Book Chapter Template

Designing Trustworthy AI in Higher Education

Sandra Rebholz, University of Education Weingarten, Ostbayerische Technische Hochschule Amberg-Weiden, Germany, rebholz@md-phw.de
Paul Libbrecht, University of Education Weingarten, IU International University, Germany, libbrecht@md-phw.de
Wolfgang Müller, University of Education Weingarten, Germany, mueller@md-phw.de

Abstract

Applying Artificial-Intelligence-(AI)-based systems and tools in the context of higher education imposes many challenges with respect to data privacy and ethics. For example, the EU AI act that was adopted in March 2024, classifies many AI systems used in education as high-risk AI systems. High-risk AI systems must follow a strict set of requirements in order to be used in practice. Beyond the legal obligations, the trustworthy use of AI systems is not yet widespread. There are already approaches for assessing the trustworthiness of AI systems that shall ensure that such systems comply with existing guidelines for ethical AI. In this chapter, we review available design approaches for building trustworthy AI systems and evaluate their applicability in the context of higher education. In the real-life use case of developing an AI-based analysis system for e-portfolios from students in introductory computing courses at university, the existing design approaches are further detailed and adapted to the specific context of higher education. Furthermore, we assess the trustworthiness of the developed AI-based analysis system using the OECD Framework for the Classification of AI systems. Based on the findings, we conclude and recommend a scenario-based design process that helps building trustworthy AI-based systems in higher education.

Keywords: AI in higher education, trustworthiness, assessment, e-portfolio

1. Introduction

Publicly available AI tools are rapidly emerging, and find immediate application in the educational field. This is especially true for Generative AI (GenAI) and corresponding technologies, but there is also an increasing number of initiatives in which independent AI developments are being developed for dedicated use in teaching as well as to support learners and teachers in learning processes. An example for such targeted development, which also motivates and underpins the analysis presented here, is the design and development of AI-based methodologies and tools to support teachers in the assessment of and the formulation of feedback on e-portfolios created by students as academic achievement [1]. Such an application has the potential to ease the assessment and the comparative evaluation

of e-portfolios, which is typically time-consuming and elaborate due to the individual character of students' e-portfolios. However, AI-based assessment of coverage of required topics, depth of treatment of individual topics and reflective linking of different subject areas also requires trust in summarized assessments generated by AI and justifications for corresponding evaluations and reasoning. This concrete example illustrates that AI-based systems may offer many potentials in the field of education, but also raises questions and poses challenges related to risks of AI and specifically to the aspect of trust that need to be addressed. The objective of developing AI-based systems needs to be to develop systems that realize the potential benefits, but at the same make sure that the systems can also be trusted. According to the EU strategy of following a human-centric approach to developing AI systems, trust is the prerequisite for this approach.

Based on the use case of supporting the assessment of e-portfolios by AI-based methods and tools as illustrated above, our work presents an in-depth analysis of how to design trustworthy AI-based systems in higher education. Specifically, the following research questions will be addressed: What are the requirements and best practices for building trust with relevant stakeholders involved in the AI-supported analysis of complex learning artifacts such as e-portfolios? How can trust be deliberately integrated into the design and development process of such an AI-based analysis system?

In order to answer these questions, we proceed as follows. After an outline of the theoretical foundations of trust in general and trustworthiness in the context of AI-based systems, we present a detailed analysis of the specific challenges of applying AI in education. Subsequently, we review existing approaches for building trustworthy AI systems that implement the Trustworthiness-by-Design paradigm. Based on the real-life use case of developing an AI-based system for analyzing e-portfolios from students at university, we adapt and refine these approaches for the context of higher education. The derived scenario-based design and development process is described in detail as well as the evaluation of the developed e-portfolio analysis system using the OECD assessment framework for trustworthy AI. Finally, the resulting findings are critically discussed, and the identified benefits and potential challenges of the proposed scenario-based approach are highlighted.

2. Trustworthiness of AI-based Systems

In a general sense, trust is the belief *„that a person (the trustee) will act in the best interests of another (the truster) in a given situation, even when controls are unavailable and it may not be in the trustee's best interests to do so.“* [2, p. 19]. According to a study conducted by [3], the most important element of trust is reliability and consistency of the trustee, followed by beneficence and transparency. Trust with respect to the use of a product or system can be defined as the *„degree to which a user or other stakeholder has confidence that a product or system will behave as intended“* [4, 3.41]. Drawing on this definition, trustworthiness is the *„ability to meet stakeholders' expectations in a verifiable way“* [4, 3.42]. Consequently, it depends on the context and type of system in order to determine the characteristics that are expected from a system and how to verify them.

In the context of AI-based systems, a variety of principles and requirements underlying trustworthy systems have been identified (e.g. [5, 6]). Trustworthy AI-based systems comprise three components: they are lawful, ethical, and robust [5]. Based on these components, the higher-level expert group on AI (AI HLEG) of the

European Commission defined a set of requirements that need to be fulfilled by a system in order to be considered trustworthy. These requirements include human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, societal and environmental wellbeing, and accountability [5]. Similar principles can be found in the Recommendation on the Ethics of Artificial Intelligence adopted by the UNESCO in November 2021 [6]. Here, the need for creating awareness and literacy among the wide public is explicitly stated and also the aspect of multi-stakeholders in AI governance is included in the principles.

In order to evaluate whether a given AI-based system adheres to the requirements of trustworthy AI, a standard set of criteria needs to be established and assessed for a given system. The assessment list by the AI HLEG for instance formulates questions for all aspects relevant to Trustworthy AI [7]. The list is intended to be used as a self-assessment list for evaluating the trustworthiness of an AI system's design, development and usage. It is important to note that not only the final system is assessed, but also the design and development process to build the AI system. Applying the assessment list in a certain domain, also requires to adapt or add elements to the list to make it fit to the specific requirements of the application context.

As an early pioneer in the field, the OECD AI group embraced the task of qualifying artificial intelligence systems so as to make them scrutinizable by policymakers. The working group defined in 2019 the AI principles which define the first policy directions needed for to create a trustworthy AI (e.g. *Shaping and enabling interoperable governance and policy environment for AI (Principle 2.3): Governments should create a policy environment that will open the way to deployment of trustworthy AI systems)*. The OECD framework for the classification of AI systems [8] presents a series of assessment questions, each of which contributes to the principles in a documented manner. This set of mostly qualitative questions give hints of how the principles are approached.

Going beyond the OECD framework, the NIST AI risk management framework [9] describes a process with more concrete steps to manage the risks so as to adhere to the principles and thus "guarantee" trustworthiness. The approach is even more operational with concrete indications on metrics, risks, playbooks, and processes.

The need for trusting artificial intelligence has grown explosively with multiple other initiatives offering approaches to evaluate (e.g. the z-inspection[1]) or to certify (e.g. the 1EdTech's TrustEd apps program[2]).

With the goal of supporting the development of trustworthy AI, the European Union adopted in May 2024 the Artificial Intelligence Act (AI Act [10]) which is the first legal framework for regulating the use of AI-based systems. The framework takes a risk-based approach and applies the classification of the OECD summarizing an AI system into four levels of risk. On the top level, systems that pose a clear threat to people and businesses are categorized as systems with unacceptable risks. The development and use of such systems are forbidden. On the next level, systems are categorized as high-risk systems. They fall under the provisions of the AI act and have to comply with strict requirements for developing, deploying, and using them.

[1] The z-inspection is an evaluation process for the trust of artificial intelligence systems, it is piloted by a non-profit association of scholars. See https://z-inspection.org.

[2] The 1EdTech TrustEd Apps program offers a management solution to educational institutions https://www.1edtech.org/program/trustedapps.

Examples of high-risk AI systems include AI-based exam scoring systems, robot-assisted surgery systems, and AI-based credit scoring systems among others. Limited-risk systems are grouped on the third level and have to adhere to certain transparency regulations. Minimal risk systems can be used without any limitations.

Despite the many guidelines, recommendations and regulations, Trustworthy AI is not the same as trusted AI [11]. Recent investigations have shown that applying up-to-date guidelines and metrics for trustworthiness does not lead to an increase in actual trust in AI systems. The authors argue that public attitudes are largely built upon the perceived trustworthiness of an AI application which in turn is influenced by typical constructs of technology acceptance such as perceived ease of use and perceived usefulness, as well as the attitude towards AI in general.

3. Challenges of Applying AI in Education

From the very beginning developments in Artificial Intelligence depicted links to the field of education, and research and developments in AI were transferred to the field of education and stimulated new research approaches in the field of education [12]. Consequently, there is a long history of applications of AI-technologies in education, and AI has been linked to numerous potentials and benefits in education (e.g., [13, 14, 15]).

Educators tend to quickly adapt all types of new technologies to enrich teaching and learning, also ones not specifically targeted to the field of education. This also applies to novel AI-related technologies. Against this background, it is not surprising that approaches to describing and classifying forms and scenarios for the use of AI in education have to be incomplete and limited. Typical approaches to classify the use of AI in education (AIED) distinguish between a) Student-focused AIED, Teacher-focused AIED and Institution-focused AIED [15, 16].

Scenarios and corresponding technological approaches in the context of Student-focused AIED include in particular personalized learning and Intelligent Tutoring Systems, while Teacher-focused AIED scenarios are often related to automatic assessment of students' learning and support in providing adequate feedback. A typical objective of Institution-focused AIED is the identification of dropouts and students-at-risk. In addition, AI-related competences and skills are considered as an important aspect, both for students and teachers. On the students' side, these are understood as a prerequisite for the effective and reflective use of generative AI (GenAI) technologies, while they are also considered a specific learning objective of AI-enhanced learning scenarios. Similarly, corresponding competencies are considered indispensable for teachers to effectively apply AI technologies in the classroom, but also for teaching fundamental skills and fostering competencies related to AI and AI technologies. Yet, in both cases the characteristics and the extent of such competencies are still objects of scientific discussions.

There is currently a consensus of opinion that such applications of AI technologies in the classroom and for learning and teaching also come with risks and challenges. This is based on general ethical concerns and requirements [18, 16]. The Beijing declaration [19] represents the first approach to list challenges and formulate policy recommendations specifically targeted to the field of AIED.

Recently, challenges for the application of AI in education were raised in a number of publications, in some cases on a more general level, in others on a more detailed one, and closely related to specific AI technologies, such as generative AI (e.g., [20, 21]).

Many of the raised challenges may be related to the aspects of trustworthiness of AI technologies, but also to trust in the use of humans [17]. Specific concerns may be

related to aspects such as privacy and security, quality and effectiveness of AI tools, trust in presented results (e.g., with respect to possible algorithmic bias), and equity in access. For instance, in Institution-focused AIED targeted to identify possible dropouts and at risk-students, the corresponding AI system requires trust in the assessment of individual students, providing a sufficient degree of transparency on how the decision was made. At the same time, privacy must be respected.

4. Trustworthy AI by Design

Despite the availability of ethical guidelines for trustworthy AI, there seems to be a gap between defining general guidelines and actually putting them into practice [22]. As the guidelines need to be applied during the whole engineering process of AI-based systems and also when deploying and using them, the design and development process of AI-based systems needs special consideration. There are various approaches on how to design and develop trustworthy AI-based systems. In the following, we present approaches that take a holistic view on developing AI-based systems and that realize a *Trustworthiness-By-Design* paradigm. All outlined approaches target and include trustworthiness as a core element of the design and development process from the beginning.

In a collection of 62 *Responsible AI Patterns*, the book [23] describes best practices in the form of solution templates for coping with the challenges associated with the design, the implementation, and the management of AI-based systems. The patterns are grouped into three categories related to product, processes, and governance considerations. Depending on the context of the application at hand, these patterns can be reused and adapted to the specific requirements of the respective domain.

In order to establish a trustworthy and responsible AI development process, [23] identifies the potential issues that can arise in the individual stages of the software development process. For each issue, the authors propose a solution to specifically address and mitigate the identified problems and risks. As an example, in the requirements phase, it is essential to collect, elicit, and document requirements with respect to trustworthy AI. As a solution, so-called *Responsible AI User Stories* can be introduced as a new type of user story. Based on predefined templates and guiding questions, the user stories are defined and worked on as part of the product backlog in an agile project. Another example on how to integrate ethical considerations in the design process, is the use of envisioning cards [24] in order to strengthen awareness and reflection on how AI systems may impact human values. Envisioning cards focus on four envisioning criteria, namely stakeholder, time, value, and pervasiveness. Each card describes a specific concept related to these criteria, and suggests design activities to initiate discussion and engagement with possible effects and implications of AI-based systems with respect to this concept.

In addition to general best-practice guidelines, there are also company-specific approaches that are published and used in practice. Examples are the Responsible AI practices recommended by Google [25] which emphasize a human-centric design approach and the importance of testing activities, the Guidelines for Human-AI Interaction [26] developed by Microsoft Research with focus on user interface design of AI-based applications, and the IBM Design for AI [27] which explains the rationales and driving forces underlying the design of AI systems.

5. Use Case: AI-based Analysis of E-Portfolios

In the following, we present a real-life example of how AI-based technology and tools can be applied in higher education. The application performs an AI-based analysis of e-portfolios and shows how both teachers and students can benefit from using such tools in teaching and learning.

E-Portfolios are collections of digital artifacts that students create for documenting their individual learning. In the e-portfolio, they present individual project results, summarize learning content, and reflections on the learning process and goals they have achieved. E-portfolios are similar to online blogs that contain a variety of multimedia content and can be highly personal. In the context of higher education, e-portfolios are generally used as a competency-based learning tool, but also as a means to perform holistic assessments of the learning process and learning outcomes. At the University of Education Weingarten, e-portfolios have been used in introductory courses in computer science and learning technologies for over ten years.

5.1 Teaching and learning scenario based on e-portfolios

A typical scenario on how to integrate e-portfolios in university courses is as follows. Students take part in the lecture and are encouraged to deepen the learning content independently. They choose their own focus topics and work on these independently. This includes researching relevant information as well as carrying out small projects to apply what they have learnt in practice. Students document the entire learning process, rephrase the knowledge they have synthesized and the developed content in their personal e-portfolio. They can share their e-portfolios: It is up to the students to decide who they grant access to the e-portfolio. By doing so, they can receive feedback on the e-portfolio presented in the composition system from their fellow students or the teacher, and use the feedback to improve the e-portfolio. At the end of the semester, students submit the completed e-portfolio. The composition system is, in the case of the University of Education Weingarten, the widely used Mahara platform.

Latest at the end of the semester, the teachers assess the e-portfolio based on predefined criteria. The assessment is typically done based on rubrics [28]. In the rubrics, all relevant assessment criteria are listed along with a description of different performance level characteristics. Figure 1 shows an extract of an example of a rubric for e-portfolio evaluation.

Criteria	Beginner	Intermediate	Advanced	Proficient
Complete presentation of relevant concepts	A subset of the relevant concepts are described, and/or relevant concepts are partly described.	Relevant concepts are stated. Descriptions are taken from the available learning materials.	Relevant concepts are described and partially enhanced with additional materials and with explanations in own words.	All relevant concepts are described technically correct and in sufficient detail. Completely independent work.
Indepently created artifacts	Artifacts (graphics, code extracts etc) are taken from the available learning materials.	Some independently created artifacts are presented. Artifacts show/apply basic concepts and their relationships.	Independently created artifacts are presented. Artifacts apply advanced concepts and their relationships.	Independently created artifacts are presented. Artifacts are fully elaborated and described proficiently.
Appropriate use of digital media	The portfolio contents is mainly presented in text format.	Some media artifacts are integrated. The artifacts are selected and included based on the thematic context.	Various media artifacts are integrated that illustrate the presented contents and enhance the understanding of the contents.	Various media artifacts are judiciously selected, well elaborated and provide new perspectives on the underlying contents.

Figure 1: A part of the rubrics to evaluate e-portfolios.

5.2 AISOP: AI-supported observation of e-portfolios

In the AISOP project, we have developed a web application that carries out an AI-based analysis of the e-portfolio contents. Every time the user accesses their e-portfolio in the composition system, they can request an automatic analysis and see the result of this analysis in the AISOP web application. The web application employs thematic classification and concept maps to allow for an interactive concept-based coverage analysis and navigation as depicted in figure 2. It will also provide different perspectives on the e-portfolio contents based on linguistic characteristics such as text complexity, lexical variety, or coherence (see [29]).

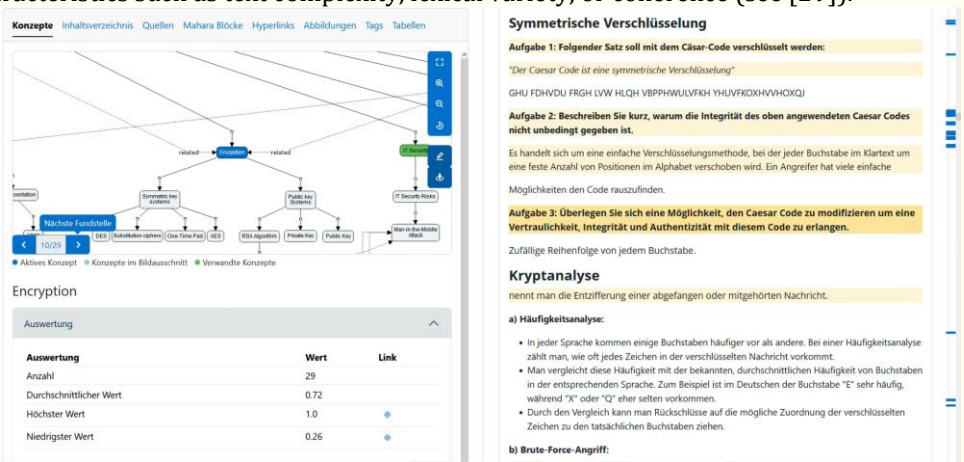

Figure 2: Concept-based navigation: An interactive concept-map generated by the result of the classification with its corresponding e-portfolio navigation.

5.3 Design and development process

The AISOP web application has been designed and developed using the scenario-based design approach as proposed by [30]. In the design process, various scenarios have been developed that illustrate the main usage scenarios for the analysis system from the perspective of the target users (see [31] for example scenarios). The resulting scenarios are the basis for system design and development, and have been used to derive test scenarios for evaluating the application in a real-life context at the university. The evaluations yielded a number of experimental results as presented in [32].

The project giving birth to the AISOP system was created so as to offer a reproducible approach which can be summarized by the following steps to obtain a similar system in other teaching opportunities based on e-portfolios (the AISOP "recipe"):

- Formulate **proposed scenarios** of use that reflect the concrete teaching situation at hand. Make sure to consider all aspects that encourage a trusting use of the web application (e.g. inspired by the key questions in the assessment list by the AI HLEG [7]).
- Have an e-portfolio **composing system ready** for the students including the possibilities of sharing with selected users.
- Make sure the e-portfolio composing system can be **interfaced** to the AISOP web app (this may need to configure web services, authorize them, or write custom interfaces). This is the step where the users will express their authorizations and thus express their trust. Thus, a clear scenario is useful to envision the trust of the authorization.

- **Identify the courses** where this is to be applied. Create concept maps for representing the knowledge domain of each course (e.g. using CmapTools[3]). Scenarios of usage of the e-portfolios in the course of a term should be available.
- **Collect** textual materials relevant to the course content such as course slides or earlier e-portfolios and make sure you are allowed to process them. This processing is necessary to generate training data for the natural language processing (NLP) pipeline incorporated in the e-portfolio analysis system. It is an internal process and can be made with protected content (copyrighted, personal-data…).
- **Extract** all the relevant text fragments within text-files (e.g. using a clipboard tracker[4]).
- Perform the manual **annotations** of the topic classification of all the fragments (e.g. using explosion AI's prodigy[5]).
- **Train the text classifier** and refine the training. This creates a classification model specialized to the course learning content (e.g. using explosion AI's prodigy).
- **Install and configure** the classification model as well as the concept map as a new course in the AISOP web application (see web-application documentation[6]).
- **Test** the system implementation based on the **proposed scenarios of use** (see step 1). Assess whether the criteria for trustworthiness are met or whether the system needs to be optimized.

All e-portfolios of the newly integrated course can now be analyzed and visualized in the web application by any user who has access to the e-portfolio composition system.

The approach applies fairly generic tools (such as the topic classification) and manages a pool of data so as to train the classifiers, one of the cornerstones of the machine learning approach to developing artificial intelligence tools. As depicted in the recipe, the process starts by defining appropriate usage scenarios considering criteria for trustworthiness and ends with a practical evaluation of the solution based on the scenarios defined beforehand. If the evaluation results do not meet the defined criteria, a new development cycle will be initiated.

This approach was elaborated and experimented in the AISOP project. Among the experiments, we ran several rounds of marking supported by the AISOP tool so as to elucidate how the tool could support the teacher's review process. Some of these experiments and their results are described in [32] and the papers cited therein. Another round of experiment is in preparation where interpretation of the students of the coloured topic maps and the induced navigation, which are a way to present the output of the text-classification, is in the focus.

5.4 OECD Assessment of the AISOP AI Service

[3] CmapTools is a freely available concept-maps editing system, https://cmap.ihmc.us.

[4] Such a clipboard tracker is available open-source at https://gitlab.com/aisop/aisop-hacking/-/tree/main/aisop-clipboard-extractor.

[5] prodigy is a commercial tool to support the annotation of texts for several classical NLP tasks, see https://prodi.gy/.

[6] The AISOP web application is available open-source at https://gitlab.com/aisop/aisop-webapp.

Before discussing the ethical and trust-building aspects of the AISOP approach, we first take the time to assess it as an artificial-intelligence application according to OECD which "provides a structured way to assess AI systems' potential to promote the development of human-centric, trustworthy AI" [8]. The complete assessment is in Annex 1.

The assessment covers the five key dimensions: People & planet, economic context, data & input, AI model, task & output. It qualifies the trustworthiness of the AI system embedded as a web application service. Being a system made for supporting the learning, the usage of the AISOP system carries core dimensions which can be reformulated as in the figure 3.

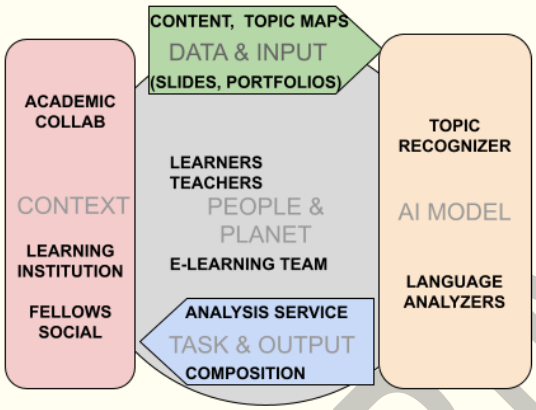

Figure 3: The cycle of usage of the AISOP web-application, adapted from the cycle of usage of a generic AI system of [8].

The highlights of this assessment include the following observations:

The service must be seen as a complete service, although the AI component (the topic recognition) is a modest part of the process: Indeed, multiple criteria such as the optionality, the impact on critical processes (such as giving a mark), the agency of persons or the personal nature of data are only valid because of the way they are assessed as a service delivering a visualization.

The transparency of the AI results cannot be offered. This lies in the relatively hidden neuronal network nature of the spaCy models but also in the probably homogeneous nature of the texts used for training the classifier. While transparency could become better defined in the scientific literature, there appears to be no pressing need to offer users a more transparent classification, instead, the paper [32] presents a study about the usability of the interactive concept map and highlights that better integrations could be closer to support the students' in their use.

Rights handling with training corpora for learning relevant material is a multi-facetted process. While a university could be handling with only its internal data, it is difficult for newcomers to embark and the re-use of existing corpora is important. For this reason, reusable corpora (e.g. from course contents and from existing e-portfolios) are desirable. Repositories such as Germany's research data repository in education[7] which allows access to limited researchers' circles can offer a solution. The highly personal nature of e-portfolios makes it that, even if anonymized, a student would recognize their e-portfolio immediately. Thus, sharing in a completely open form is rare and needs an explicit permission. However, there is no

[7] The research data repository in education is accessible from https://www.forschungsdaten-bildung.de/.

risk in terms of possible divulgation of training content in the AISOP case (contrary to the case of generative AI systems).

Finally, it is important to acknowledge that the AISOP recipe involves course-specific data and thus a course-specific classification system. This implies that each application of the recipe operates for its restricted focus. We claim that, beyond the cold-start problem mentioned above, this allows every institution to carry responsibility for the relevance of the AI service, which is a fundamental value of a course and the services around it.

The assessment of the OECD has given us the opportunity to ask ourselves how the software's artificial intelligence dimensions (such as the flow of data and the personal-data-protection aspects) are being monitored. Based on this assessment, one can easily answer the European Union's AI Act's classification of the service [10]. This leads us to evaluate the AISOP web application as a limited risk system for which there may be registration requirements.

6. Making a Trustworthy Process

The proposed recipe to create an artificial intelligence is rather following common steps: It involves the re-use of software, the re-use of data-sets (NLP corpora), the enrichment with context-specific data, and the interfacing so that students submit their own data and obtain, thus, a service powered by artificial intelligence. It can be seen as a typical system without employing large artificial intelligence models of which only a few exist on earth and for which the privacy terms are rarely respectful.

The process can be assessed as trustworthy and respectful of the AI goals of the OECD. That means showing what data is stored where, where it is transmitted, and how it is being analyzed. We claim that the AISOP experience has proven that the use of scenarios makes it clear how a user perceives how their data is exchanged and processed. This is a very important lever to attract trust and is somewhat independent of the assessment of the governance of data and algorithms. However, both are of fundamental importance to be able to offer a trustable service.

As shown in the OECD assessment, it appears just as fundamental to give users freedom of choice as it is to show how the data is exchanged in a transparent and comprehensible manner.

While some of the AISOP experiments have shown very little concern about privacy on the part of the students, the respect for privacy can be the subject of a sudden breach of trust, that would have a fatal impact on the use of a service. Thus, it appears fundamental to be able to express properly which data is transferred? how much? to whom and for which purpose? and to what extent the user is obliged to use the service. Due to the interplay of multiple systems, it is not uncommon that users feel overwhelmed by the selection options and simply click on 'ok' in an authorization dialogue, without actually understanding what they are giving their consent to. But this may stop at any time (e.g. when the news arouses mistrust about a certain aspect, which generally provokes the entire rejection reaction) and only a careful explanation may convince them otherwise. We claim that assessing the trust in workflows through scenarios even before a finalized software is available is an appropriate method to ensure long-term trust and long-term evolution of the software.

7. Conclusion

In this paper we have attempted to define trustworthiness and trustability for artificial intelligence applications based on the definitions of the literature. The wide spectrum of contributions and recommendations that we could encounter have not yet provided methodologies that have proven themselves as applicable in practice for learning systems.

We have described the design and development process that we followed to build an AI-based web application service of which one can assess the trustability. Through the use of scenarios, we have been able to highlight challenging points of trust and on the presentation of what to expect of a system and thus make sure that they are clear to all stakeholders.

In our process, we realized that some uses by students or teachers may have been missed by our scenarios. While it is good for a development to limit its scope, some scenarios are unavoidable as they are fundamental to building trust and some are even an obligation by law. Examples include the scenarios to operate in case of a request to be forgotten (as is a fundamental right) or the reactions as a teacher against fears of using the service (teachers could explain the web application's privacy guarantees better but they could also adjust the configuration).

Among the few discoveries which appeared is the establishment of principles of "who should be able to decide whether an AI system is used to analyze an e-portfolio?": While it appears natural to leave this choice to the authors of the e-portfolios, this is not what is done in practice: Any person who is reviewing an e-portfolio is in a position to submit the content to an AI system. Expressing this possibility (or its prohibition) as a scenario is an excellent way to evaluate its desirability.

Acknowledgments

This research was partially funded by the grant 16DHBKI015 (AISOP) of the German Federal Ministry of Research and Education.

We wish to thank the collaborators of the AISOP team including Thierry Declerck†, Alexander Gantikow, Andreas Isking, Pierre Günthner and Simon Ostermann as well as the many students who helped by lending their e-portfolios.

References

[1] Gantikow A, Isking A, Libbrecht P, Müller W, Rebholz S. On the Creation of Classifiers to Support Assessment of E-Portfolios. International Workshop on Multimedia in Technology Enhanced Learning; Laguna Hills CA; 2023.p. 297-302. DOI: 10.1109/ISM59092.2023.00057

[2] Marsh S, Dibben, MR. Trust, Untrust, Distrust and Mistrust - An Exploration of the Dark(er) Side. In: Herrmann P, Issarny V, Shiu S, editors. International Conference on Trust Management. Berlin, Heidelberg: Springer; 2005. p. 17-33

[3] Slade S, Prinsloo P, Khalil M. "Trust us," they said. Mapping the contours of trustworthiness in learning analytics. In: Information and Learning Sciences, Vol. 124 No. 9/10; 2023. p. 306-325

[4] ISO/IEC. ISO/IEC 24028:2020. Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence. 2020.

[5] HLEG - High-Level Expert Group on Artificial Intelligence set up by the European Commission. Ethics guidelines for trustworthy AI. 2019. Available from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [Accessed: 2024-10-17]

[6] UNESCO. Recommendation on the Ethics of Artificial Intelligence. Adopted on 23 November 2021. Paris: United Nations Educational, Scientific and Cultural Organization; 2022. Available from: https://www.unesco.org/en/articles/recommendation-ethics-artificial-intelligence [Accessed: 2024-10-17]

[7] HLEG - High-Level Expert Group on Artificial Intelligence set up by the European Commission. The Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self assessment [Internet]. 2020. Available from: https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment [Accessed: 2024-10-17]

[8] OECD. OECD Framework for the Classification of AI systems. OECD Digital Economy Papers, No. 323, Paris: OECD Publishing; 2022. DOI: https://doi.org/10.1787/cb6d9eca-en.

[9] NIST. Artificial Intelligence Risk Management Framework (AI RMF 1.0 [Internet]. 2023. Available from: https://airc.nist.gov/AI_RMF_Knowledge_Base [Accessed. 2024-10-06]

[10] AI Act. Regulation (EU) 2024/1689 of the European Parliament and of the Council [Internet]. 2024. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689 [Accessed: 2024-10-18]

[11] Knowles B, Richards JT. ACM TechBrief. Trusted AI. ACM Technology Policy Council, Issue 9, Fall 2023. New York, NY, USA: ACM; 2023. DOI: 10.1145/3641524

[12] Doroudi S. The Intertwined Histories of Artificial Intelligence and Education. International Journal of Artificial Intelligence in Education. 2023; 33(4), 885–928. DOI: https://doi.org/10.1007/s40593-022-00313-2

[13] Schank R., Edelson D. A role for AI in education: Using technology to reshape education. Journal of Artificial Intelligence in Education. 1989; 1, 3–20.

[14] Baker RS. Artificial intelligence in education: Bringing it all together. Digital Education Outlook: Pushing the Frontiers with AI, Blockchain, and Robots. 2021. p. 43–54.

[15] Anderson JR, Kline PJ. A Learning System and Its Psychological Implications. IJCAI. 1979. p. 16-21

[16] Holmes W, Tuomi I., State of the art and practice in AI in education. European Journal of Education. Vol 5 Num 4.; 2022. p. 542-570. DOI: 10.1111/ejed.12533

[17] Vincent-Lancrin S., van der Vlies R. Trustworthy artificial intelligence (AI) in education: Promises and challenges (No. 218; OECD Education Working Paper). OECD. 2020. DOI: 10.1787/19939019.

[18] Coeckelbergh M. AI Ethics. MIT Press; 2020. ISBN 9780262538190.

[19] UNESCO. Beijing Consensus on Artificial Intelligence and Education [Outcome document of the International Conference on Artificial Intelligence and Education 'Planning education in the AI era: Lead the leap']. Paris: United Nations Educational, Scientific and Cultural Organization; 2019. Available from: https://unesdoc.unesco.org/ark:/48223/pf0000368303 [Accessed: 2024-11-25]

[20] Michel-Villarreal R., Vilalta-Perdomo, E, Salinas-Navarro DE, Thierry-Aguilera R, Gerardou FS. Challenges and Opportunities of Generative AI for Higher Education as Explained by ChatGPT. Education Sciences. 2023; 13(9), 856. DOI: 10.3390/educsci13090856

[21] OECD. Initial policy considerations for generative artificial intelligence. Paris: OECD Publishing; 2023. DOI: 10.1787/fae2d1e6-en .

[22] Li B, Qi P, Liu B, Di S, Liu J, Pei J, Yi J, Zhou B. Trustworthy AI: From Principles to Practices. ACM Comput. Surv. 2023; 55, 9, Article 177 (January 2023). 46 p. DOI: doi.org/10.1145/3555803

[23] Lu Q, Whittle J, Xu X, Zhu L, Responsible AI: Best Practices for Creating Trustworthy AI Systems. ISBN: 9780138073947. Addison-Wesley Professional; 2023. 291 p.

[24] Friedman B, Hendry D. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12). New York, NY, USA: ACM; 2012. p. 1145–1148. DOI: doi.org/10.1145/2207676.2208562

[25] Google AI. Responsible AI practices [Internet]. 2024. Available from: https://ai.google/responsibility/responsible-ai-practices/ [Accessed: 2024-09-19]

[26] Amershi S, Weld D, Vorvoreanu M, Fourney A, Nushi B, Collisson P, Suh J, Iqbal S, Bennett PN, Inkpen K, Teevan J, Kikin-Gil R, Horvitz E. Guidelines for Human-AI Interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19). New York, NY, USA: ACM; 2019. Paper 3, 1–13. DOI: 10.1145/3290605.3300233.

[27] IBM. Design for AI [Internet]. 2022. Available from: www.ibm.com/design/ai [Accessed: 2024-09-19]

[28] Brookhart SM, Chen F. The quality and effectiveness of descriptive rubrics, Education Review. 2015; Vol 67, number 3. DOI: 10.1080/00131911.2014.929565.

[29] Günthner P, Rebholz S. Implementierung von Textanalyse in der E-Portfolio-Bewertung. EPEPLA Workshop DELFI 2024. 2024.

[30] Rosson MB, Carroll JM, Usability engineering: scenario-based development of human computer interaction. San Francisco, CA, USA: Morgan Kaufmann; 2002. ISBN 978-0-08-052030-8.

[31] Isking A, Libbrecht P. Szenarien für eine automatische Analyse von E-Portfolios. EPEPLA Workshop DELFI 2024. 2024.

[32] Gantikow A, Durski S, Isking A, Libbrecht P, Müller W, Ostermann S, Rebholz S. KI-basierte Analyse von E-Portfolios. In: Proc. DELFI 2024, Fulda: GI e.V.; 2024.

Annex 1: OECD Assessment of the Artificial Intelligence AISOP Web-Application

People and Planet		
Characteristic	**Question**	**Response**
Users of AI system	*What is the level of competency of users who interact with the system?*	Amateur / Apprentices
Impacted Stakeholders	*Who is impacted by the system?*	Students, Teachers
Optionality and redress	*Can users opt out, e.g. switch systems? Can users challenge or correct the output?*	Usage is optional Correction requires re-running steps.
Human rights and democratic values	*Can the system's outputs impact fundamental human rights?*	No because usage is optional and the information is only indicative for teachers.
Well-being, society and the environment	*Can the system's outputs impact areas of life related to well-being (e.g. job quality, the environment, health, social interactions, civic engagement, education)?*	Enhancement of the review process (support of the education process)
Displacement potential	*Could the system automate tasks that are or were being executed by humans?*	(only enhance the time taken)

Economic Context		
Characteristic	**Question**	**Response**
Industrial sector	*Which industrial sector is the system deployed in (e.g. finance, agriculture)?*	education
Business function	*What business function(s) or functional areas is the AI system employed in (e.g. sales, customer service, human resources)?*	learning, teaching
Business model	*Is the system a for-profit use, non-profit use or public service system?*	part of the teaching process (public/for-profit)
Impacts critical functions / activities	*Would the disruption of the system's function or activity affect essential services?*	no
Breadth of deployment	*Is the AI system deployment a pilot, narrow, broad or widespread?*	narrow (becoming broad)

Data & Input		
Characteristic	**Question**	**Response**
Detection and collection	*Are the data and input collected by humans, automated sensors, both?*	human
Provenance of data and input	*Are the data and input from experts; provided, observed, synthetic or derived?*	yes (collected using, e.g. copy-and-paste)
Dynamic nature	*Are the data dynamic, static, dynamic updated from time to time or real-time?*	No, except enhancements by teachers.
Rights associated with data and input	*Are the data proprietary (privately held), public (no intellectual property rights) or personal data (related to identifiable individual)?*	Fragments need allowance but could be shared in corpora. Fragments can be personal. Anonymized fragments can be used and shared for training.
Identifiability of personal data	*If personal data, are they anonymised or pseudonymised?*	personal or *pseudonymised*
Data quality and appropriateness	*Is the dataset fit for purpose? Is the sample size adequate? Is it representative and complete enough? How noisy are the data?*	according to cross validation: fit for the purpose (~90%). But course specific.
Structure of the data and input	*Are the data structured, semi-structured, complex structured or unstructured?*	semi-structured (lines of texts, annotated)
Format of data and metadata	*Is the format of the data and metadata standardised or non-standardised?*	standardised for the annotations.
Scale	*What is the dataset's scale?*	small

AI Model		
Characteristic	**Question**	**Response**
Model information availability	*Is any information available about the system's model?*	limited (spacy's classification model, probably a simple multi-layer-perceptron
AI model type	*Is the model symbolic (human-generated rules), statistical (uses data) or hybrid?*	statistical
Rights associated with model	*Is the model open-source or proprietary, self or third-party managed?*	inference tools are open-source (spacy), preparation tools are proprietary (prodigy)
Discriminative or generative	*Is the model generative, discriminative or both?*	discriminative (and generates visualizations)

AI Model		
Characteristic	**Question**	**Response**
Single or multiple model(s)	*Is the system composed of one model or several interlinked models?*	one model
Model-building from machine or human knowledge	*Does the system learn based on human-written rules, from data, through supervised learning or through reinforcement learning?*	supervised learning based on an annotated corpus
Model evolution in the field (applicable only to machine-learning systems)	*Does the model evolve and / or acquire abilities from interacting with data in the field?*	no But an actualized run of the recipe can be done at the end of the semester..
Central or federated learning (applicable only to machine-learning systems)	*Is the model trained centrally or in a number of local servers or edge devices?*	centrally (based on existing language models)
Model development and maintenance	*Is the model universal, customisable or tailored to the AI actor's data?*	customizable
Deterministic and probabilistic	*Is the model used in a deterministic or probabilistic manner?*	probabilistic
Transparency and explainability	*Is information available to users to allow them to understand model outputs?*	No. (a score is provided)

Task & Output		
Characteristic	**Question**	**Response**
Task(s) of the system	*What tasks does the system perform (e.g. recognition, event detection, forecasting)?*	topic recognition.
Combining tasks and actions into composite systems	*Does the system combine several tasks and actions (e.g. content generation systems, autonomous systems, control systems)?*	yes (recognition is followed by presentation on the interactive concept map)
Action autonomy	*How autonomous are the system's actions and what role do humans play?*	analysis is on request of the user (reviewer) The system only generates visualizations.
Core application area(s)	*Does the system belong to a core application area such as human language technologies, computer vision, automation and / or optimisation or robotics?*	human language technologies and visualizations
Evaluation methods	*Are there standards or methods available for evaluating system output?*	unclear