Incorporating Social Responsibility in Artificial Intelligence Systems: A Framework of Essential Aspects*. 161–173.

[5] Kudina, O., & van de Poel, I. (2024). A sociotechnical system perspective on AI. *Minds and Machines*, *34*(3), 1–9.

[6] Lin, S. S., Shen, S. L., Zhou, A., & Xu, Y. S. (2021). Risk assessment and management of excavation system based on fuzzy set theory and machine learning methods. In *Automation in Construction* (Vol. 122). Elsevier B.V.

[7] Mirea, N., Anghel-Drugarin, C.-V., & Draghici, A. (2021). A STUDY OF USING GOOGLE CLASSROOM PLATFORM IN THE CASE OF A RURAL PRE-UNIVERSITY SMALL EDUCATION UNIT. *ACTA TECHNICA NAPOCENSIS SERIES-APPLIED MATHEMATICS MECHANICS AND ENGINEERING*, *64*(4), 597–606. https://atna-mam.utcluj.ro/index.php/Acta/article/view/1673

[8] Mirea, N., Palade, M., & Gaureanu, A. (2024). STEM Education, Artificial Intelligence, and Ethical Challenges. *Artificial Intelligence for Human-TechnologiesEconomy Sustainable Development*, 251–260. https://eng.bigai.ai/

[9] Munstermann, B., Janoskuti, L., Dosa, M., Kitipova, N., Tisler, O., Weber, T., & Horacek, M. (2025). *AI in Romania: The potential for the public sector*. https://www.mckinsey.com/industries/pub

lic-sector/our-insights/the-transformative-potential-of-ai-in-romanias-public-sector?utm_source=chatgpt.com

[10] Rosa, T., Pereira, L., Crespo de Carvalho, J., Vinhas da Silva, R., & Simões, A. (2025). From risk to reward: AI's role in shaping tomorrow's economy and society. *AI and Society*, 1–25. https://doi.org/10.1007/S00146-025-02428-1/TABLES/5

[11] Sarfraz, M., & Ivascu; Larisa. (n.d.). *Risk Management Edited by Muddassar Sarfraz and Larisa Ivascu*. Retrieved October 29, 2025, from https://biu.primo.exlibrisgroup.com/discovery/fulldisplay/alma9926858161605776/972BIU_INST:972BIU