Report

Organization: **Fayston Preparatory School**

https://fayston.org/

Publication date: Apr 22, 2025, 09:02 AM PDT

Reporting period: 2025

Section 1 - Risk identification and evaluation

a. How does your organization define and/or classify different types of risks related to AI, such as unreasonable risks?

I have developed an original classification system tailored for K–12 education, defining “unreasonable risks” as those that pose unacceptable threats to student safety, institutional credibility, or public trust. Risks are categorized into ethical (e.g., algorithmic bias), operational (e.g., system misuse), and systemic (e.g., governance failure) types, informed by the OECD AI Principles and the G7 Hiroshima AI Process Code of Conduct.

b. What practices does your organization use to identify and evaluate risks such as vulnerabilities, incidents, emerging risks and misuse, throughout the AI lifecycle?

Although the framework has not yet been deployed, I have designed a lifecycle-based risk evaluation process, including pre-use rubrics, misuse scenario logs, and oversight workflows. These are intended to help institutions identify and monitor vulnerabilities, misuse, and emerging risks in school contexts before and after AI system deployment.

c. Describe how your organization conducts testing (e.g., red-teaming) to evaluate the model's/system's fitness for moving beyond the development stage?

I have structured a red-teaming methodology focused on educational AI systems. This includes simulation of edge cases such as biased feedback loops, adversarial misuse, and unintended student interactions. These tests are intended to be implemented during pilot phases and reviewed by independent advisors.

d. Does your organization use incident reports, including reports shared by other organizations, to help identify risks?

No

e. Are quantitative and/or qualitative risk evaluation metrics used and if yes, with what caveats? Does your organization make vulnerability and incident reporting mechanisms accessible to a diverse set of stakeholders? Does your organization have incentive programs for the responsible disclosure of risks, incidents and vulnerabilities?

I have developed both qualitative tools (e.g., stakeholder interviews, perception audits) and quantitative methods (e.g., deviation scoring, impact thresholds). Reporting mechanisms are designed to be multilingual, anonymous, and inclusive. While no monetary incentives are in place yet, the framework encourages recognition-based disclosures via student and staff governance channels.

f. Is external independent expertise leveraged for the identification, assessment, and evaluation of risks and if yes, how? Does your organization have mechanisms to receive reports of risks, incidents or vulnerabilities by third parties?

Yes. I have consulted with international accreditation and AI policy experts in the development phase. The framework includes pathways for independent third-party reporting, along with secure intake protocols to ensure confidentiality and transparency.

g. Does your organization contribute to the development of and/or use international technical standards or best practices for the identification, assessment, and evaluation of risks?

Yes. The framework is explicitly aligned with the OECD AI Principles, the G7 Hiroshima AI Process Code of Conduct, and informed by best practices from UNESCO, NIST, and ISO/IEC standards. It is designed to translate these technical and policy frameworks into practical tools for school-level governance.

h. How does your organization collaborate with relevant stakeholders across sectors to assess and adopt risk mitigation measures to address risks, in particular systemic risks?

I am in the process of establishing cross-sector partnerships to pilot the framework. It is designed to support collaboration between school leaders, AI developers, policymakers, and accreditation bodies. Mitigation strategies will be co-developed during implementation, grounded in real institutional needs.

Any further comments and for implementation documentation

This is an original, independently developed governance framework authored by me, Timothy Kang. It is currently pre-deployment and was created to fill a policy gap in the application of international AI principles to school systems. I welcome collaboration with OECD and G7 stakeholders to pilot, refine, and scale this contribution.

Section 2 - Risk management and information security

a. What steps does your organization take to address risks and vulnerabilities across the AI lifecycle?

I have designed an AI governance framework that outlines clear mitigation protocols for each stage of the AI lifecycle: pre-deployment (design and consent checks), deployment (limited-use environments and monitoring), and post-deployment (incident tracking and audits). Each phase includes documented steps for addressing foreseeable vulnerabilities, particularly in educational AI contexts.

b. How do testing measures inform actions to address identified risks?

Testing measures such as scenario analysis, misuse simulations, and ethical stress testing are built into the framework as tools to surface latent risks. Once piloted, these will guide mitigation through model revision, policy updates, and adjusted access or usage permissions.

c. When does testing take place in secure environments, if at all, and if it does, how?

The framework prescribes testing in restricted sandboxed environments prior to any live deployment. These secure environments will simulate school use cases, preventing real-world impact during the testing phase. All tests will be logged and reviewed by governance teams.

d. How does your organization promote data quality and mitigate risks of harmful bias, including in training and data collection processes?

Data protocols emphasize contextual relevance, de-identification, and auditability. The framework includes safeguards to prevent bias amplification through diverse data review panels, documented provenance, and policy alignment with OECD fairness principles. While no data has yet been collected or processed under this framework, the structures are in place.

e. How does your organization protect intellectual property, including copyright-protected content?

The framework emphasizes strict use of licensed, open-source, or institutionally owned datasets. All AI-assisted tools will be required to respect copyright constraints, with detailed guidance provided to school leaders and developers on identifying, attributing, and avoiding unauthorized content use.

f. How does your organization protect privacy? How does your organization guard against systems divulging confidential or sensitive data?

I have designed this framework to operate under strong privacy-first principles. Student and staff data must be anonymized, access-limited, and compliant with applicable data protection laws. Governance protocols include user consent tracking, redaction procedures, and internal audits to prevent inadvertent data disclosure.

g. How does your organization implement AI-specific information security practices pertaining to operational and cyber/physical security?

- **i. How does your organization assess cybersecurity risks and implement policies to enhance the cybersecurity of advanced AI systems?**
- **ii. How does your organization protect against security risks the most valuable IP and trade secrets, for example by limiting access to proprietary and unreleased model weights? What measures are in place to ensure the storage of and work with model weights, algorithms, servers, datasets, or other relevant elements are managed in an appropriately secure environment, with limited access controls in place?**
- **iii. What is your organization's vulnerability management process? Does your organization take actions to address identified risks and vulnerabilities, including in collaboration with other stakeholders?**
- **iv. How often are security measures reviewed?**
- **v. Does your organization have an insider threat detection program?**

i. The framework includes a cybersecurity risk checklist based on ISO/IEC 27001 and NIST AI RMF. These tools are intended to be used during vendor review and internal system evaluation stages.

ii. While I do not currently manage proprietary model weights, the framework calls for restricted-access storage, encrypted cloud environments, and multi-factor authentication for all sensitive elements in schools using third-party or open AI models.

iii. A staged vulnerability log and triage system is designed to flag, classify, and escalate risks to an oversight team, with external support if necessary. Risk resolution protocols are outlined in detail for future implementation.

iv. The framework mandates quarterly review of AI tools and security protocols by designated school governance teams, and external audit every 12 months once deployed.

v. Yes, a policy is drafted within the framework that defines misuse monitoring, role-based access controls, and internal reporting pathways for potential insider risk—particularly relevant in educational staff usage scenarios.

h. How does your organization address vulnerabilities, incidents, emerging risks?

A proactive incident management structure is included, featuring a disclosure hotline, stakeholder communication guidelines, and a classification framework for prioritizing emerging risks. While not yet active, this system is ready for deployment during the pilot phase.

Any further comments and for implementation documentation

This security and risk management protocol is part of an original, independently developed governance framework authored by me, Timothy Kang. It is designed for integration into educational institutions and accreditation systems, and aligns fully with HAIP principles. I welcome collaboration to test and refine the model in live pilot settings.

Section 3 - Transparency reporting on advanced AI systems

a. Does your organization publish clear and understandable reports and/or technical documentation related to the capabilities, limitations, and domains of appropriate and inappropriate use of advanced AI systems?

- i. How often are such reports usually updated?
- ii. How are new significant releases reflected in such reports?
- iii. Which of the following information is included in your organization's publicly available documentation: details and results of the evaluations conducted for potential safety, security, and societal risks including risks to the enjoyment of human rights; assessments of the model's or system's effects and risks to safety and society (such as those related to harmful bias, discrimination, threats to protection of privacy or personal data, fairness); results of red-teaming or other testing conducted to evaluate the model's/system's fitness for moving beyond the development stage; capacities of a model/system and significant limitations in performance with implications for appropriate use domains; other technical documentation and instructions for use if relevant.

I. I have developed comprehensive templates and reporting structures as part of my AI governance framework. While they have not yet been deployed, the framework is designed to produce and maintain clear technical documentation and usage guidelines for advanced AI tools used in education.

II. Major system updates or vendor changes will trigger an immediate addendum to the documentation, which must be reviewed by governance teams before deployment.

III. The framework is designed to include:

- Evaluation results for safety, security, and societal risks
- Risk assessments related to human rights, fairness, and bias
- Red-teaming/test results for readiness
- Documentation of capabilities and limitations
- Instructions for safe and appropriate use

b. How does your organization share information with a diverse set of stakeholders (other organizations, governments, civil society and academia, etc.) regarding the outcome of evaluations of risks and impacts related to an advanced AI system?

The framework outlines a stakeholder communication protocol for educational AI deployments, including shared reporting with accrediting agencies, peer institutions, and policy research networks. I intend to release white papers and participate in forums such as OECD.AI, GPAI, and MSA as part of its dissemination strategy.

c. Does your organization disclose privacy policies addressing the use of personal data, user prompts, and/or the outputs of advanced AI systems?

Yes. I have drafted a set of privacy guidelines covering student and teacher data, AI-generated outputs, and prompt history. These are written to align with FERPA, OECD Data Governance Principles, and emerging AI-specific privacy standards, and will be published alongside system documentation.

d. Does your organization provide information about the sources of data used for the training of advanced AI systems, as appropriate, including information related to the sourcing of data annotation and enrichment?

While I do not personally train foundation models, the framework includes mandatory disclosure requirements for any vendor-provided systems, including training data sources, labeling practices, and data enrichment procedures. Schools adopting third-party AI tools will be required to make this information publicly accessible.

e. Does your organization demonstrate transparency related to advanced AI systems through any other methods?

Yes. The framework includes a transparency index for each AI tool used, governance dashboards for school boards, and opt-in/opt-out features for users. I have also structured feedback loops through student councils, staff committees, and community forums to ensure multi-stakeholder transparency beyond documentation alone.

Any further comments and for implementation documentation

This section is based on an independently developed, pre-deployment AI governance framework designed by me, Timothy Kang. It is built to model transparency as a core value of institutional AI adoption in K–12 systems and is intended for international sharing, piloting, and policy alignment under HAIP and OECD efforts.

Section 4 - Organizational governance, incident management and transparency

a. How has AI risk management been embedded in your organization governance framework? When and under what circumstances are policies updated?

I have developed a governance framework that integrates AI risk management into school-level leadership structures, accreditation workflows, and policy review cycles. Risk policies are designed to be reviewed quarterly and updated in response to emerging threats, regulatory changes, or audit findings. The framework includes predefined triggers for policy revisions, such as incidents, new deployments, or material vendor updates.

b. Are relevant staff trained on your organization's governance policies and risk management practices? If so, how?

Although not yet implemented, the framework includes a comprehensive training protocol for school leaders and faculty. It features role-specific guidance, annual refreshers, onboarding briefings for new staff, and scenario-based exercises to promote governance fluency in AI contexts.

c. Does your organization communicate its risk management policies and practices with users and/or the public? If so, how?

Yes. The framework is designed to include a public-facing AI governance statement, user rights documentation, and risk management summaries published through school websites, accreditation reports, and parent/community briefings. These communications are intended to build institutional transparency and public trust.

d. Are steps taken to address reported incidents documented and maintained internally? If so, how?

Yes. The framework includes an internal incident management protocol that requires all steps—from report intake to resolution—to be logged in a secure risk registry. Templates for incident documentation, analysis, and resolution tracking are built into the system for consistent accountability.

e. How does your organization share relevant information about vulnerabilities, incidents, emerging risks, and misuse with others?

A tiered sharing protocol is built into the framework. Critical incidents would be escalated to governance boards and external evaluators. Summaries of non-sensitive incidents are intended for inclusion in public risk updates, and emerging trends would be shared with peer institutions and policy networks where applicable.

f. Does your organization share information, as appropriate, with relevant other stakeholders regarding advanced AI system incidents? If so, how? Does your organization share and report incident-related information publicly?

Yes. The framework includes a policy for responsible disclosure of significant incidents to stakeholders such as accrediting bodies, education ministries, and peer organizations. While privacy is prioritized, non-identifiable summaries will be shared publicly to contribute to collective learning and risk prevention.

g. How does your organization share research and best practices on addressing or managing risk?

I intend to publish insights and implementation guides through international forums (e.g., OECD.AI, UNESCO, GPAI), as well as contribute to accreditation and education leadership networks. The framework is designed to support case study publication and inter-school collaboration on AI governance.

h. Does your organization use international technical standards or best practices for AI risk management and governance policies?

Yes. This framework is grounded in best practices drawn from the OECD AI Principles, G7 Hiroshima AI Process Code of Conduct, UNESCO's AI in Education recommendations, and the NIST AI Risk Management Framework. It adapts these standards to fit the operational realities of schools and educational institutions.

Any further comments and for implementation documentation

This section reflects an original, independently authored governance framework created by me, Timothy Kang. Though pre-deployment, it was intentionally built to scale across schools globally and contribute meaningfully to the advancement of transparent and responsible AI risk governance.

Section 5 - Content authentication & provenance mechanisms

a. What mechanisms, if any, does your organization put in place to allow users, where possible and appropriate, to know when they are interacting with an advanced AI system developed by your organization?

As part of my independently developed governance framework for education systems, I have integrated a requirement for AI-use disclosure protocols. These include clear user-facing indicators—such as labels, prompts, and usage disclaimers—whenever users interact with or receive output from advanced AI systems. These disclosures are designed to appear within learning platforms, parent notifications, and student interfaces to ensure transparency, particularly for minors in school environments.

---

b. Does your organization use content provenance detection, labeling or watermarking mechanisms that enable users to identify content generated by advanced AI systems? If yes, how? Does your organization use international technical standards or best practices when developing or implementing content provenance?

Yes. Although the system is pre-deployment, I have included in the framework a policy requiring AI-generated content to be labeled or watermarked, where technically feasible. This includes instructional content, assessments, and communications generated by AI tools. These mechanisms are designed to comply with emerging international standards, such as W3C provenance models, ISO/IEC 23053, and OECD recommendations. In addition, all AI-generated materials used in classrooms are expected to carry metadata or visual markers to distinguish them from human-created content.This section reflects provisions within an original, pre-deployment governance framework authored solely by me, Timothy Kang. It is designed to meet and extend HAIP-aligned best practices for AI transparency and provenance in the education sector and will evolve alongside future global content authentication standards.

Section 6 - Research & investment to advance AI safety & mitigate societal risks

a. How does your organization advance research and investment related to the following: security, safety, bias and disinformation, fairness, explainability and interpretability, transparency, robustness, and/or trustworthiness of advanced AI systems?

I have developed a governance framework that explicitly integrates research-backed criteria for fairness, explainability, and AI trustworthiness in K–12 education. Although the framework is not yet deployed, it incorporates safeguards such as bias impact rubrics, explainability checklists, and documentation protocols intended to strengthen institutional AI safety. The framework is grounded in ongoing review of global best practices and is structured to inform future investment and iterative refinement.

b. How does your organization collaborate on and invest in research to advance the state of content authentication and provenance?

While the framework is currently self-funded and independently authored, I actively contribute thought leadership through my involvement in international policy and accreditation forums. I am seeking collaboration with institutions, researchers, and technology developers to pilot the provenance labeling and audit mechanisms embedded in the framework—especially in contexts involving minors, educational integrity, and instructional content verification.

c. Does your organization participate in projects, collaborations, and investments in research that support the advancement of AI safety, security, and trustworthiness, as well as risk evaluation and mitigation tools?

Yes. As a Commission Representative for an international accreditation body and a member of multiple governance networks (including Center for AI Safety), I participate in cross-sectoral working groups that explore scalable AI safety policies. My framework is intended to serve as a use-case for integrating AI governance into school evaluations and to catalyze further research into context-sensitive risk mitigation tools.

d. What research or investment is your organization pursuing to minimize socio-economic and/or environmental risks from AI?

The framework is designed with a focus on equitable access and digital inclusion, particularly in under-resourced K–12 environments. It emphasizes transparency in algorithmic decision-making that may affect student placement or performance. While environmental metrics are not yet embedded, future iterations will consider carbon-aware AI adoption policies for schools using cloud-based or compute-intensive systems.

Any further comments and for implementation documentation

This framework is an original, pre-deployment initiative authored independently by me, Timothy Kang. It is positioned to serve as a globally scalable governance reference for schools and accreditation systems, offering a values-driven, education-specific contribution to OECD and G7 priorities in AI safety, fairness, and societal responsibility.

Section 7 - Advancing human and global interests

a. What research or investment is your organization pursuing to maximize socio-economic and environmental benefits from AI? Please provide examples.

I have developed a governance framework aimed at reducing inequality in AI access, safeguarding student rights, and ensuring ethical oversight in education—especially in international schools and emerging economies. The framework promotes cost-effective, low-compute AI applications in learning environments, with future iterations planned to incorporate environmental sustainability metrics to guide responsible procurement and infrastructure usage.

b. Does your organization support any digital literacy, education or training initiatives to improve user awareness and/or help people understand the nature, capabilities, limitations and impacts of advanced AI systems? Please provide examples.

Yes. The framework embeds AI literacy modules for students, faculty, and leadership. These include policy briefings, ethical simulation workshops, and responsible use protocols. I have also contributed to national and international digital literacy conversations and am preparing guidance aligned with UNESCO's AI competencies for educators. Once piloted, this initiative will provide a replicable model for school-based AI literacy worldwide.

c. Does your organization prioritize AI projects for responsible stewardship of trustworthy and human-centric AI in support of the UN Sustainable Development Goals? Please provide examples.

Yes. This framework directly supports SDG 4 (Quality Education), SDG 9 (Industry, Innovation, and Infrastructure), and SDG 16 (Peace, Justice, and Strong Institutions). It promotes human-centric governance, safeguards against biased or opaque AI decision-making, and ensures that AI use in schools respects the dignity and rights of all learners. It was developed as a policy-to-practice bridge for human-centric and inclusive AI adoption.

d. Does your organization collaborate with civil society and community groups to identify and develop AI solutions in support of the UN Sustainable Development Goals and to address the world's greatest challenges? Please provide examples.

While formal pilots are pending, I have designed the framework to be adaptable in collaboration with ministries of education, accreditation bodies, and civil society actors concerned with equity in tech adoption. Ongoing dialogues with governance experts and academic networks help ensure the framework remains inclusive, cross-cultural, and globally relevant.

Any further comments and for implementation documentation

This final section reflects my commitment—as the sole author of this framework—to shaping AI governance through the lens of educational equity, human rights, and long-term sustainability. I welcome collaboration with OECD, G7, and UN-aligned partners to advance this work from concept to implementation.