Safe AI in Education Manifesto

Safe AI in Education Manifesto

View on GitHub

Safe AI in Education Manifesto

version 0.4.0 8 October 2024

If you want so sign the manifesto, please fill in this form.

Spanish version

Table of Contents

Preamble

- Principle 1: Human Oversight and Accountability
- Principle 2: Guaranteeing Confidentiality
- Principle 3: Alignment with Educational Strategies
- Principle 4: Alignment with Didactic Practices
- Principle 5: Accuracy and Explainability
- Principle 6: Comprehensive Interface and Behavior
- Principle 7: Ethical Training and Transparency

Conclusion

Authored by

Signatories

How to cite the Safe AI in Education Manifesto

Preamble

In the rapidly evolving landscape of education, Artificial Intelligence (AI) offers great prospects for improving learning experiences, streamline administrative tasks, and support both students and educators. However, with these advances come significant responsibilities.

The Safe AI in Education Manifesto establishes the foundational principles that ensure AI is deployed in educational settings in a manner that is ethical, secure, and aligned with the core objectives of education.

These principles are grounded in the belief that AI should always be at the service of people, enhancing human capabilities rather than replacing them. AI should act as a tool that empowers students, educators, and institutions to achieve their full potential while preserving the human-centric nature of education.

These principles are intended to provide a framework to guide educators, educational institutions, developers and AI vendors in their process to decide wether and how AI should be used in education. Although other considerations about AI in education are important, we believe that these principles are relevant and actionable in this context.

Artificial Intelligence (AI) is a field of research, not a specific set of technologies or even a specific approach. This manifesto concerns mainly the current crop of AI technologies being introduced in the educational landscape in the mid-2020s, most of which are based on some form of machine learning and usually called Generative AI. As new breakthroughs in AI technologies emerge, this manifesto will be updated to reflect the new state of the art.

The principles in the manifesto have inspired a Checklist that can be used to evaluate and integrate AI tools in education.

This is a living document and it will be updated as the community and the technology mature.

Principle 1: Human Oversight and Accountability

AI tools in education must always complement, not replace, human educators. While AI can assist with administrative tasks like grading or providing feedback, all decision-making processes must remain under human supervision. AI-driven decisions should be explainable, and students must have the right to appeal these decisions through human-led processes. This ensures fairness, maintains the role of teachers as mentors, and protects the integrity of the educational process.

Declaration: An AI system cannot be responsible for the education of students. Any decisions made by AI, particularly those related to grading and assessment, must be transparent and fully accountable to human oversight. Students must retain the right to appeal AI-driven decisions, which must be evaluated and addressed by human educators. AI's role should be to enhance, not diminish, the human connection essential to education.

Principle 2: Guaranteeing Confidentiality

We commit to safeguarding the privacy and confidentiality of all student data, including identities, roles, academic records, and interactions. AI systems must be designed and implemented with stringent security measures to protect student information. Educational institutions should either own and operate the technology stack or require strict privacy compliance from AI vendors, ensuring no data is exposed to unauthorized parties.

Declaration: We caution against the use of free tools that require students to register with third-party services as a mandatory component of their education. Students should retain full control over their personal data, and the institution needs to ensure that all technologies used in the educational environment are secure, transparent, and under the institution's direct control. This prevents any compromise to the privacy or rights of students.

Principle 3: Alignment with Educational Strategies

AI tools must be in harmony with the educational strategies and IT governance of the institutions they serve. They should support learning objectives without introducing undue complexity or facilitating unethical practices such as cheating or plagiarism. AI should be a tool that enhances educational outcomes, not one that complicates the learning process.

Declaration: We recognize the risk of using general-purpose AI tools that are not specifically designed for educational contexts. The complexity of such tools, together with the risk of their misuse, undermines the educational process and adds to the background noise that impedes proper cognitive processing. We advocate for the use of AI systems that are tailored to fit the specific needs and goals of educational institutions.

Principle 4: Alignment with Didactic Practices

The deployment of AI in education must be grounded in established didactic practices. Educators need to have a clear understanding of how AI tools will integrate with their instructional design and learning objectives. AI should support teaching methodologies rather than disrupt them.

Declaration: We emphasize the need for AI tools to be adaptable to various instructional designs. Whether through specialized interfaces or specific configurations, AI should be a seamless extension of the teaching-learning process, providing support without imposing additional burdens on educators or students.

Principle 5: Accuracy and Explainability

Accuracy is paramount in educational contexts. AI systems must prioritize delivering precise, explainable and relevant information, especially given the risks of hallucinations and errors inherent in current AI technologies. These risks can be mitigated by deploying AI in narrowly defined application contexts and ensuring that AI tools reference their sources, allowing users to verify the information provided.

Declaration: We encourage rigorous quality assessments for AI tools used in education. The reliability of AI systems is non-negotiable, and continuous evaluation is necessary to maintain the integrity of the educational process.

Principle 6: Comprehensive Interface and Behavior

AI systems must have interfaces that are transparent and easily understood by students and educators. The behavior of these systems should clearly communicate their intended use and limitations, avoiding any pretense of infallibility or omniscience. AI generated content should be always clearly marked as such.

Declaration: We advocate for AI tools that are designed with clarity and transparency at their core. These tools must explicitly convey their limitations and avoid presenting erroneous information with undue confidence. In doing so, they will support a learning environment where students can trust, but also critically assess, the AI's outputs.

Principle 7: Ethical Training and Transparency

AI models used in education must be trained in an ethical manner, with a clear commitment to transparency regarding the sources of training data and the methodologies used. It is essential that these models actively work to minimize biases and provide transparency about their training processes, allowing educators and students to understand the limitations and considerations involved in the AI's outputs.

Declaration: We insist that all AI tools in educational contexts must be developed and deployed with a commitment to ethical standards. This includes openly declaring the sources of training data, ensuring that the models are designed to minimize biases, and providing transparency about the potential limitations inherent in the AI's decision-making processes.

Conclusion

The integration of AI into education holds great potential. However, this integration must be guided by principles that prioritize the well-being, privacy, and educational success of students. The Safe AI in Education Manifesto is a commitment to these principles, ensuring that instututions and educators introduce AI in education in a way that enhances the educational experience without compromising ethical standards or educational integrity.

Authored by

Marc Alier Forment, Francisco Garcia Peñalvo, Maria José Casañ Guerrero, Juan Antonio Pereira and Faraón Llorens Largo. With additional insights from Roberto Rodríguez Echeverría, Ramón Martí, Maria Pilar Almanjano Pablos, Antoni Hernández-Fernández and Jordi Cortadella .

How to cite the Safe AI in Education Manifesto

APA Style:

In-text citation: (Alier Forment et al., 2024)

Reference list: Alier Forment, M., Garcia Peñalvo, F., Casañ Guerrero, M. J., Pereira, J. A., & Llorens Largo, F. (2024, October 8). *Safe AI in Education Manifesto* (Version 0.4.0). *Safe AI in Education Manifesto*. https://manifesto.safeaieducation.org

MLA Style:

In-text citation: (Alier Forment et al.)

Works Cited: Alier Forment, Marc, et al. *Safe AI in Education Manifesto*. Version 0.4.0, 8 Oct. 2024, https://manifesto.safeaieducation.org.

Chicago Style:

In-text citation: (Alier Forment et al. 2024)

Bibliography: Alier Forment, Marc, Francisco Garcia Peñalvo, Maria José Casañ Guerrero, Juan Antonio Pereira, and Faraón Llorens Largo. *Safe AI in Education Manifesto*. Version 0.4.0, October 8, 2024. https://manifesto.safeaieducation.org.
