Assessing the Fairness of AI Systems for Education

Velislava Hillman[1], Katarzyna Barud[2], Ibrahim Sabra[2,*], Clara Saillant[2],
Syed Zulkifil Haider Shah[2], Lukas Faymann[2], Edoardo Pareti[3], Leo Bianchi[4] and
Manuele Barbieri[3]

[1]*London School of Economics I& Political Science, United Kingdom*
[2]*University of Vienna, Department of Innovation and Digitalisation in Law, Austria*
[3]*RINA Consulting S.p.A.,Italy*
[4]*Institute of Theoretical Astrophysics, University of Oslo, Blindern, Oslo, Norway*

Abstract

As the EU AI Act comes into force, AI systems in education, particularly those influencing learning outcomes and student pathways, can be considered 'high-risk' as these systems can significantly impact individuals' educational trajectories by profiling students and educators, predicting outcomes, and inadvertently perpetuating biases or limiting future opportunities. Given the rising integration of such systems into education, it becomes imperative to assess their accuracy, trustworthiness and fairness. This paper presents a comprehensive framework integrating legal, socio-ethical and technological dimensions to assess AI trustworthiness. We apply this framework to two case studies—an assessment platform (Thrively) and an adaptive learning tool (Century Tech)—and demonstrate their utility. Using a Natural Language Model to automate developer documentation analysis, we reveal gaps in fairness and data security. Our contribution offers timely insights to guide the ethical development of educational AI and support compliance with the EU's AI Act (2024).

Keywords

AI in education, EU AI Act, Trustworthy AI, AI assessment frameworks

1. Introduction and research context

Artificial intelligence in education (AIED) is likely designated as 'high-risk' under EU AI Act (s.6(2), Annex 3(3), and recital 56 of the AI Act) [1] due to their potential impact on students' rights, data privacy, and educational outcomes. The Act, entering into force progressively until 2027, already applies certain provisions including those related to definitions, AI literacy and prohibited practices. As a result, many AIED applications will be subject to strict regulatory requirements. Relevant prior work in AIED and policy already highlights both the opportunities and risks emanating from such fast-advancing technologies including learning-sciences-driven design of educational technology (EDtech) and cautionary analyses of AI's use in schools [2], [3] as well as European Commission guidance to help teachers address AI misconceptions and promote ethical use [4]. AIED promises inclusivity [5], timely assessment [6], and alternative learning provisions [7]. These systems increasingly rely on predictive models to support students, teachers, and administrators [2], including identifying at-risk students, flagging concerns based on predicted disengagement or difficulties [8]. However, AIED remains technically immature and often over-promoted [3]. Risks range from basic yet fundamental issues like data privacy loss and cybersecurity [9] to unpredictable risks and errors such as bias and social

TRUST-AI: The European Workshop on Trustworthy AI. Organized as part of the European Conference of Artificial Intelligence - ECAI 2025. October 2025, Bologna, Italy.
*Corresponding author.
v.hillman@lse.ac.uk (V. Hillman); katarzyna.barud@univie.ac.at (K. Barud); ibrahim.sabra@univie.ac.at (I. Sabra); clara.saillant@univie.ac.at (C. Saillant); syed.zulkifil.haider.shah@univie.ac.at (S. Z. H. Shah); lukas.faymann@univie.ac.at (L. Faymann); edoardo.pareti@rina.org (E. Pareti); l.a.bianchi@astro.uio.no (L. Bianchi); manuele.barbieri@rina.org (M. Barbieri)
https://www.lse.ac.uk/ (V. Hillman); https://id.univie.ac.at/en/ (K. Barud); https://www.rina.org/ (E. Pareti); https://www.mn.uio.no/astro/english/ (L. Bianchi)
0000-0002-3661-0914 (V. Hillman); 0000-0001-9882-2712 (K. Barud); 0009-0008-7703-1143 (I. Sabra); 0009-0004-8761-1441 (S. Z. H. Shah); 0009-0004-9205-8409 (L. Faymann); 0009-0006-8604-3985 (E. Pareti); 0009-0002-6351-5426 (L. Bianchi); 0000-0003-4310-8356 (M. Barbieri)

injustice [10]. Without clear governance, educators and students must navigate these tools without established frameworks for trustworthiness and accountability. The high-risk status of AIED stems from its capacity to influence educational trajectories, with opaque algorithms that may profile students and educators and unintentionally reinforce inequality. A recent bibliometric analysis identified key AI assessment frameworks that attempt to address trustworthiness, fairness, bias mitigation, and socio-ethical compliance such as the Assessment List for Trustworthy Artificial Intelligence (ALTAI)[11], developed by the EU Commission's High-Level Expert Group on AI. Another relevant framework is the IEEE P7003 Standard for Algorithmic Bias Consideration [12], which focuses on detecting and mitigating bias in AI models. However, these guidelines remain generic and do not provide sector-specific support. In this work we present the Trustworthy AI Assessment Support Design Framework or TAI-SDF, which aims to fulfil the gap by providing a structured, evidence-based methodology for evaluating fairness, transparency, and regulatory compliance in AIED to ensure they are:

- Legally compliant (aligned with the EU AI Act and frameworks like the GDPR);
- Ethically robust (addressing fairness, trustworthiness, explainability and security); and
- Technically sound (incorporating robustness, reliability, and privacy safeguards).

The original contribution of our work is two-fold. We introduce the TAI-SDF framework as a multidisciplinary evaluation tool. Second, we apply it to two real-world case studies using an AI-based assistant to analyse developer documentation and identify potential gaps in fairness and data security in such systems. The first case study has been conducted on Thrively [13], a platform that assesses student behaviour, socio-emotional being and academic performance, and the second on Century Tech [14], an adaptive learning system that personalises educational content based on student performance. This work contributes to the ongoing development of ethical, legal, and technical standards in AIED.

2. Methodology

2.1. Legal Framework for Trustworthy AI

Under the EU´s AI Act (2024), AIED are classified 'high risk' if used to determine admission, evaluate learning outcomes, assess appropriate education levels, and monitor behaviour (Annex III, Section 3 [1]). Given their impact on students' rights and futures, these systems must operate transparently, lawfully, and fairly. The TAI-SDF supports such evaluations using a checklist drawn from the ALTAI principles, which categorises these requirements under seven principles, namely: human agency and oversight; technical robustness and safety; privacy and data governance; transparency; diversity, non-discrimination and fairness; societal and environmental wellbeing; and accountability.

2.1.1. Who must comply?

Responsibility lies with (1) AI providers—persons or entities that develop an AI system and place it on the market or put into service (e.g., Century Tech, etc.) (Art 3(3) [1]); and (2) AI deployers—persons or entities that use an AI system placed on the market (Art 3(4) [1]) such as schools using an AI-powered adaptive learning system like Thrively.

2.1.2. What does compliance require?

AI providers and deployers must document trustworthiness measures related to design, deployment, and standards. TAI-SDF supports automated analysis of this documentation using LLMs by flagging gaps across legal, ethical, and technical dimensions. Users can supplement automated outputs with manual inputs where needed. For example, a university developing and deploying an AI admissions system would need to assess transparency, fairness across socioeconomic groups, and appeal mechanisms. While some exceptions exist (e.g., for research-only systems), compliance becomes mandatory once systems enter the market. In a word, preparatory alignment with legal requirements becomes mandatory.

2.2. Technical approach to assessing AIED trustworthiness: TAI-SDF

Developed under Horizon Europe 'TRUSTEE' project, TAI-SDF formalizes and streamlines human-centric trustworthiness assessments for AI systems. Based on ALTAI principles and other international standards, the framework is structured around four core components:

- Personas – Different users who interact with TAI-SDF:
  - Model Provider (builds AI systems)
  - Deployer (uses AI systems in real-world settings)
- Aspects of AI trustworthiness considered:
  - Privacy, security, fairness, robustness and explainability
- Lifecycle phases:
  - Scope and plan to understand the interfaces, types of AI feedback, and gather requirements in compliance with regulatory and ethics principles
  - Data and algorithm management
  - Verification and validation to test the system functionalities and security features
- Evidence collection for compliance:
  - Documentation and structured user prompts (tasks, constraints, examples)

2.3. The TAI-SDF as an assessment tool

TAI-SDF is not only a theoretical framework and a set of questionnaires covering different aspects and phases of AI products; it has been implemented as a suite of practical software tools that support the actions associated with the obligations of the AI Act for the High-Risk AI Systems (such as the AI based education applications). By combining structured questionnaires with AI-driven analysis, TAI-SDF enables compliance verification for AI systems in education and beyond. Its building blocks include: a questionnaire-based assessment, which allows the user to create questionnaires related to the following aspects: fairness, privacy, security, social impact and explainability; an AI-driven assistant that supports users in assessing compliance of documentation with trustworthiness questions and requirements. In particular, TAI-SDF tools support the Case Studies in the following:

- A set of questions are created with the Questionnaire based assessment;
- Deployer documentation is loaded on the AI based assistant;
- The AI assistant provides answers/evidence related to the questions previously created.

2.3.1. Questionnaire based assessment

The questionnaire-based assessment within TAI-SDF is a qualitative and evidence-based software tool designed for self-assessment (for AI developers) and third-party evaluations (for deployer, institutions or regulators). Specifically, the Model Provider Self-Assessment allows the AI developer to perform a Trustworthiness Analysis during the AI design and development lifecycle. This requires access to the source code, design documentation and test results. The Deployer Assessment allows evaluators or end-users to carry out a third-party assessment, relying only on the AI system user experience and any documentation that is shared by the Model Provider. From a deployer perspective (i.e., a school), TAI-SDF provides the capability to access predefined questionnaires compiled by domain experts or create new reusable ones. Both can benefit from the additional support of the AI driven assistant, which is described next in 2.3.2.

2.3.2. AI-driven assistant

The TAI-SDF framework includes an AI-driven web service which allows the user (e.g., a school) to assess the completeness of product´s (technical) documentation against a set of requirements that can be retrieved from the TAI-SDF knowledge database itself, via an intuitive and essential user interface. Moreover, such a tool can assist the user with generating trustworthiness-related product documentation if missing in the first place. Writing software documentation compliant to AI trustworthiness standards must consider some typically overlooked considerations [15] which this tool aims to spot. Users can upload and edit one or multiple PDF files, refine them, and build structured prompts based on prompt engineering best practices to maximize the effectiveness of AI analysis [16]. This set-up aligns with advances in few-shot prompting [17], zero-shot reasoning [18] and retrieval-augmented generation for knowledge-intensive tasks [19]. While the use cases shown in the present work rely on publicly available documents, in real-world scenarios, third-party deployers should have access to provider documentation for a more thorough assessment. To input documentation and requirements, the user prompt includes:

- Tasks: Instructs the AI model on what should be done;
- Constraints: Instructs the AI model on the required response format;
- Examples: Shows the model some examples to enhance its reasoning capabilities;
- Additional Information: provide contextual details to enhance assessment accuracy.

The AI-driven assessment can be conducted using a locally hosted private model (e.g., Ollama) for data security, or remote LLM services, (e.g. OpenAI), for scalability and ease of access. The result is a set of reports detailing compliance based on the available documentation against each of the selected trustworthiness requirements. Gaps in AI transparency, fairness and security policies are also generated as a result. Two examples of educational software documentation assessment are shown in Sections 3.1 and 3.2.

3. Results: Fairness and Security evaluations using TAI-SDF

In this paper we aimed to demonstrate a novel tool for assessing trustworthiness and fairness of AIED. Theoretically, the tool is informed by the new European legislature [1] and integrated with innovative technological solutions. Its applicability is demonstrated in the unique domain of education, where AI is increasingly normalised in everyday operations. To demonstrate the practical application of TAI-SDF, we conducted two case studies, as previously mentioned, with AI-powered educational technologies: Thrively [13] and Century Tech [14]. According to the provisions of Annex III of the AI Act [1], each of these products would qualify as high-risk. Thrively uses algorithmic assessments to evaluate students' socio-emotional being and academic performance, while also recommending career and academic pathways. In short, its algorithms can directly impact a student's educational and vocational trajectory. Similarly, Century Tech can be classified as 'high-risk' since it uses algorithms to evaluate learning outcomes and impact the students´ rights.

3.1. Case study 1: Fairness assessment of Thrively

3.1.1. Case Study Procedure

A fairness questionnaire (10 key questions) was generated using TAI-SDF. These focused on bias detection, fairness metrics and transparency in Thrively's AI system. The AI-driven tool analysed Thrively's publicly available documentation to determine whether fairness principles were addressed. The obligation to examine potential biases and to implement mitigation measures stems from Art 10(2)(f) and (g) AI Act and the Diversity, Non-discrimination and Fairness requirement of the ALTAI principles [11].

Table 1
Fairness Assessment

Questions	Answers (before fairness review)	Answers (after fairness review)
Does it mention fairness metrics?	No evidence of fairness metrics or bias mitigation strategies	"Yes, Disparate Impact Discrimination Index (DIDI)"
Does it explicitly state the consequences of AI-based children profiling?	No mention of the consequences of children profiling.	"Yes, possible negative outcomes on the students resulting from AI-based profiling are explicitly stated."
Does it address bias mitigation?	No mention of strategies to prevent Ai discrimination	No, still lacks description of bias detection and mitigation
Does it provide transparency about AI-derived decisions?	Lacks explanation of how student assessments are made	No, still lacks decision-making process and criteria

3.1.2. Findings and analysis

Can TAI-SDF support the Deployer Assessment? Yes, but with limitations. As described in the previous sections, the aim has been to assess the completeness of the documentation covering the given topic. The fairness assessment was based on publicly available documentation. The absence of developer documentation, model details or dataset descriptions prevented a fuller fairness evaluation, as shown in Table 1. An additional section, tailored to fairness assessment, has been added to the available documentation to show how the same questions can be compared against an updated documentation. Such documents are scanned by the natural language model wrapped by the tool to assimilate their content and assist the final user to check or generate product documentation by providing assessment reports analysing their semantic coverage of the TAI-SDF questions. The assessment results for each question, embedding the reasoning procedure of the model, are shown in the form of a long and verbose output text file [20].

3.1.3. Results of AI-driven assessment tool

The AI-driven assessment tool achieved a 100% true negative rate (TNR), meaning, the considered documentation does not address any of the considered TAI-SDF fairness questions. After fairness updates of the available documentation, TAI-SDF was able to detect fairness compliance, which confirmed its ability to enhance documentation quality when supported with relevant information. This simple comparison between the original true negative (from documentation without any fairness coverage) and the new true positive (from fairness-tailored simulation) shows that the approach of using a reasoning model to scan documentation for weak points (i.e. nonconformities with respect to TAI-SDF knowledge base) can assure a complete view on the potential strengths and vulnerabilities of the current documentation. Additionally, the tool offers detailed explanations of the reasoning process (due to limited space, full findings can be found at [20]), which make the assessment assistance process transparent and objective, instead of serving as a mere black-box tool. Crucially, TAI-SDF successfully identified fairness gaps in the original documentation. Although assessment is limited when fairness-

Table 2
Security Assessment

Assessment criteria (Questions)	Results (Answers)
Does it mention retention time of users data?	Yes, both retention time and policies are explicitly stated.
Does it mention encryption for data security?	No, doesn't mention encryption, pseudonymisation or anonymization
Does it describe privacy-by-design measures?	No, lacks detailed technical safeguards beyond claims about general compliance
Does it address compliance with AI Act, the GDPR or similar?	No, it states compliance but lacks details around how these are achieved.

related disclosures are missing from AI providers, with updated documentation, AI models could confirm fairness compliance.

3.2. Case study 2: Data Security of Century Tech

3.2.1. Case Study Procedure

The second case study focused on the data security assessment of Century Tech (CT), chosen due to the relative availability of its public documentation. Again, the assessment involved two steps: generating 13-question security questionnaire with TAI-SDF—covering privacy-by-design, encryption, and anonymisation—and using the AI assistant to analyse CT's public privacy policy for compliance. Each question was processed against the available documentation [20], in line with ensuring compliance with the Technical Robustness and Safety requirement of the ALTAI principles [11] while the corresponding provision of Art 15 AI Act obligates the developer and deployer of high-risk AI systems to provide a "appropriate level of accuracy, robustness, and cybersecurity, and that they perform consistently in those respects throughout their lifecycle."

3.2.2. Findings and Analysis

Can TAI-SDF support the Deployer Assessment? Yes, as in Thrively's case, however with limitations due to the limited privacy and security documentation available, as highlighted in Table 2. Since CT's developer documentation (regarding their AI models) was unavailable, the assessment was based on their security and privacy policy. Using what was available, TAI-SDF effectively identified privacy gaps (e.g., missing details on encryption, anonymisation, risk mitigation). Trustworthiness concerns were raised due to unclear security protocols and lack of explicit privacy-by-design measures.

3.2.3. Results of AI-driven assessment tool

A score of 100% for the TNR, the precision and the recall metrics have been obtained. The small number of true positives with respect to the number of true negatives shows that the considered CT privacy documentation does address some of the TAI-SDF privacy requirements 'as-is', but it can be definitively improved or integrated. In other words, two security elements were correctly identified (as a basic compliance mention); 11 security gaps were confirmed which includes missing encryption and details around data privacy. Lastly, no information relating to the AI functions of CT suggests that much improvement is needed to achieve transparency, fairness and trustworthiness.

4. Findings and conclusions

The integration of AI into education introduces high stakes for fairness, trustworthiness, security and privacy for students, teachers, and society at large. This calls that key stakeholders urgently address measures to not only assess fast-encroaching AIED across fairness and trustworthiness principles but ensure that they fully meet them. This paper explored how the TAI-SDF framework—an AI trustworthiness evaluation tool—supports legal and ethical compliance, technical assessment, and contextual application in AI-powered educational technologies. Trustworthy AI systems in education require a multifaceted approach where legislation, technology, and contextual considerations are aligned. TAI-SDF is built on a multidisciplinary knowledge base with legal, ethical, industry, and academic experts, to align with current standards and technical best practices. The TAI-SDF framework shows its capabilities to support both model providers and deployers in assessing various aspects of the trustworthiness of AIED during their full lifecycle. The introduction of a local AI-based assessment assistant boosted by reasoning capabilities helps both self- and third-party assessors in spotting current documentation weak points and evaluating new documentation versions, all while fully demonstrating the logical process and guaranteeing no external leakage of product documentation. It is expected that this holistic and domain-agnostic approach can help both technical and non-technical professionals in a deeper understanding of the key concepts of trustworthiness of complex systems, improve the overall quality of the documentation and streamline the development of fair, ethical, and secure AI systems. By applying TAI-SDF to two real-world AI products-Thrively (for fairness in student assessment) and Century Tech (for data security and privacy in adaptive learning), we demonstrated both the potential and challenges of AI accountability measures in education. In Thrively's case, TAI-SDF successfully identified missing fairness metrics in the company's original documentation. Once fairness-related details were added, the AI assistant detected and validated fairness principles confirming that TAI-SDF can enhance transparency when AI providers disclose relevant information. For Century Tech, the tool flagged missing security and privacy details such as encryption methods and privacy-by-design practices. These flags should inform companies of the key actions to ensure their systems adhere to legal and socio-ethical standards and prevent the risk of harm in education. The findings have also demonstrated that TAI-SDF can help AI system providers and deployers assess and document the fairness of their AI systems. This structured assessment supports compliance with the AI Act, particularly Articles 10 and 11, by providing evidence of how fairness-related requirements are addressed and ensuring comprehensive technical documentation of such evidence. In doing so, the tool also aligns with the ALTAI principles, facilitating transparent documentation and enabling organisations to demonstrate that their AI systems uphold fairness and accountability standards. That said, a key limitation of this work is TAI-SDF's reliance on available documentation. The framework and its tools can only assess what is documented; if AI providers fail to disclose fairness or security measures, a complete evaluation is not possible. While TAI-SDF helps deployers, end-users and even regulators flag compliance gaps, this does not guarantee that an AI system is fair or trustworthy. A full audit remains essential, which necessitates direct access to internal development documents (e.g., source code, test reports, datasets diversity, etc.). Future work should focus on AI providers' willingness to adopt structured documentation formats that support assessments of their systems' fairness, trustworthiness, and security. Moreover, any future work shall also augment the relevance and robustness of this study by validating the results of these tools. Finally, TAI-SDF is adaptable to domains beyond education. Improving the AI model's ability to interpret loosely structured or incomplete documentation is also necessary for better real-world assessments. Ultimately, the work around TAI-SDF along with AI systems in or outside the education domain will encourage AI developers to align with AI regulatory frameworks and prioritise their systems' transparency and trustworthiness.

Acknowledgments

This work was produced under the European Project TRUSTEE (No: 101070214).

Declaration on Generative AI

During the preparation of this work the authors used a small, locally-hosted, pretrained LLM, LLama3.1 8B Instruct. The LLM has been used in order to:

- analyse the documentation to determine whether fairness principles were addressed (see Case study 1);
- and analyse the documentation to determine whether security principles were addressed (see Case study 2);

Afterwards, the authors reviewed and edited the content as needed and take full responsibility for the paper's content.

References

[1] Ai act regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence, 2024. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng.
[2] R. Luckin, Designing educational technologies in the age of ai: A learning sciences-driven approach. british journal of educational technology, 50, 2824–2838., 2019. URL: https://doi.org/10.1111/bjet.12861.
[3] N. Selwyn, The future of ai and education: Some cautionary notes. european journal of education, 57(4), 531–540., 2022. URL: https://doi.org/10.1111/ejed.12532.
[4] Guidelines to help teachers address misconceptions about artificial intelligence and promote its ethical use, 2022. URL: https://ec.europa.eu/commission/presscorner/detail/en/ip_22_6338.
[5] M. P. Lin, Exploring inclusivity in ai education: Perceptions and pathways for diverse learners. in sifaleras, a., lin, f. (eds.), generative intelligence and intelligent tutoring systems. its 2024. lecture notes in computer science (vol. 14799). springer, 2024. URL: https://doi.org/10.1007/978-3-031-63031-6_21.
[6] H. C. Chu, Roles and research trends of artificial intelligence in higher education: A systematic review of the top 50 most-cited articles. australasian journal of educational technology, 38(3), 22, 2022.
[7] C. M. Pontikas, A map of assistive technology educative instruments in neurodevelopmental disorders. disability and rehabilitation: Assistive technology, 17(7), 738–746., 2022.
[8] B. Prenkaj, A survey of machine learning approaches for student dropout prediction in online courses. acm computing surveys, 53(3), article 57., 2020. URL: https://doi.org/10.1145/3388792.
[9] V. Hillman, The state of cybersecurity in education: Voices from the edtech sector. lse media communications department working paper series., 2022. URL: https://www.lse.ac.uk/media-and-communications/assets/documents/research/working-paper-series/WP72.pdf.
[10] C. Dutton, Are colleges' predictive analytics biased against black and hispanic students? the chronicle of higher education., 2024. URL: https://www.chronicle.com/article/are-colleges-predictive-analytics-biased-against-black-and-hispanic-students.

[11] E. Commission, The assessment list for trustworthy artificial intelligence (altai)., 2020. URL: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.
[12] A. Koene, Ieee p7003tm standard for algorithmic bias considerations. 2018 ieee/acm international workshop on software fairness (fairware), 38–41, 2018. URL: https://doi.org/10.23919/FAIRWARE.2018.8452919.
[13] Thrively, 2025. URL: https://www.thrively.com/.
[14] Century tech, 2025. URL: https://www.century.tech/.
[15] F. Königstorfer, Software documentation is not enough! requirements for the documentation of ai. digital policy, regulation and governance, 23(5), 475–488., 2021. URL: https://doi.org/10.1108/DPRG-03-2021-0047.
[16] J. Wei, Chain of thought prompting elicits reasoning in large language models., 2022.
[17] A. Fisch, Making pre-trained language models better few-shot learners. proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural langua gao ge processing, 3816–3830, 2021.
[18] T. Kojima, Large language models are zero-shot reasoners., 2022.
[19] P. Lewis, Retrieval-augmented generation for knowledge-intensive nlp tasks., 2021.
[20] RINA-C, Results assessing ai trustworthiness of ai systems for education, 2022. URL: https://github.com/EPL-education/Assessing-AIeD.