1. Healthy Workplaces Summit 2025: discover key takeaways, photos and resources on safe digital work / EU-OSHA. Bilbao, 2025. URL: https://osha.europa.eu/en/highlights/healthy-workplaces-summit-2025-discover-key-takeaways-photos-and-resources-safe-digital-work (дата звернення: 06.05.2026).
2. Mishiba, Takenori. 2024. “Transforming Occupational Health and Safety Regulation: Strategic Pathways in the Era of Industry 4.0.” Journal of Occupational Health Law and Emerging Vision 3, no. 2: 151–169. https://doi. org/10.57523/jaohlev.pp.24-016 (дата звернення: 06.05.2026).
3. World Day for Safety and Health at Work 2025: Revolutionizing Health and Safety: The Role of AI and Digitalization at Work / International Labour Organization (ILO). 2025. URL: https://www.ilo.org/safeday (дата звернення: 06.05.2026).

Mitigating AI-driven security risks in educational software systems

UDK 621.395.7 (043.2)

Stepan Prokipchyn[1]

State University of Information and Communication Technologies,
[1]s.prokipchyn@stud.duikt.edu.ua

The growing use of autonomous AI agents in educational software introduces new cybersecurity challenges due to their ability to interact with external systems and act on behalf of users. The objective of this work is to analyze access control-related risks in such systems and propose practical mitigation strategies. The relevance of the study is driven by the increasing integration of AI into critical educational processes. The scientific novelty is the structured analysis of these risks across different layers of AI usage within educational systems.

The vulnerability surface of AI systems is often defined as prompt injection, data poisoning and hallucination [1, 2]. However, any LLM is prone to these kinds of risks. What makes agentic systems especially vulnerable to these attacks is the main strength of the ReAct pattern – the ability to interact with external systems (load data, perform actions).

In the context of educational systems, the work covers three classes of AI agent use. Internal automation agents – purpose-built agents that automates operational scenarios, running scripted flows (e.g., automated syllabus review per instructional design, suggesting additional materials for students based on their results, AI-assisted learning, etc.). AI-assisted coding and engineering – in 2026 this has become an industry standard, with a significant portion of code generated by AI agents. AI-generated code in system-critical domains like security can lead to vulnerabilities. AI features in the product – production-side features powered by AI agents that students and educators interact with directly. Exposing AI to actual users without proper guardrails not only allows the system to be tricked or abused by dishonest individuals but can also lead to unexpected costs and overall system instability. All three layers are prone to access control-related security risks which are not new but are rather elevated by AI [3].

Credential and token leakage. AI agents can read files, project configuration, MCP definitions, and environment variables. Tokens are especially dangerous: they can be used from anywhere, are often broader-scoped than intended, and persist long after they're useful. Before AI, tokens could be leaked due to security misconfiguration or application bugs. Nowadays, an agent running within the security perimeter can read the token and publish it due to hallucination or prompt injection [2]. To mitigate this risk, long-lived access tokens must be avoided in favor of temporary credentials, SSH keys and other authentication mechanisms limited in lifetime and scope of use.

Over-privileged access. Even short-lived, machine-scoped credentials have some access assigned. Often "minimal" permissions can be broader than intended. Hallucinations can cause the agent to deviate from the intended task to unexpected destructive or corruptive actions (e.g., removing a student user instead of sending an assignment) [2]. LLMs are stochastic in principle. Usually this is a strength, but in some rare cases this can lead to unwanted behavior. Mitigation always requires following the principle of the least privilege. Mitigation includes stricter permission scoping for agent service accounts, the use of fine-grained tokens, and tracking of AI identities. A comprehensive audit layer adds observability to the security plane.

Uncontrolled AI actions. Agents may execute unintended actions, especially when used in auto-approve mode. Bugs in agent tools or in their underlying MCPs can lead to unauthorized data changes. For agents to be useful, they must be provided a certain level of write access. We can forbid destructive actions via fine-grained permission configuration, but most tasks will still require data modification, email sending, API requests, etc. Even within a hardened security perimeter, a hallucinating agent can do harm. There is no general solution to this risk. However, secure-by-design implementation of tools and MCPs can significantly reduce it. In addition to that, a secondary observer AI agent can be running in the background and validating the primary agent intent. Such observers are less prone to hallucination because their context is often limited to an action that the primary agent is trying to execute and a brief explanation of intent. If the observer is unsure or the potential risk of the action is higher than a certain threshold – a human-in-the-loop pattern can be triggered, requiring explicit operator approval.

The work shows that key security risks: credential leakage, over-privileged access, and uncontrolled actions are amplified by agent autonomy. Mitigation requires least-privilege access, short-lived credentials, audit mechanisms, and human-in-the-loop controls. These results provide practical guidance for securing AI-enabled educational systems and form a basis for further research on controllable agentic architectures.

1. OWASP Gen AI Security Project. OWASP Top 10 for LLM Applications 2025. URL: https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025 (date of access: 08.05.2026)
2. Tang Y., et al. Security of LLM-based agents regarding attacks and defenses. Journal of Network and Computer Applications, 2025. DOI: https://doi.org/10.1016/j.inffus.2025.103941.
3. Gao Y., Wu S. A Four-Layer Security Governance Framework for LLM-Based AI Agents. Journal of Artificial Intelligence Practice, 2025. DOI: http://dx.doi.org/10.23977/jaip.2025.080406.