An initiative of the

European Commission



Explainable AI in education: Fostering human oversight and shared responsibility

by the European Digital Education Hub's squad on explainable AI in education

EUROPEAN **DIGITAL EDUCATION HUB**

Education and Training

EUROPEAN
DIGITAL
EDUCATION
HUB



Authors:
Francisco Bellas
Jeroen Ooge
Lezel Roddeck
Hasan Abu Rashheed
Marjana Prifti Skenduli
Florent Masdoum
Nurkhamimi bin Zainuddin
Jessica Niewint Gori
Eamon Costello
Lidija Kralj
Deepti Teresa Dcosta
Dora Katsamori
Darren Neethling
Sarah ter Maat
Roy Saurabh
Rena Alasgarova
Elena Radaelli
Ana Stamatescu
Arjana Blazic
Graham Attwell
Giedrė Tamoliūnė
Theodora Tziampazi
Moritz Kreinsen
António José Alves Lopes
Jose Viñas Diéguez
Barbara Loranc
Cristina Obae

The European Digital Education Hub (EDEH) is an online community for practitioners from all sectors of education and training aiming to contribute to improving digital education in Europe. To achieve this goal, EDEH is not only a place for exchange and discussions but also offers a variety of different events and activities. These activities include the squads that are online working groups where community members can collaborate on a specific topic of digital education. This report is the result of the work of the EDEH squad on explainable AI in education.

EUROPEAN
DIGITAL
EDUCATION
HUB

This document has been prepared for the European Commission and for the European Education and Culture Executive Agency (EACEA), however it reflects the views only of the authors, and the European Commission and EACEA are not liable for any consequence stemming from the reuse of this publication.

More information on the European Union is available on the internet (http://europa.eu).

© European Union, 2025



Except otherwise noted, the reuse of this document is authorised under a Creative Commons Attribution 4.0 International (CC-BY 4.0) licence (https://creativecommons.org/licenses/by/4.0/). This means that reuse is allowed provided appropriate credit is given and any changes are indicated.

The reuse policy of European Commission documents (applicable also to documents of the European Education and Culture Executive Agency) is implemented based on Commission Decision 2011/833/EU of 12 December 2011 on the reuse of Commission documents (OJ L 330, 14.12.2011, p. 39).

For any use or reproduction of elements that are not owned by the European Union, permission may need to be sought directly from the respective rightholders. The EU does not own the copyright in relation to the following elements, which are used under their respective licences:

- *Cover page image – © Freepik 2025 I Freepik.*
- *Text of the report – © Francisco Bellas, Jeroen Ooge, Lezel Roddeck, Hasan Abu Rashheed, Marjana Prifti Skenduli, Florent Masdoum, Nurkhamimi bin Zainuddin, Jessica Niewint Gori, Eamon Costello, Lidija Kralj, Deepti Teresa Dcosta, Dora Katsamori, Darren Neethling, Sarah ter Maat, Roy Saurabh, Rena Alasgarova, Elena Radaelli, Ana Stamatescu, Arjana Blazic, Graham Attwell, Giedrė Tamoliūnė, Theodora Tziampazi, Moritz Kreinsen, António José Alves Lopes, Jose Viñas Diéguez, Barbara Loranc, Cristina Obae. Licensed under CC-BY-NC-SAA 4.0.*
- *Figure 1 on page 10 - © Authors' own work I The 4 core concepts in XAI, organized in technical and human dimensions. Licensed under CC-BY-NC-SAA 4.0.*
- *Figure 2 on page 18 - © Authors' own work I The 4 core concepts in XAI, organized in technical and human dimensions. Licensed under CC-BY-NC-SAA 4.0.*
- *Figure 3 on page 21 - © Authors' own work I The 4 core concepts in XAI, organized in technical and human dimensions. Licensed under CC-BY-NC-SAA 4.0.*
- *Figure 4 on page 57 - © Ooge, 2023 I AI-based e-learning platform assigns exercises to the learner. Licensed under CC-BY-NC-SAA 4.0.*
- *Figure 5 on page 58 - © Kim et al, 2020 I AExplanations of estimated scores in the Santa tutoring system. Licensed under CC-BY-NC-SAA 4.0.*

Table of Contents

1. Introducing explainable artificial intelligence and its implications in education 6

1.1. Background 6

1.2. Technical issues 8

1.3. Basic definitions 10

1.4. Main features of explanations in AI systems 15

1.5. Perspectives and tiers of XAI 16

1.6. XAI in education 17

1.7. Contribution and organisation 23

2. Navigating compliance with the AI Act, the GDPR and related digital laws 24

2.1. Background 24

2.2. Primers 25

2.3. Use scenarios 42

2.4. How to implement responsibly 54

2.5. Key takeaways and implementation concerns 54

3. XAI in education from the perspective of different stakeholders **56**

3.1. Background 56

3.2. Visual explanations 57

3.3. Use case 1: AI-powered intelligent tutoring system 59

3.4. Use case 2: AI-powered lesson plan generator 66

3.5. Stakeholder's intervention level and points of attention 74

3.6. Ensuring human-centred explainability in AI for education: roles, responsibilities, and the need for oversight 76

4. Defining educators' competences for and towards XAI **77**

4.1. Background 77

4.2 Foundations for AI regarding XAI 78

4.3. Core competences and principles for integrating XAI in education 80

4.4. Competences for the key dimensions of XAI 85

4.5. Practical implementations 86

4.6. Recommendations for different stakeholders 95

4.7. Summary and final considerations 96

5. Conclusion **97**

1. Introducing explainable artificial intelligence and its implications in education

1.1. Background

Explainable artificial intelligence (XAI) is a sub-field of artificial intelligence (AI), which aims to provide explanations about the reasons why an AI-based system takes a decision or provides an output (TechDispatch, 2023). The search for meaningful explanations is not new in the field of AI, but it has been mainly a technical issue for developers who were looking for reliability in the results obtained by their AI systems, so they could be accepted by end users of specific areas (Ali et al, 2023). The great advance of AI technology in the last years has turned these systems into general-purpose digital tools, and new considerations have arisen in this realm.

In terms of ethical AI, the Ethics guidelines for trustworthy AI published in 2019 by the High-Level Expert Group on AI of the European Commission established seven key requirements for trustworthy AI: (1) human agency and oversight, (2) technical robustness and safety, (3) privacy and data governance, (4) transparency, (5) diversity, non-discrimination and fairness, (6) societal and environmental wellbeing, and (7) accountability. In this general scope report, we can find this paragraph:

> *For an AI system to be trustworthy, we must be able to understand why it behaved a certain way and why it provided a given interpretation. A whole field of research, Explainable AI (XAI) tries to address this issue to better understand the system's underlying mechanisms and find solutions.*

Therefore, XAI is a key field towards trustworthy AI, and throughout this report it will become clear that it provides the practical support to most of the previous ethical requirements.

Regarding the AI Act, it does not explicitly stipulate that AI must be explainable. Instead, human oversight, data governance, cybersecurity and transparency are referred to, alongside the rights of explanation of individual decision-making.

Consequently, the relevance of XAI has increased enormously, and new research and discussion articles studying its impact in different fields have arisen in the short-term (Longo et al, 2024). Such relevance seems to concern policy-makers, AI developers and technology experts, but most end users are not yet aware of it, and need simple answers to these questions: *Why are these explanations necessary? Why is this advanced technology not trustworthy?*

The rapid advance of computing over the last 30 years has taken us from the first personal computers with applications for calculation, document creation and information storage, to currently having different types of

devices with permanent global communication and a multitude of tools that solve increasingly sophisticated tasks. But we as users of such tools do not demand explanations. For instance, no one questions why a photo editing software removes the background in a specific way, or why a particular emoji is suggested when typing a word in a chat. To understand why AI technology is different and explanations are required, we can analyse a general definition of AI, included in Article 3 of the AI Act:

> *'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.*

Two key ideas arise from this paragraph. First, AI systems can operate autonomously and adaptively, facing tasks previously delegated to humans, compromising human agency. Second, AI systems can combine multiple inputs and generate complex outputs that could be difficult for humans to perform and/or understand. But they are less able to reason like a human, and to show empathy, sensitivity or cultural nuance. Therefore, if AI systems were to take decisions on our behalf, and humans may not fully understand them, the involved risk is clear, and also the need to obtain a clear explanation to guarantee that AI is assisting and not deciding.

In addition, it is important to be aware that the output of AI systems could be inaccurate (UNU, 2024). The complexity of the problems that are faced and the internal operation of some techniques used in this field, imply that we cannot trust the provided response completely. For common users of digital technology this is a new scenario, as traditional applications do their task or not, but there is not a probability of success. Therefore, AI systems must include appropriate explanations about the accuracy of their outputs, so their trustworthiness can be reinforced.

The current reality is that most users are unaware they need XAI, and that future AI tools will include it due to ethical and legal reasons. The general public does not know what questions should be raised to AI systems nor do they have the training to properly understand the explanations. This is where education comes in, providing them with the skills and knowledge required to evaluate the trustworthiness of AI systems, fostering critical thinking and agency. This aligns with the Ethical Guidelines on the use of AI in teaching and learning for educators (2022), which highlight the importance of empowering learners through education to critically engage with AI systems in informed and responsible ways. The guidelines are being revised in 2025.

As will be detailed throughout this report, the impact of XAI in education goes beyond capacity building. But to properly frame it, we first need to analyse some core technical issues.

1.2. Technical issues

It must be pointed out that this is an educational report, not a technical one, but there are some technical aspects that must be clarified to understand the singularities of this field. Providing an explanation about the output of standard software is straightforward for developers, as they are based on traditional computer programming, made up of a set of commands that allow to analyse the logic behind a provided output. But in the case of AI systems, the situation is more complicated.

From a general perspective, we can distinguish two main technical approaches to AI: knowledge-based and data-driven (Holmes & Tuomi, 2022). In the first one, human knowledge and expertise are represented in a way that can be processed by computer programmes, mainly through logic rules and probabilistic reasoning. These systems were very popular in the 1980's, but their difficulties to scale up to complex and real problems have restricted their application to controlled domains. Obtaining explanations from knowledge-based AI is simple, as in standard software. This is the reason why it has been the most common approach in the field of AI in education until the emergence of generative AI (Tuomi, 2018). For example, several Intelligent tutoring systems (ITS), in which a personalised student learning path is autonomously created, are based on rule-based reasoning and fuzzy logic. These techniques allow to include detailed dashboards for teachers and students, which provide visual explanations and tendencies about the learning progress, increasing the trustworthiness and usefulness of the systems (Mousavinasab et al, 2018).

On the other hand, data-driven AI is based on the idea that knowledge can be extracted directly from the data corresponding to a given problem, by analysing patterns and making inferences, taking advantage of the high computational power of today's computers. Within this approach, machine learning (ML) is the specific field of study in which algorithms and statistical models are developed that computer systems can use to make predictions or take decisions without using explicit instructions (Marsland, 2011). We now have very reliable algorithms that adjust models to 'learn' the patterns hidden in the data. Such a learning process results in a set of numerical parameters that characterise the model, and which define its response. Obtaining explanations from a set of numbers is not as straightforward as obtaining them from a set of logic rules or commands written in standard language, as in the case of knowledge-based AI. Moreover, the larger the number of numerical parameters, the larger the complexity of the model, and consequently, the larger the complexity of obtaining proper explanations from it.

Readers can find details about the specific techniques, algorithms and models that correspond to each of these two approaches in classical references such as Poole & Mackworth (2023) or Russell & Norvig (2020). But, without relying on a deep understanding of them, a simple example can be presented to clarify the commented differences in XAI terms:

Imagine that a mechanic uses an AI assistant tool to diagnose a car problem. How would this tool work if it is based on knowledge-based AI or in data-driven AI? For the first approach, a technique called case-based reasoning (CBR) is used, and for the second one, an artificial neural network (ANN) model is applied. The following table summarises the main aspects involved in the decision process:

	Case-based reasoning (CBR) for knowledge-based AI	**Artificial neural network (ANN) for data-driven AI**
Scenario	Diagnosing a car problem based on a database of past cases.	Diagnosing a car problem using input features and ANN predictions.
Symptom reported	Clicking sound when turning.	Clicking sound when turning.
Input data	Description of the sound and context (e.g., clicking sound, occurs while turning).	Numerical encoding of features: • *Sound type:* clicking, whining, thudding (encoded numerically as 1, 2, 3). • *Car action:* turning, accelerating, hitting bumps (encoded numerically as 1, 2, 3). • *Car age:* numerical value (e.g., 5 years).
Process	Matches input to past cases and applies rules: • Rule 1: clicking sound + turning → joint issue. • Rule 2: whining sound + accelerating → transmission belt issue. • Rule 3: thudding sound + hitting bumps → suspension issue.	Processes input through weighted connections and activations: • *Input layer* (3 nodes): sound type, car action, car age. • *Hidden layer* (4 nodes): calculates activations based on weights and biases. • *Output layer* (3 nodes): provides a prediction. Probability of failure on joint, transmission belt, suspension.
Example match	Input: "clicking sound when turning" matches rule 1. Decision: joint issue based on past case.	Input: "clicking sound" (1), "turning" (1), "car age: 5." Weights between input and hidden layer: • Sound type → hidden node 1: **0.8.** • Car action → hidden node 2: **-0.3.** • Car age → hidden node 3: **0.5.** Weights between hidden and output layer: • Hidden node 1 → CV joint: **0.6.** • Hidden node 2 → transmission belt: **0.2.** • Hidden node 3 → suspension: **0.4**. Weighted sums and activations lead to an *output probability*: • Joint: 85%, transmission belt: 10%, suspension: 5%.
Explanation	**Easy to explain:** "In a previous case with the same symptoms, the problem was the joint. **Solution: replace joint.**"	Difficult to explain: Depends on how weights and activations interacted (e.g., sound type weight: 0.85 to joint node). **Solution: 85% probability of replacing joint.**

Table 1: Key differences between CBR and ANN in an AI-based car diagnosis tool example.

This example clearly contrasts the two approaches in terms of XAI. It must be pointed out that not all data-driven models have the same opacity level, as the number of numerical parameters is not the only feature to consider, but the model structure itself (Dwivedi et al, 2023).

However, the improvement of ML techniques and their success when applied to solve real-world problems has been so significant in recent years that data-driven AI has dwarfed knowledge-based AI. But when it comes to explainability, things are still unclear, and some critical fields, such as education and healthcare, prioritise proper explainability even if they imply a lower model performance (Loh et al, 2022; Khosravi et al, 2022). Consequently, many ML researchers and developers are intensively working on computational techniques that allow to obtain explainability from complex ML models (Bennetot et al, 2024). This is ongoing research, and new improvements will be obtained in the near future, so we must be careful when discarding complex data-driven models in light of XAI.

1.3. Basic definitions

It is necessary to establish some core concepts in the realm of XAI that will be used throughout the report. It is out of scope to provide original definitions here, as this is an open issue in the field, but we will rely on those already existing in the bibliography that better fit to the goal audience of this work.

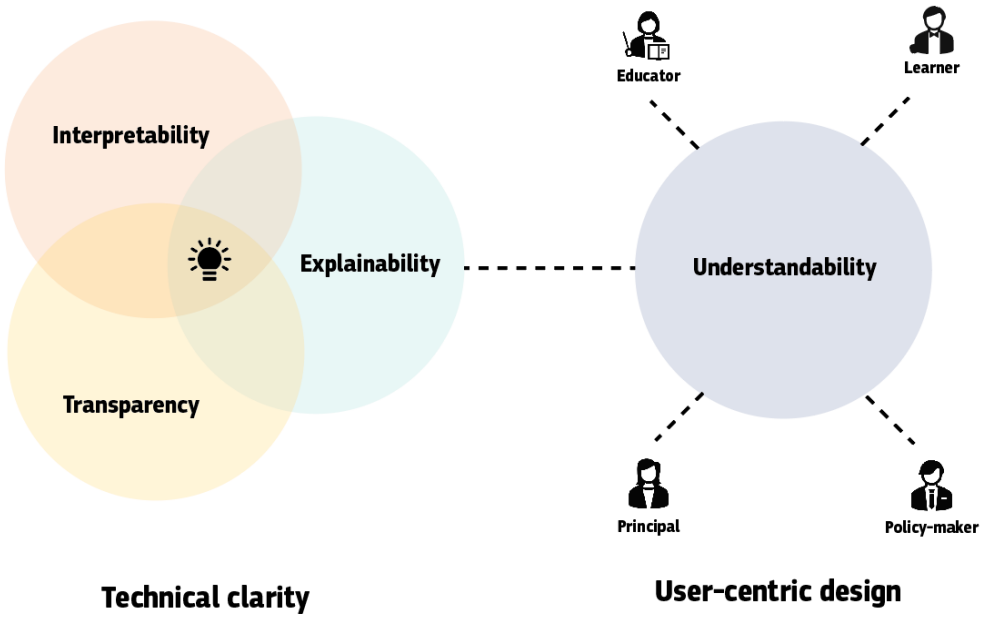


Figure 1: The 4 core concepts in XAI, organized in technical and human dimensions. Source: authors' own work.

These concepts are: (1) transparency, (2) interpretability, (3) explainability and (4) understandability, and they will be defined in detail in the following sections. What is relevant at this point is to clarify that the first two concepts belong to the technical dimension of AI while the last two belong to the human dimension, as illustrated in figure 1. To support the latter, the main goal in XAI, the developer must first include the former in the AI system.

The same perspective argued in (Chaudhry et al, 2022) is followed here, taking transparency as the core ethical dimension of AI, acting as a central link with others like safety, accountability or fairness. Two equivalent definitions are taken here for this concept:

Transparency

> 'Transparency in AI refers to a process with which all the information, decisions, decision-making processes and assumptions **are made available** to be shared with the stakeholders and this shared information enhances the understanding of these stakeholders. (Chaudhry et al, 2022)
>
> Under the EU AI Act, Recital 27, it says 'transparency means that AI systems are **developed** and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their right'.

Therefore, *transparency relies mainly*[1] on the developer, who must develop the AI system in a way that it can be interpreted and understood by the user. This is not an exclusive feature of AI, and this kind of recommendation can be found in a more general scope in the realm of open science practices established by UNESCO, which encourage researchers and developers to share the details of their studies and findings to advance of equity and inclusion in AI.

Five key aspects must be considered by the developer in terms of transparency on AI:

1. **Data:** Providing information about the datasets used to train AI models, including their sources, quality, and any preprocessing steps. This helps in assessing potential biases and the representativeness of the data.
2. **Model:** Offering insights into the AI model's architecture, algorithms, and decision-making processes. This allows stakeholders to understand how inputs are transformed into outputs, facilitating trust and accountability.

[1] As it will be explained in the next chapter, the AI Act confers a right on individuals (end-users) to obtain clear, and meaningful explanations from the deployer on how the AI system was involved in the decision-making process. This could be seen as a non-technical level of transparency. In the educational scope, we would talk about educators' transparency, related to their ability to explain to learners, parents, or peers why they are using a certain AI tool (which ties in with item 5 above). This is an ethical consideration under the term 'justification of choices', and it is very relevant in the scope of AI in education.

3. **Process:** Documenting the development and deployment procedures of AI systems, including design choices, testing protocols, and updates. This ensures that the AI's lifecycle is open to scrutiny and aligns with ethical standards.
4. **Outcome:** Clearly communicating the results produced by AI systems, along with their confidence levels and potential limitations. This helps users interpret AI outputs appropriately and make informed decisions.
5. **Purpose:** Clarifying the intended use and scope of the AI system, including its goals and the context in which it operates. This ensures that stakeholders understand the AI's role and limitations.

Assuming that such transparency is provided, and considering the technical issues established in the previous section, a key dimension of XAI arises, *interpretability*.

Interpretability

> 'Interpretability enables **developers** to delve into the model's decision-making process, boosting their confidence in understanding where the model gets its results'. (Ali et al, 2023)
>
> 'Interpretability refers to how easily humans can understand how a **model works** or makes decisions'. (Ooge, 2023)

In this sense, AI systems can have two different main levels of interpretability (Ooge, 2023):

1. **Inherently Interpretable models:** These are models that are simple enough for humans to understand directly, such as those used in knowledge-based AI (like CBR or logic rules), and some simple models used in data-driven AI (like decision trees or Bayesian models). They provide transparency by showing the logic behind their decisions, which is easy to interpret, and are referred to as white-box models in the field of XAI (Ali et al, 2023).
2. **Complex (opaque) models:** These models, such as neural networks or ensemble methods, are highly accurate but difficult to interpret because of their complexity. Their internal workings are based on large sets of numerical parameters which are adjusted using complex algorithms that take long periods of time and lots of calculations. They are usually referred to as black-box models in the field of XAI (Ali et al, 2023).

In the case of opaque models, as commented above, the field of XAI is being really active on the development of *post-hoc explainability*. That is, interpretability can be added after training using techniques like visualisations, feature importance analysis, or approximations to explain the model's behaviour or individual predictions (Ooge, 2023). Some of these techniques will be explained in more detail in the next chapter.

In general, there is a trade-off between model performance and interpretability, with simpler models being more interpretable but less accurate for complex tasks. See fig. 4 in Ali et al (2023) for a deeper explanation of this issue.

Explainability

> 'Explainability provides insight into the AI system decision to the **end-user** in order to build trust that the AI is making correct and non-biased decisions based on facts'. (Ali et al, 2023)
>
> 'Explainability in AI concentrates on providing clear and coherent explanations for specific model predictions or decisions. It aims to answer questions like "Why did the AI system make this particular prediction?" by offering **human-understandable** justifications or reasons for a specific out-come'. (TechDispatch, 2023)

Therefore, interpretability is used when we are talking about making AI systems transparent by design instead of opaque, and explainability when we mean justifying an AI system's behaviour to end-users (Hamon et al, 2022; Panigutti et al, 2023). In this way, interpretability is a passive characteristic: Any AI model is inherently interpretable or not to a certain degree (Barredo Arrieta et al., 2020). Explainability, however, is an active characteristic: *AI models are explainable when they do something to clarify or detail their internal functions such that humans can understand them more easily* (Barredo Arrieta et al, 2020).

Transparency, interpretability and explainability are three core concepts of XAI that fall on the developer side. The two first rely on the technical features of the approach, while the third is targeted to the user. For the final goal of increasing the trustworthiness of the AI system, the developer must develop it with the highest transparency level while keeping in mind the trade-off between performance and interpretability, because the system must be useful to make sense. Finally, the developer must consider the singularities of end-users, as the explainability should be targeted to them. This last idea links to a fourth XAI dimension, understandability.

Understandability

> 'The degree to which the provided insights can make sense for the **targeted audience's** domain knowledge' (Saeed & Omlin, 2023)
>
> 'The degree of **human comprehensibility** of an AI system decision' (TechDispatch, 2023)

This concept[2] illustrates how well the final user can comprehend an explanation that is targeted to them, so it is a measure of the real utility of XAI. Therefore, understandability is a human-centric dimension that goes beyond the technical characteristics of AI models. It emphasises the need for AI explanations to align with human cognitive and contextual needs, ensuring that people can grasp the system's behaviour and

[2] Understandability is sometimes defined as equivalent to interpretability in the scope of technical XAI literature (Saeed & Omlin, 2023; Chaudhry et al, 2022). These authors consider the developer as the end user, so the interpretability of the models is related with their understandability. But here we assume that understandability depends on the explanation that is tailored to the type of end-user, while interpretability is more general, and it depends on the type of AI model and the provided transparency.

outcomes (Ooge, 2023)[3].

To sum up, XAI encompasses two main perspectives of development: the *technical* and the *human* one. To clarify with a simple case how the four core concepts of XAI (transparency, interpretability, explainability, and understandability) affect an AI system development, the following table continues with the previous example of an AI-based car diagnosis tool, illustrating how a developer can integrate them to foster trustworthiness:

	Technical perspective	Human perspective
Transparency	Use clear and well-documented datasets, such as repair histories and sensor data from various car models.	Inform mechanics and car owners about the data sources used for diagnosis (e.g., "based on 10,000 car repairs").
	Share the types of issues the tool can diagnose (e.g., engine faults, battery health) and its limitations.	Publish a manual explaining the tool's scope and ensure users understand it is an assistive tool, not a final authority.
Interpretability	Choose an interpretable model for simpler issues (e.g., decision trees for battery health).	Provide mechanics with tools that show clear decision paths, like "Low voltage in cell 3, has a 75% probability of being a failing battery."
	For more complex diagnostics (e.g., engine misfires), use feature importance tools to highlight key factors.	Train mechanics on how to interpret and verify the AI's model confidence levels and sensitivity data with physical checks or further testing.
Explainability	Include explanation tools that show why the system suggests specific issues (e.g., "based on engine RPM fluctuations").	Offer visuals, like annotated diagrams, explaining affected car parts (e.g., "The AI detected a leak in the fuel injector system, with high confidence").
	Use counterfactual explanations: "If the spark plug voltage were higher, this issue might not occur."	Ensure explanations are easy for car owners to understand, focusing on what actions to take (e.g., "Replace spark plug, the accuracy of this prediction is high").
Understandability	Simplify language in the interface (e.g., "Fault detected in exhaust system" instead of "Exhaust gas recirculation issue").	Provide a clear, user-friendly app or dashboard for car owners to view diagnostics with severity levels (e.g., "critical, needs repair, high confidence").
	Use visuals (e.g., system diagrams) to highlight problem areas.	Involve mechanics and car owners during testing to ensure the explanations are useful and actionable.
Integration	Continuously monitor tool performance with mechanic feedback, Update models as needed to reduce misdiagnoses.	Provide ongoing training for mechanics and customer support for car owners. Regularly gather feedback to improve clarity and functionality.

Table 2: Technical and human perspective for the AI-based car diagnosis tool example.

1.4. Main features of explanations in AI systems

It is important to converge around a set of core characteristics that explanations should ideally exhibit to be

[3] With regards to the concept of *non-technical transparency* introduced above, understandability is very relevant, as it allows deployers to comprehend the AI system's purpose and, consequently, to be transparent with the end-user. In the scope of education, it means that an appropriate explanation for educators supports them to be transparent with learners, parents or peers.

both ethically responsible and truly helpful for end-users. Developers should be aware of them when defining their AI systems. A "minimal set" of these features could be summarised as follows:

Category	Sub-feature	Description	Example (education)
Clarity (to foster understandability)	**Plain language**	The explanation should avoid technical terms	*"Our system looked at your recent quiz scores and noticed you had difficulties with algebraic equations. You might benefit from reviewing those specific concepts."*
	Tiered tetail	Different users may need more or less detail, so explanations should offer basic information at first, with an option to delve deeper into the technical or data-driven aspects if desired	*Basic level: "We used your quiz results to identify areas for improvement." Detailed level: "We combined multiple quiz scores and weighted each question based on difficulty to determine that algebraic factoring is your weakest skill."*
Relevance (in relation with context)	**Plain language**	Explanations should be meaningful within the context of application	*"Because your essay had repeated grammar errors, the system suggests extra practice on sentence structure, which is crucial for this English writing course."*
	Tiered tetail	The explanation should help the user take practical next steps or make decisions	*"Based on your quiz results, the system recommends reviewing Chapter 4 of the textbook and completing the practice exercises by Friday."*
Specificity (related to AI technology)	**Model process or reasoning**	The user should know the system uses certain data inputs and a machine learning (ML) or rule-based model to generate the recommendation or decision.	*"We trained an ML model on past students' quiz scores and final grades. Your current performance data was compared to similar student profiles to suggest targeted study areas."*
	Limitations and uncertainties	The system's explanation should state that it can be uncertain. It may include confidence levels or mention situations where data might be incomplete or biased.	*"This recommendation may not fully reflect your understanding if you have not completed all quizzes yet. The confidence level in the recommendation is an 86%"*
Traceability (for accountability)	**Who is responsible**	The user should be able to identify who (or what organisation) is responsible for the system's outcomes and whom to contact for clarification.	*"This recommendation system is maintained by the Office of Learning Analytics. If you have any questions or concerns, please contact them at [email]."*
	Auditability	The system should record its decision-making steps or data so that an internal or external audit can verify how conclusions were reached.	*"All data used in generating your recommendation is logged. An academic integrity committee can review this log to ensure that your suggestions were produced fairly and accurately."*

Consistency (reliability)	**Stable explanations**	Explanations for similar cases or inputs should not vary wildly; they should follow the same logic or rules.	*"Other students who struggled specifically with factoring polynomials received the same study module recommendation, ensuring consistency across similar profiles."*
	No contradictory messages	If multiple "layers" of explanation exist (basic vs. detailed), they should not conflict.	*"The high-level overview states that algebra is your main challenge, and the detailed breakdown confirms that factoring polynomials is the key skill area needing review."*

Table 3: Core features of explanations in AI.

1.5. Perspectives and tiers of XAI

From what has been discussed up to now, it seems clear that to advance on XAI implies involving different actors and stakeholders and foster their collaboration. As a starting point, one can follow the approach proposed by Saeed & Omlin (2023), that contemplates five perspectives, and associate them to three main types of stakeholders with a different role on XAI:

1. **Lawmakers and policy-makers:** with regulatory and social perspectives. The EU's approach to digital regulation is informed by the fundamental rights of individuals, while aiming to encourage innovation by promoting human-centric and trustworthy AI. It is a balanced approach, promoting the development of safe and ethical AI with a focus on the intended purpose of the system. So, their role is not only about development control but also, and more importantly, on governance and human supervision of an overall AI system. They can define laws and policies that frame the development and use of XAI, which contribute ensuring transparency, explainability and understandability of AI systems.

2. **Researchers, practitioners and developers:** with industrial (professional) and model development perspectives. These stakeholders are the technical force behind XAI, and the main responsible for the advance of the field. Commercial interests must be handled at this level, so a balance between restrictions and opportunities must be reached. They must contribute promoting transparency, interpretability, explainability and understandability of AI systems.

3. **End-users:** with industrial (professional) and social perspectives. This group includes a heterogeneous target audience, which can apply AI systems for their professional development or their particular issues. They must contribute promoting explainability and understandability of AI systems. As commented above, only this last dimension depends on the end-user, but it is the most relevant one, as the target goal of trustworthiness lies here. The education and training community institutions would fall under this type.

From a practical perspective, the two first types of stakeholders shown above are end-users of AI too. Consequently, they are affected by the explainability and understandability of the systems they promote and develop. In this sense, three tiers of XAI can be distinguished, adapted from Ooge, 2023:

AI novices: Individuals impacted by AI systems with little to no technical expertise in AI. They need explanations to understand AI models better, ensuring fairness, trust, and data privacy.

Examples: patients, loan applicants, regulatory bodies, administrative staff, students, teachers.

Advanced AI users: Professionals like data scientists and domain specialists who use AI for analysis and decision-making but lack deep AI technical expertise. They require advanced tools to assess model trustworthiness, tune, and compare models.

Examples: physicians, loan officers, managers, judges, social workers, educational researchers, school IT systems managers, informatics teachers.

AI experts: Specialists who build and deploy AI models or develop explainable AI techniques. They focus on interpreting and improving their models to ensure proper functionality.

Examples: AI researchers, engineers.

1.6. XAI in education

The previous sections provided a general overview of the field of XAI. But this is a report about *education*, so XAI must be framed in this scope.

AI is increasingly seen as a promising tool in education, with the potential to enhance learning experiences, tackle challenges, and personalise instruction to better support the diverse and evolving needs of learners, though its full capabilities and practical impact are still emerging. For learners, AI systems aim to provide personalised learning experiences by analysing individual strengths and weaknesses, delivering tailored content, offering real-time feedback, and identifying areas for improvement to bridge knowledge gaps effectively. For educators, AI has the potential to handle routine tasks like grading and lesson planning, giving educators more time to engage in interpersonal interactions with learners. AI can promote inclusivity by supporting learners with disabilities, multilingual learners, and those requiring alternative learning formats. For a sound introduction to the field of AI in education, see the report of the previous EDEH squad on AI which serves a foundation for the present report.

Maxims of AI explainability

1) Be transparent

Transparency in AI involves articulating its functionality and decision-making processes. Organisations must disclose AI usage and provide understandable explanations for diverse audiences while adhering to regulations like GDPR. This fosters trust and enables stakeholders to comprehend systems outputs.

2) Be accountable

Accountability in AI governance specifies roles and responsibilities throughout the AI lifecycle. organisations need to justify their design and deployment choices and allow stakeholders to contest decisions such as the EU AI Act and GDPR, thus enhancing ethical oversight.

3) Consider context

Explanations must be tailored to the audience's needs, whether they are educators, students, or policymakers. This contextual adaptability ensures that the communication around AI systems is relevant and meaningful, taking into account the expertise and application domain of stakeholders.

4) Reflect on impacts

AI systems should focus on human and societal well-being throughout ongoing assessments of their benefits and risks. Organisations must implement strategies to mitigate potential harms and promote inclusivity, ensuring that AI systems align with ethnical goals and practical objectives.

Figure 2: Maxims of AI explainability.
Source: authors' own work.

Ensuring the explainability of AI systems is fundamental in the development of ethical and reliable AI systems, especially within the education sector where decisions have significant effects on both individuals and society as a whole. Consequently, a number of intricate legal, ethical, and governance challenges have emerged, particularly concerning transparency, fairness, and accountability ("TFA"). In this line, the Alan Turing Institute has presented the AI Explainability in Practice framework, which introduces four guiding principles to establish a strong basis for ensuring that AI systems are transparent, accountable, and in line with the needs of various stakeholders (see figure 2). These principles bridge the gap between legal mandates, ethical considerations, and practical implementation, fostering trust and usability across different contexts. They also offer clear guidance on effectively communicating AI-assisted decisions to individuals, ensuring explanations are both meaningful and aligned with diverse stakeholder needs.

Specific impact of XAI in education

XAI impacts education in many ways. Broadly, we could consider two main aspects:

Capacity building: All educational stakeholders require appropriate AI competencies, including knowledge, skills, and attitudes (Vuorikari et al, 2022). It is necessary to show critical thinking and to maintain the personal agency when using AI tools for teaching, learning, or administration. Moreover, in the case of educators, capacity building for XAI is also required to properly understand the explainability features of the tools they use in their classrooms.

Development of AI tools for education: developers of AI tools in Europe must comply with the transparency and explainability mandates established by the AI Act. In the case of education, such requirements must be discussed and agreed upon with educators and pedagogues, as for the case of learners, the understandability and reliability of the explanations may interfere with the learning process. Self-regulated learning, as the active process where learners utilise their cognitive and physical abilities to acquire skills relevant to specific tasks, could be compromised if XAI is not properly implemented (Azfaal et al, 2023). Therefore, for education authorities, it is necessary to consider the learners' singularities in terms of explainability and understandability when selecting the AI tools to be introduced in learning environments. These ideas reinforce the conclusion of the previous section: XAI in education requires tight cooperation between stakeholders in the development stage and active supervision by academic authorities in the deployment stage.

More specifically, education has distinctive needs for XAI (Khosravi et al, 2022):

Accountability

Educators must be accountable to learners, parents, or the administration when using AI systems for teaching, learning analysis or task recommendations.

Transparency

Educators must explain how decisions are reached and how AI systems used in institutions are used especially when processing personal information, and when using AI systems as defined in the current regulations.

Metacognition and agency

Explanations must empower students to take greater control of their learning by promoting self-reflection, planning, and decision-making.

Legal data compliance

Providers and deployers of AI systems must adhere to legal standards such as GDPR, ensuring transparency about data collection, usage, and retention, paying special attention to the protection of minors.

Handling noisy and complex data

From the technical perspective, educational data related to the student's learning process often originates from diverse sources, including digital interactions, assessments, and behavioural observations, which can be noisy and require careful interpretation.

Misconceptions

AI systems in education must be designed to prevent the introduction of misconceptions or undesired learning behaviours, due to inaccuracies or bias in the output. This is a consequence of the complexity of the data introduced above.

Pedagogical-centred design

Explanations should be aligned with pedagogical goals and learning sciences to maximise educational outcomes.

Specific stakeholders[4]

End users: Educators and learners need simplified and clear explanations to understand AI recommendations and act on them effectively. Educators must be able to critically evaluate outputs, while learners require actionable feedback that fosters trust, engagement, and ownership of their learning process. Even parents could be included here (in line with the educational context during the COVID-19 pandemic), and they benefit from transparent insights into their child's performance and progress to better support their education

Education authorities: Education leaders, principals and policy-makers require insights that balance global and local explanations to make well-informed decisions.

Developers: Developers (or providers, including certain importers or distributors defined under the AI Act, particularly for high-risk systems) must ensure AI educational tools provide practical, user-friendly explanations that are meaningful to non-technical users. Collaboration with educators is essential to ensure that outputs are clear, actionable, and aligned with real-world educational needs.

[4] In the following chapters, these main types of stakeholders in education will be referred to, but in each specific context, it could be necessary to adopt a slightly different terminology (e.g., legal terms could refer to learners and educators as end users, or administrators as education authorities).

Challenges, limitations and opportunities

The following SWOT diagram provides a comprehensive summary of the main challenges, limitations and opportunities of the integration of XAI in education, which will be discussed throughout this report.

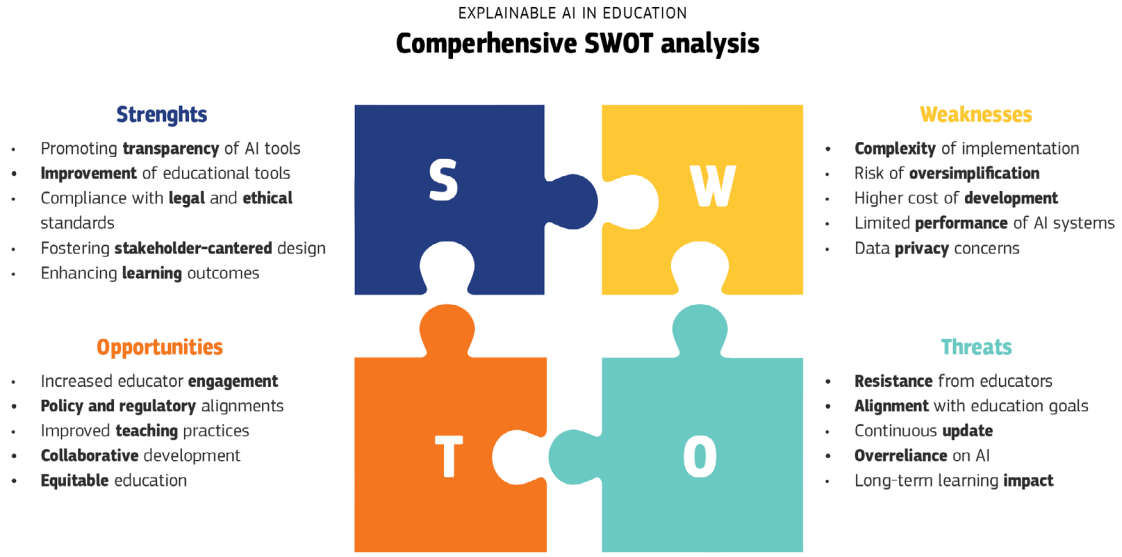


Figure 3: Comprehensive SWOT analysis. Source: authors' own work.

A taxonomy of XAI in education

Table 4 contains a taxonomy that aims to outline the **key dimensions of AI explanations, exemplified for educational contexts**, comprising two levels of explanation that relate to (1) the properties of an AI model or system, and (2) the presentation of explanations to data subjects or users. Each dimension ("scope", "depth", "alternatives" and "flow") identifies specific ways in which explanations can be tailored to meet the needs of different stakeholders. The taxonomy has been adapted from Kesari et al (2024), and will be referenced in the coming chapters of the report.

Scope: Whether explanations are generalised or localised	**Global explanation** Gives a comprehensive understanding of the behaviour of the model across a wide range of scenarios E.g., understanding the appropriateness of an educational tool for operational requirements in line with institutional policies, as a demand from the school principal E.g., assessing whether an AI grading platform demonstrates consistent bias across different student demographics, to be used by a policymaker	**Local explanation** Focuses on understanding the model's behaviour for a specific instance or a small set of instances E.g., explaining to a teacher why a particular individual student received a zero for a grade due to an AI-based assessment system

Depth: **Level of selectiveness in explanations**	**Comprehensive explanations** Transmit comprehensive evaluations for in-depth system reviews	**Selective explanations** Simplified insights for immediate feedback
Alternatives: **Whether explanations are contrastive or non-contrastive**	**Contrastive explanations** Highlights the difference between what happened and what was expected, focusing on alternative outcomes E.g., a student asks why they received a lower grade compared with a peer on an AI-graded assignment. E.g., an educator wants to know why a particular course recommendation was provided to one student but not another.	**Non-contrastive explanations** Provides insight into the model's behaviour without reference to an alternative outcome E.g., educators requiring understanding the factors or features the AI considers most important when assigning grades to students E.g., analysing the general principles behind an AI tool's suggestion for curriculum development
Flow: **How explanations are conveyed (that is, as conditions or patterns)**	**Conditional explanations** Rule-based explanations for targeted decisions (if-then formats show when specific outcomes occur). Makes explanations clearer and easier to understand. Useful for simple, clear and actionable guidelines. The problem is that they oversimplify complex relationships and do not capture subtleties in variables. E.g., a personalised learning system recommends additional practice based on a pre-defined score threshold 60%: *„If the student's quiz score is <60% on topic A, THEN assign additional exercises for topic A."*	**Correlational explanations** Useful for understanding how changes in input data affect the model's output. Helpful for trend analysis and probabilistic insights. However, difficult to interpret. E.g., an automatic grading system shows that higher scores are strongly correlated with time spent on practice assignments, helping educators understand systemic trends. E.g., a personalised learning platform is shown to recommend more reading materials if a student's quiz performance is decreasing. This would show the educator a correlation between increasing the number of recommended exercises and the decreasing quiz scores.

Table 4: Key dimensions of explainability in educational AI systems.

The difference between tables 3 and 4 must be clear. Table 3 focuses on what makes an explanation both ethically sound and user-friendly. It emphasises best-practice "checklist" qualities that any AI explanation should strive to incorporate. It is helpful for educational deployers who need to quickly judge if an AI's explanation meets key ethical and pedagogical standards. On the other hand, table 4 offers a broad view of possible explanatory strategies, enabling educational deployers to pick the approach that best suits their needs (e.g., a quick "local" explanation for an individual learner's grade vs. a "global" overview for institutional policy). Both tables can guide developers in designing AI systems offering clear, relevant, and educational-appropriate explanations.

1.7. Contribution and organisation

The current report aims to contribute to the educational community by providing a formal and updated analysis about the implications of XAI in education. The focus will avoid specific tools and address broader and systemic educational issues, to achieve a strong and future-oriented approach, including practical recommendations for all the stakeholders involved in education: developers, educational authorities, educators and learners.

The report is organised into three main chapters. The first one is focused on the legal aspects of XAI in education. Given the evolving regulatory landscape, it is essential to consider how the AI Act, GDPR and other digital regulatory frameworks influence th e adoption of transparent and explainable tools in education. This chapter is intended to deepen the understanding of how XAI impacts education, helping readers appreciate the complexity of implementing AI in compliance with European laws. Stakeholders such as educators, learners, developers and policy-makers gain clarity on their roles and responsibilities regarding AI explainability. Developers can identify opportunities for creating AI systems that meet both technical excellence and explainability standards, enhancing trustworthiness. Overall, this chapter prepares readers to anticipate and navigate future regulatory and technological changes in the digital education landscape.

The second chapter faces the issues related with the *perspectives of different users*. The significance of XAI in education is emphasised by highlighting its role in fostering trust, transparency, and accountability across diverse stakeholders. It is a practical chapter, which focuses on two main AI applications, intelligent tutoring systems (ITS) and AI-driven lesson plan generators (LPG) and analyses the perspective of different users when using them. From such analysis, it becomes clear that achieving explainability in AI for education requires a collaborative and human-centred approach involving all stakeholders.

The last chapter is targeted to *AI literacy and critical thinking*. It emphasises fostering critical thinking as a fundamental educational goal, using XAI to enhance understanding and transparency. Core contributions include proposing the core competences for educators across all educational levels for understanding, evaluating, and implementing XAI in educational contexts. The chapter includes, as a new aspect, a set of teacher competences to comprehend the key dimensions of XAI shown in table 4. Practical examples illustrate integrating XAI into curricula, from primary to higher education and vocational training, with activities designed to demystify AI processes and foster critical engagement.

2. Navigating compliance with the AI Act, the GDPR and related digital laws

2.1. Background

The EU Artificial Intelligence Act (AI Act) and the General Data Protection Regulation (GDPR) and related laws[5] regulate digital spaces where stakeholders such as learners, educators, edtech companies and education authorities meet. AI systems operate within complex ecosystems where these diverse stakeholders need varying levels of explanations. Learners may seek straightforward and accessible rationales for decisions, such as why they received a new exercise of the same level instead of promoting to the next one, while educators may require particular insights that will enable them to take action to align AI recommendations with pedagogical objectives. On the other hand, edtech developers and regulators mandate comprehensive process-based explanations to ensure technical accuracy and adherence to ethical and legal standards. Thus, a single explanation cannot suffice for all these requirements; instead, explainability frameworks must accommodate multiple approaches. Moreover, there is the technical challenge of translating the complex workings of algorithms, especially those based on advanced techniques such as neural networks, into explanations laypersons can understand.

To ensure these legal principles are operationalised by stakeholders, AI systems must be understandable, actionable and relevant to the intended recipient. Thus, developing methods to make their logic accessible to non-experts, without oversimplifying or misrepresenting the underlying processes, is a significant hurdle. Consequently, addressing these challenges in educational AI systems requires a balanced approach. *Explanation methods must simplify complex AI techniques in ways that help decision subjects – such as learners, educators and administrators – understand and trust AI-driven decisions. At the same time, these explanations must comply with legal requirements while meeting the specific needs of the education sector.* This chapter is structured to first introduce the educational context and practical needs of AI use in educational institutions. It then moves to the legal obligations – particularly under the AI Act and related GDPR provisions – which frame what is permissible and required. Finally, it explores the technical aspects needed to meet these legal and educational goals in practice. This progression reflects the order in which many educational institutions should approach operationalising AI-powered tools: starting with goals, then checking legal constraints, and finally implementing or procuring technical solutions. Examples in this chapter are hypothetical and used for illustrative purposes. Given the limited availability of tested, publicly documented cases in this area, hypothetical examples serve to illustrate common risks and guide the development of best practices.

[5] Digital Services Act, Digital Market Act, Data Act, Data Governance Act, Cybersecurity Act, and Cyber Resilience Act. Table 6 includes a summary of these laws.

Against this background, the current section begins with a primer on the main educational, legal and technical concepts underpinning explainability in AI. Next, it will be shown how these concepts apply in three fictional use cases – namely, automated grading, intelligent tutoring and AI-generated content detection tools. This will be done by analysing the educational, legal and technical aspects; addressing the associated challenges; and providing recommendations to relevant stakeholders. The section concludes with key takeaways and implementation concerns. The taxonomy of the key dimensions of explainability in educational AI systems shown in table 4, provides further context for the regulatory framework and it will be mentioned throughout the sections.

2.2. Primers

Educational

When implementing XAI in educational settings, it is essential to prioritise and promote transparency so AI systems can provide clear explanations for their decisions, allowing educators and learners to understand and trust the outcomes (Maity & Deroy, 2024). Selecting AI models that balance performance with interpretability further enhances this trust. But transparency requires explanations that are task-specific and actionable to be useful. For example, in the case of an automated grading system, educators need local explanations to understand how the AI system operates and assigns grades (Messer et al, 2024). Such educators must provide non-contrastive explanations to learners, detailing the key factors or features considered by the AI, such as the rubric used and its descriptors and weightings. This ensures that both educators and learners can trust the AI's reasoning and outputs.

In addition, accountability must be clearly defined, with processes specifying whether developers, educators or other operators are responsible for the AI's outcomes, and robust error-handling protocols should be set up to address and rectify any mistakes, ensuring human oversight remains in place and is integral to the system. Developers should be responsible for designing transparent algorithms, minimising bias, and providing detailed documentation of system operations. Educators, as end users, are responsible for interpreting AI results, validating them against human judgment and ensuring AI recommendations are aligned with educational goals. System operators must oversee the continuous monitoring of system performance, address errors and ensure compliance with ethical and privacy standards.

In addition, addressing bias and promoting fairness means checking how the AI systems make decisions, for example, by testing whether learners from different backgrounds receive similar feedback. Educators can support this process by identifying patterns that the system may miss. When bias is found, developers can

adjust the data or scoring methods accordingly. To effectively identify bias and promote fairness, educational decision-makers and system operators require tailored explanations that meet the specific needs of the stakeholders involved. Such explanations (Kim et al, 2024) promote trust and fairness, and transparency in AI systems. For instances, explanations based on counterfactual approaches show why one learner received a different result than another (Miller, 2018). Feature-based explanations (Ribeiro et al, 2016) help identify potential sources of biases, while procedural explanations can show how a decision was made, which is useful for audits or fixing errors. Clear and customised explanations help everyone understand and evaluate AI decisions, address disparities, and uphold trust in educational contexts (Binns et al, 2018).

Addressing bias and promoting fairness also requires analysing both training data and system outputs to identify patterns of bias that should be addressed through strategies such as rebalancing datasets, adjusting algorithmic weights, or incorporating fairness constraints. This must be carried out by the developers during the validation stage of their AI system. Equally important is transparent communication about how data is collected, used and accessed. Educators, learners and parents should understand what types of data the AI system uses (e.g., test scores or participation metrics), how this informs decisions, and who has access to that data. Compliance with relevant regulations, such as the GDPR, ensures privacy and data protection are maintained, while clear consent protocols and anonymisation practices prevent the misuse of sensitive information.

Developing comprehensive policies on the use of technology in educational institutions includes setting up robust data privacy and security measures, such as detailed policies on data handling and strong cybersecurity protections to safeguard sensitive information. Ethical use of technology should be guided by a clear code of conduct to prevent misuse, such as cyberbullying or unauthorised data access, and policies should govern the ethical deployment of AI to align with the institution's values and educational goals (Paschal, 2023). Fostering accessibility and inclusivity involves providing equal access to necessary technology for all learners, including assistive technologies, ensuring an inclusive learning environment. Different examples of the use of AI systems in education can be found in the AI report by the first EDEH squad on AI in education.

Professional development for educators should involve ongoing training on effectively integrating technology and understanding AI tools, as well as ensuring all stakeholders are aware of and understand that the established technology policies can promote a culture of ethical and effective technology use.

There could be several focus areas for effectively integrating AI into education, addressing different yet interconnected dimensions of technology application:

Transformational use of AI (SAMR Framework): One focus area emphasises the progression of technology use in transforming educational practices. By analysing how AI tools can enhance, modify or redefine learning activities, educators are encouraged to move beyond basic substitution of traditional methods and explore how AI can fundamentally improve or create new learning experiences. This perspective is critical for leveraging the unique capabilities of AI, such as personalised feedback, adaptive learning systems, and the ability of such systems to model complex concepts in ways previously thought impossible.

Interdisciplinary expertise (TPACK Framework): Another focus area highlights the interplay between an educator's technological, pedagogical and content knowledge. Understanding how these domains intersect is vital for designing AI-driven learning experiences that are both meaningful and pedagogically sound. This approach ensures that the integration of AI is not just technically proficient but also aligned with the content being taught and the strategies used to deliver it. It prioritises the educator's role in crafting lessons that effectively harness AI's potential to enhance learner understanding and engagement.

Systems and external influences (SETI Framework): A further focus area adopts a broader systems perspective, recognising that technology integration is influenced by a variety of external factors beyond the classroom. This includes the availability of infrastructure, institutional support, policies, and the socio-cultural environment. By considering these elements, this perspective ensures educators are not working in isolation, and that the necessary support systems – such as training, leadership guidance and equitable access to AI tools – are in place. It also highlights the importance of addressing societal attitudes and cultural norms around technology, which can significantly affect its acceptance and efficacy in educational settings.

Together, these focus areas can highlight important points of attention for the successful use of AI in education and encourage educators to think critically about how AI can transform learning, as well as ensure its integration is pedagogically grounded and sustainable.

Legal

Fundamental rights are deeply embedded in the constitutional fabric of the EU, serving as a value-based foundation for European integration and providing a normative framework for the EU's legislative agenda[6]. Thus, the EU's approach to digital regulation is informed by the fundamental rights of individuals but simultaneously aims to foster innovation by promoting human-centric and trustworthy AI (European Commission, 2018). Against this backdrop, the AI Act sets out rules for the development, marketing and use of AI across the EU, complementing the GDPR and other EU digital laws. While the AI Act applies to AI systems and general-purpose AI models (GPAIMs), the GDPR applies to the processing of personal data. If

[6] Article 2 of the Treaty of the European Union. See further Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law (2024) (note that the Council of Europe is independent from the EU).

an AI system or GPAIM processes personal data, both the GDPR and AI Act apply. Both the AI Act and GDPR are sector-agnostic.

A cornerstone of the AI Act is its risk-based framework, which classifies AI systems into four distinct **risk categories**: *prohibited, high, minimal, low*, based on the potential risks to individuals and society.

What is covered by the AI Act?

The AI Act regulates two types of technology, AI systems and GPAIMs:

AI Systems[7]**:**	**General-Purpose AI Models (GPAIMs):**
Any machine-based system (software) that is designed to operate with varying levels of autonomy to infer how to generate outputs (such as predictions, content, recommendations or decisions) that can influence physical or virtual environments, and that may continue to adapt after deployment. (Article 3(1), Recital 12)	An AI model that displays significant generality, is capable of competently performing a wide range of distinct tasks and can be integrated into a variety of downstream systems or applications. (Article 3(63), Recitals 97, 98, 99)

For instance, an open-source AI tool designed to assist educators in grading essays may initially seem exempt, but since it directly interacts with student submissions and influences individual evaluations, it must comply with transparency, fairness and risk-management requirements. Similarly, a GPAIM released by a university for research into adaptive learning methods may qualify for some exemptions under open-source provisions, but if it is integrated into a commercial learning management system for personalised education, it must adhere to the obligations of the AI Act, including those aimed at ensuring accuracy and non-discrimination. Finally, a GPAIM that informs large-scale educational strategies (e.g., funding decisions) will not be exempted owing to its systemic impact, and it must therefore follow regulations for high-risk AI systems (Article 55, Recitals 114, 115, AI Act). AI systems and models 'specifically developed and put into service for the sole purpose of scientific research and development' (Article 2(6), Recital 25) are exempt. For instance, if a university develops an AI tutoring tool for research purposes, it is exempt from the AI Act during the research and development phase.

[7] See European Commission, *Annex to the Communication to the Commission: Approval of the Content of the Draft Communication from the Commission – Commission Guidelines on the Definition of an Artificial Intelligence System Established by Regulation (EU) 2024/1689 (AI Act)*, C (2025) 924 final, 6 February 2025.

Key terms for understanding scope and applicability

Understanding the key terms *placing on the market, making available on the market, and putting into service* is essential for understanding the scope and applicability of the AI Act, as they define critical stages in the lifecycle of an AI system or GPAIM.

Placing on the market (Article 3(9))	The first time an AI system or GPAIM is made available in the EU. This triggers initial compliance requirements for manufacturers and developers.
Making available on the market (Article 3(10))	Supplying an AI system or GPAIM for distribution or use in the EU as part of a commercial activity, whether for payment or free of charge. This expands the scope to cover the entire supply chain.
Putting into service (Article 3(11))	Supplying an AI system for first use to a deployer or for use in the EU for its intended purpose. This marks the start of operational deployment, emphasising compliance for deployers and end-users.

Who is covered by the AI Act?

The AI Act identifies key actors (collectively "operators")[8] involved in the AI lifecycle:

Providers (Article 3(3))	Developers of AI systems or GPAIMs such as natural or legal persons, public authorities or agencies (e.g., edtech companies, universities and research institutions,[9] government agencies, or departments developing AI systems for use in public education), publishers (educational publishers creating AI-driven content platforms, interactive textbooks, quiz generators, etc.), and AI-as-a-service providers.
Deployers (Article 3(4), Recital 13)	Supplying an AI system or GPAIM for distribution or use in the EU as part of a commercial activity, whether for payment or free of charge. This expands the scope to cover the entire supply chain.
Importers (Article 3(6))	Entities placing AI systems on the market, often working in conjunction with providers.
Distributors (Article 3(7))	Entities who dispense and administer AI systems in a supply chain, other than the provider or importer, or who make an AI system available on the EU market.

Who is liable under the AI Act?

The AI Act applies mainly to *providers* (e.g., edtech developers who develop or commission the development of an AI system or GPAIM). Examples include major technology companies, cloud and infrastructure providers, open-source AI communities, and academic research institutions. However, the AI Act also places obligations on *deployers* – individuals or organisations using or operating an AI system in a professional context such as educators, educational institutions or system operators. Therefore, the deployer could be a school using an

[8] Article 3(8) provides that "operator" means a provider, product manufacturer, deployer, authorised representative, importer or distributor.
[9] If the system is developed solely for research, it may be exempt under Article 2(6). However, if it is commercialised or widely deployed, the university becomes a provider under the AI Act.

AI system for automated grading, or a system operator at a tertiary institution managing tools that monitor students' conduct via webcams and microphones during online tests. Identifying where each stakeholder in an institution fits within these categories is critical for assessing their rights and obligations, particularly for prohibited and high-risk AI systems.

For example, when an educational leader such as a principal of a school (deployer) evaluates the purchase of an off-the-shelf solution (e.g., a personalised learning platform), that person would need the provider of the AI system to be able to demonstrate compliance via documentation and conformity assessments and would need to interpret global, comprehensive and contrastive explanations of its functionalities to verify the provider's compliance with the AI Act and also to ensure that the solution aligns with institutional policies. Conversely, an educator (e.g., a teacher (deployer) using an automated grading system in a course) may need local, selective and conditional explanations to quickly understand and address why a specific student received a particular grade. These tailored explanations ensure transparency and support informed decision-making across various educational roles. Importantly, the *intended purpose*[10] of an AI system refers to the use specified by the provider, including the context and conditions outlined in the system's instructions, promotional materials and technical documentation. To comply with these specifications, deployers could rely on global explanations to understand the system's overall behaviour, capabilities and limitations across diverse scenarios. For instance, a tertiary institution using an AI system to detect academic dishonesty must ensure the intended purpose of detecting academic dishonesty is achieved without unfairly targeting students because of linguistic differences in writing style or legitimate collaborative practices.

Further, comprehensive explanations in technical documentation should support a deeper understanding of the system's design and limitations, such as detailing how an AI-powered grading system evaluates assignments across various subjects. Conditional explanations clarify how the system operates under specific conditions – e.g., explaining the rules behind the triggering of additional practice recommendations in a personalised learning platform or the logic used by an attendance monitoring system to flag absences. By integrating these dimensions of explainability, deployers can ensure the system operates transparently and ethically. With context-specific explanations, deployers can clarify decisions made by the AI system to affected parties (e.g., students or parents). This transparency fosters trust and allows deployers to perform their oversight function sufficiently, while also aligning with rights afforded to individuals regarding automated individual decision-making, including profiling under data protection laws (Article 22, Recitals 71, 72, GDPR).

[10] The intended purpose is how the AI system is meant to be used, including the context and conditions of its use (Article 3(12), AI Act).

Unacceptable risk: prohibited AI practices in the AI Act[11] (Chapter II, Article 5, AI Act)

Certain AI applications are strictly prohibited, as they risk violating fundamental rights and ethical standards. Practices such as *social scoring* (Article 5(1)(c)), whereby systems rank or score learners or staff based on behavioural traits or personal characteristics (e.g., facial expressions or voice tones), or AI systems placed on the market and put into service specifically or used to infer the emotions of a natural person in educational institutions are prohibited (Article 5(1)(f), Recital 44). These tools – designed for the purpose of detecting, or used to detect, emotions in the classroom during assessments or during educator-learner interactions – raise serious concerns about privacy[12], the need for consent, and the accuracy of their interpretations. However, the same emotional detection system (face-detection technology) can be applied for vastly different ends – social scoring (prohibited) or simple verification methods (low risk). This duality underscores the importance of contextual regulation. Explainability in this scenario is less about determining whether a system's design is inherently compliant and more about the context in which it is implemented. It is therefore the responsibility of the deployer of the system to implement and monitor the system's intended use or purpose to ensure compliance. However, deployers must inform individuals that emotion-recognition technologies are being applied to them, even if their use aligns with other permitted purposes (Article 50(3), Recital 132, AI Act).

Explainability can help clarify how the tool is being used and whether it is being used ethically. The deployer needs to assess various key risk factors, including:

1. **Data input:** What (end) user (e.g., student) information is being fed into the system (e.g., grades, attendance or behavioural patterns)?

2. **Data output prediction:** What predictions or decisions are being made (e.g., recommending tutoring or categorising students in ability groups)?

3. **Input-output correlation:** How does (end) user data affect decisions (e.g., does a student's attendance correlate unfairly with academic ability prediction)?

As introduced, explainability is not merely about making algorithms interpretable, but understanding and communicating the role of the system within its broader context of deployment.[13]

[11] Entry into force six months after AI Act on 2 February 2025. See Commission Guidelines on prohibited artificial intelligence practices established by the AI Act.

[12] Emotion-detection systems in educational contexts face significant legal challenges under the GDPR, particularly regarding privacy, consent and data accuracy. Key provisions include Article 5 (which mandates lawful, transparent and purpose-limited data processing, meaning that collecting facial expression data without informing learners or parents of this breaches transparency) and Article 9 (which restricts the use of sensitive data such as inferred emotional states without explicit consent). Article 22 prohibits automated decision-making with significant effects, such as profiling learners based on emotions, without meaningful human oversight. Institutions are required to perform data protection impact assessments (DPIAs) under Article 35 to evaluate risks and ensure compliance. Furthermore, consent must be informed, specific and revocable, as outlined in Article 7, while systems must prioritise privacy by design and default under Article 25.

[13] See, further, the Digital Markets Act (DMA), which mandates gatekeepers to increase transparency in their policies and algorithms, including public compliance reports. These reports can provide critical insights for evaluating educational tools, particularly regarding profiling and personalisation practices.

High-risk AI systems in the AI Act (Chapter III, AI Act)

For high-risk AI systems (Article 6(2), AI Act, Annex III, Recital 56), such as those AI systems that determine access to vocational education and training institutions and evaluate learning outcomes or admissions, institutions must comply with strict transparency, human oversight and accountability measures. These provisions will apply to high-risk Annex III systems starting in August 2026 Most of the compliance requirements fall on the providers (e.g., edtech developers). Under Article 6(3) of the AI Act, certain high-risk systems may be exempt from full compliance if providers can self-assess that they pose no significant risk to fundamental rights or do not materially influence decision-making (see Recital 53). However, there are also far-reaching obligations on deployers (e.g., educational leaders, educators and other operators) who use these systems in a professional context. These duties include deployers taking appropriate technical and organisational measures to ensure the AI system is used in keeping with the instructions for use that accompany the system, as well as monitoring the operation of the system, implementing competent human oversight to the extent the deployer exercises control over the system. Further, this human oversight function includes (1) deployers ensuring relevant and appropriate robustness and cybersecurity measures are regularly monitored for effectiveness, and are regularly adjusted or updated; (2) ensuring input data is relevant and sufficiently representative to the extent the deployer exercises control over the input data; (3) maintaining the logs automatically generated by the AI system to the extent they are under their control; (4) consulting workers' representatives and informing the affected employees they will be subject to the system prior to the putting into service or use of a high-risk AI system in the workplace; (5) informing people they are subject to the use of a high-risk AI system and their right to an explanation if the system is being used to make decisions or assist in making decisions related to natural persons; and (6) performing an assessment of the system's impact in the specific context of its use.

Human oversight (Article 14, AI Act)

The human oversight function or human in the loop (HITL) described above comprises two aspects of AI oversight: first, AI development and second, operational phase. The "loop" refers to stages in the AI system lifecycle where human oversight may be needed to avoid risks. Some loops do not require HITL intervention. For example, AI systems that affect learners' academic progress, admissions or assessments (high-risk AI systems) require human oversight whereas those AI systems performing routine tasks like automating administrative workflows do not necessarily require oversight. Thus, it is necessary to delineate the scope of these loops appropriately to avoid excessive oversight. To do so, deployers must identify the context and impact on decision-making these AI systems have. For example, in a school using AI systems to make

learning recommendations, loops may include training the model, integrating it into the learning platform, and real-time interactions with students. Among these, real-time interactions may require HITL oversight, while simpler interactions may only need periodic reviews.

Moreover, it is critical to choose the right expert for HITL. The expertise needed depends on the purpose of the oversight. For instance, if the goal is accuracy in grading essays, the human should be a subject matter expert and understand the AI tool's evaluation methods. However, if the object is to ensure fairness in admissions decisions, the human must understand fairness (equity) principles and the metrics used by the AI system. Explainability supports defining the goals of HITL oversight. By clarifying what the system does and how it does it, institutions can better align oversight roles with the specific risks or objectives associated with the AI's use. For instance, if the AI system is designed to predict academic success, explainability helps determine whether the human overseeing it needs expertise in academic data analysis, ethical guidelines, or institutional policies. It is important to note that a HITL as a solution is only as good as the loop it is in. To appropriately perform the human oversight function, it is important to define the loop, understand why human oversight is needed, and have a process to eliminate systemic risks. Without these measures, the HITL will not work.

The role of explainability and HITL becomes even more critical when applied to biometric systems in educational settings[14]. Biometric systems such as facial recognition, are often employed to monitor attendance, enhance campus security in classrooms or examination halls, or identify individuals during large student gatherings or on online learning platforms. These applications are useful but should be deployed with care. Biometric systems used solely for verification purposes, such as login mechanisms to confirm a student's identity when the student accesses resources or platforms, are excluded from the high-risk classification. However, deployers must still inform individuals that biometric categorisation systems are being applied to them (Article 50(3), AI Act). Moreover, broader applications such as tracking or surveillance are considered high risk owing to the potential for them to be misused, which would occur when there is unauthorised tracking of students or staff, breaches of privacy, or the sharing of biometric data without proper consent.

Deployers of high-risk categorisation systems must inform individuals exposed to these systems (Article 50(3), Recital 132, AI Act) and process the data generated by them in compliance with the GDPR. Transparency is essential, with institutions under an obligation to communicate the purpose of collecting and storing such biometric data. Moreover, students and staff must be provided with mechanisms to opt out of non-essential uses, ensuring their rights are protected and fostering a sense of trust within the educational environment. Explainability safeguards that the operation of these systems is transparent, allowing deployers and oversight personnel to understand how the biometric data is collected, processed, and used, and whether

[14] High-risk biometric categorisation system" under Article 6(2) and set out in Annex III refers to systems used for purposes such as identifying sensitive or protected attributes, as these have the potential to cause significant harm or influence decision-making outcomes.

it aligns with the system's intended purpose. For example, explainability allows the human overseeing a biometric attendance system to verify that the data collected is solely used for attendance purposes and not for unauthorised tracking or profiling.

For high-risk systems, e.g., biometric systems, there is an obligation to conduct a fundamental rights impact assessment (FRIA) (Article 27, Recitals 93 and 96, AI Act). Deployers of high-risk systems, including public sector bodies and private entities providing public services, must complete a FRIA before deploying such systems. This undertaking involves identifying affected individuals, assessing risks to fundamental rights, and implementing oversight and mitigation strategies. The FRIA process enhances understanding of the AI system and its data, promoting transparency for stakeholders, and providing a framework for embedding explainability and HITL into the operation of high-risk AI systems.

Minimal risk AI systems in Article 50, AI Act (individual-user-facing AI)

Transparency obligations for certain AI systems are set out in Article 50 of the AI Act. Providers must inform individuals they are interacting with some AI systems, such as chatbots. Generative AI, whether in the form of synthetic audio, images, video or text (e.g., deepfakes), must be marked in a machine-readable format and identifiable as artificially generated. This requirement is key in educational contexts, where generative AI might be used to create learning materials, feedback, or communication. Clear labelling helps maintain trust and prevents misuse.

Right to explanations

Furthermore, individuals have a right to obtain clear, and meaningful explanations from the deployer on how the AI system was involved in the decision-making process (Article 86, AI Act). These obligations complement data protection principles[15] on transparency such as the general necessity of transparent communication right (Articles 12 to 14, GDPR), which requires that educational institutions (as data controllers) provide information on processing activities in a 'concise, transparent, intelligible and easily accessible form, using clear and plain language',[16] especially in cases involving information addressed to a child. Students and other individuals have the right to be informed when, for example, automated decision-making processes, such as algorithmic grading or personalised learning systems, are used in educational settings. Furthermore, individuals have a right of access to information about automated decisions affecting them, including details about the decision-making logic and the implications of this for their educational experience (Article 15, Recitals 63 and 71, GDPR). In addition, individuals have the right to object to the processing of their personal

[15] Article 5, GDPR: (1) lawfulness, fairness, and transparency; (2) purpose limitation; (3) data minimisation; (4) accuracy; (5) storage limitation; (6) integrity and confidentiality.
[16] A privacy statement should be linked at the bottom of every website page. A layered privacy notice should be short, condensed and comprehensive. See Art 29 Working Party, Opinion 10/2004.

data, including profiling (Article 21, GDPR). This entitlement is particularly relevant in education, when profiling is used for purposes, such as tracking academic performance or predicting student behaviour. For direct marketing purposes, such as promoting additional services to learners, this right is absolute. Finally, Article 35 mandates that educational institutions conduct data protection impact assessments (DPIAs) before implementing AI systems that process individual data in ways that may pose a high risk to the rights and freedoms of individuals. For example, DPIAs are necessary for evaluating automated systems used in admissions, grading and learner support, as these processes can significantly impact learners' educational trajectories.

The below XAI-Ed compliance map (table 5) outlines tailored obligations for deployers and providers, aligning educational AI systems with XAI principles. This mapping is intended to serve as a high-level referencing tool.

High risk AI systems: Chapter III, AI Act		
Key obligation	**Provider**	**Deployer**
Risk management systems Article 9	Implement risk management for biases, fairness, and transparency in AI tools.	-
Data and data governance Article 10	Stipulates requirements for training, validation and testing data sets. Must be relevant, sufficiently representative and free of errors and complete in view of the intended purpose.	-
Technical documentation Article 11	Sets standards for creating technical documents for high-risk AI systems before being placed on the market.	-
Record keeping Article 12	Establishes the rules for the automatic recording of events, or logs, over the lifetime of an AI system.	-
Transparency and provision of information to deployers Article 13	AI systems must be designed to be transparent, so deployers can understand and use them correctly. Instructions must be clear and include information about the provider, the system's capabilities and limitations, and risks. They must explain how to interpret the system's output, any pre-determined changes to the system, and how to maintain it. Instructions should describe how to collect, store and interpret data logs.	Share understandable explanations with end-users (e.g., students, parents, and staff).
Human oversight Article 14	Design systems that allow effective human oversight. These measures should match risk and context and be built into the system by the provider. AI systems must include mechanisms to guide and inform a person to whom human oversight has been assigned to make informed decisions about when and how to intervene.	Ensure effective oversight of AI systems used in the operations aiming to prevent or minimise risks according to its intended purposes or reasonably foreseeable misuse.

Accuracy, robustness, and cybersecurity Article 15	Design robust AI for education with mechanisms to handle inaccuracies and bias outputs. Secure AI systems against unauthorised third parties' attack.	-
Provider obligations Article 16	Comply with conformity assessments and maintain documentation. CE marking compliant.	-
Deployer obligations Article 26	-	Imposes obligation to take appropriate technical and organisational measures and to assign human oversight, e.g., implement measures to ensure safe and fair use of AI in educational settings.
Fundamental rights impact assessments (FRIA) Article 27	-	Deployers that are bodies governed by public law, or are private entities providing public services must assess the impact on fundamental rights that the use of such a system may produce.
Post-market monitoring Article 72	Monitor AI tools' performance post-deployment.	Report issues and review system effectiveness.
Reporting of serious incidents Article 73	Must report any serious incidents to the market surveillance authority within specified time frames. Must submit a report. Must investigate the incident promptly, identify the root cause, and work with the relevant authorities to ensure resolution and prevent recurrence.	Establish mechanisms to monitor and detect serious incidents; escalate suspected incidents to the provider and report them to the relevant authorities if necessary.
Right to explanation Article 86	-	Gives any affected person subject to certain decisions by deployers the right to obtain „clear and meaningful explanation" from the deployer. E.g., students and parents have the right to get information about AI-driven decisions.
Low risk AI systems		
Key obligation	**Provider**	**Deployer**
Transparency obligations for providers and users of certain AI systems Article 50	Providers must inform users that they are interacting with an e.g., chatbot or emotion recognition system or viewing outputs from e.g., deepfakes. AI systems that create content, including general-purpose AI systems, must mark their outputs in a machine-readable format.	Deployers of an emotion recognition system or a biometric categorisation system must inform people of how it operates and process their data in line with data protection (GDPR) obligations.
Transparency obligations for providers and users of certain AI systems Article 50 Voluntary codes of conduct Article 95	The EU's AI Office and member states will encourage the creation of codes of conduct for AI systems. These codes will promote voluntary adherence to certain standards, considering technical solutions and industry best practices.	Deployers can choose to follow voluntary codes of conduct.

Table 5: XAI-Ed compliance map.

In addition to the AI Act and the GDPR, several other EU digital laws are relevant to the educational sector. The table 6 below provides an overview of these legislative enactments.

Regulation	Focus area	Educational relevance	Explainability connection
Digital Services Act (DSA) (entered into full force on 17 February 2024 and is applicable to very large online platforms (VLOPs) and very large online search engines (VLOSEs) since 25 August 2023)	Transparency in platform algorithms, user rights and content moderation.	Transparency in algorithms used for online learning platforms (e.g., content curation, moderation and recommendation systems). Article 28 ensures protection for minors. VLOPs and VLOSEs (e.g., YouTube and Google) must protect user data and curb illegal/inappropriate content. Prohibits targeted advertisements to minors or using sensitive personal data.	Requires platforms to provide clear explanations about how algorithms function in curating and moderating content. Helps educators and learners understand processes such as content recommendation and online classroom moderation. Researchers have access to the data of key platforms to scrutinise how they work. Transparency reporting for intermediary services, hosting services, online platforms, and VLOPs.
Digital Markets Act (DMA) (entered into force on 1 November 2022)	Fair competition and data portability in digital markets.	Applies to very large tech companies (gatekeepers such as Alphabet, Amazon, Apple, Meta, Microsoft, etc.) Ensures fair access to educational platforms/tools. Mandates data portability for educational institutions switching platforms (e.g., moving between learning management systems).	Mandates that gatekeepers clarify how platforms process and store data. Facilitates interoperability by ensuring transparent data handling when institutions switch platforms.
Data Act (entered into force on 11 January 2024. Will apply from 12 September 2025)	Secure data sharing and interoperability, particularly for non-personal data.	Encourages innovation in edtech by enabling institutions and researchers to securely access and understand non-personal educational data. It regulates who can use what data and under which conditions. Emphasises the importance of data literacy.	Ensures that data processors provide clear explanations about data handling, enabling accurate and fair analysis for research or educational insights.
Data Governance Act (DGA) (entered into force on 23 June 2022, applicable since September 2023)	Transparency in data-sharing mechanisms and trusted data intermediaries.	Promotes transparent data-sharing mechanisms for educational institutions. Enables universities to develop AI-driven curricula using shared data sets while understanding how data influences decision-making.	Trusted intermediaries must provide clear information about data processing and sharing, fostering transparency in AI-driven educational applications.

Cybersecurity Act (entered into force on 27 June 2019)	Security certification for ICT providers and systems.	Enhances cybersecurity of digital learning environments and tools. Allows institutions to assess certified tools' security measures, ensuring student and staff data protection.	Raising awareness of cyber-security and promoting cyber literacy in educational institutions. Certification schemes require providers to document and communicate security protocols, ensuring a clear understanding of protection measures.
Cyber Resilience Act (CRA) (entered into force on 10 December 2024)	Secure-by-design principles for connected devices and software. It complements the NIS2 Directive 2022.	Encourages innovation in edtech by enabling institutions and researchers to securely access and understand non-personal educational data. It regulates who can use what data and under which conditions. Emphasises the importance of data literacy.	Ensures that data processors provide clear explanations about data handling, enabling accurate and fair analysis for research or educational insights.

Table 6: Overview of EU digital laws relevant to the educational sector.

Technical

The technical dimension of designing and implementing explainable solutions for educational stakeholders does not only face challenges from the technological software development side, but it also involves the balance between the effective practices of developing complex AI systems, and the user-centered practices that aim to make those systems as transparent and interpretable as possible. Providers of AI systems are expected to include clear documentation written in an accessible language. This means that instructions should be concise, complete, correct and clear – offering information that is relevant, understandable, and usable for educators, learners and other stakeholders. Visualisation tools, like dashboards and progress indicators, have the potential to make data insights and system performance easy to interpret (see next chapter). Another part of the software development practice that is required to ensure explainable and effective AI systems in education is validating the technical solutions in educational context with the corresponding stakeholders, to establish effective communication and adaptation of XAI techniques to each use case.

In education, the role of XAI extends beyond technical challenges to address the diverse needs of stakeholders, including learners, educators, administrators and legal entities. Educational AI systems often consist of complex structures with multiple AI models working together, necessitating explainability that goes beyond individual model outputs. Effective XAI must deliver transparent and comprehensible explanations of AI decisions tailored to the needs of each stakeholder, as it will be detailed in the next chapter.

From a technical perspective, developing XAI in education requires a solid understanding of stakeholder needs, and these must be translated into explicit technical requirements that ensure accountability and are lawful.

This process faces challenges, including integrating diverse stakeholders into XAI design, addressing their varying AI literacy levels, and accounting for pedagogical and legal aspects. Additionally, defining the end users is crucial for selecting suitable explanation algorithms and formats (e.g., textual, visual, feature-based, example-based, etc.). The chosen XAI techniques need to align with the specific use case and stakeholder needs, such as selecting global explanations for policy-makers to understand overall system behaviour, and local explanations for students seeking clarity on individual outcomes.

Technology stakeholders often lack the knowledge about education to design the optimal content and format of AI explanations. Therefore, co-design approaches using communication interfaces and channels among stakeholders from the different disciplines are needed. To that end, XAI design can be supported by (1) clear definitions of the terms and vocabulary used in the corresponding disciplines, and (2) clear requirement lists, functions, and features defined for the use case. Through this communication, developers can assist education stakeholders in translating pedagogy requirements into technical functions and features of the XAI system.

What to explain? How to explain?

Technical explanations of AI models focus on the mechanisms that led to generating a prediction structure, performance and training data, but educators require deeper insights into the design assumptions, reasoning and input-output relationships of the models. In current XAI discussions, there is a focus on explaining AI models themselves, rather than the environment in which they were developed, which includes their design assumptions, data collection principles, data interpretation, labelling of training data, as well as other connected services, such as model hosting. For example, when a model uses eye movements for gaze tracking, an educator may need clarity on how eye-movement data correlates to detecting screen presence–which stems from assumptions made during data preparation, not just the model's technical workings. In other words, this expands model explainability to include process transparency, requiring AI solutions to provide understandable and pedagogically oriented explanations. AI-assisted educational systems must demonstrate clear data flows and provide robust support for audits.

A set of XAI techniques is available to the developers to select from (Bennetot, 2024). While this selection is greatly influenced by the use case in hand, there are certain requirements on what must be explained by AI systems, which include, but not limited to, system documentation and system transparency, in Articles 11 and 13, respectively, in the AI Act, Chapter III on high-risk systems. Among other legal requirements, developers are expected to monitor the risk level of the system they provide and are obliged to fulfil the legal requirements in terms of its explainability, transparency, and understandability.

To generate an explanation of the inner workings of an AI system, from the approaches introduced in the section 1.1., developers usually use a "post-hoc" approach to explain black-box models, whereas open-box ones are explained using an "ante-hoc" approach:

Post-hoc explanation methods: Post-hoc methods aim to explain black-box models after they are built without altering their structure. These methods provide insights into how models work and why specific decisions are made.

Feature relevance techniques: These evaluate the influence of individual features on the model's predictions.

- SHAP (Shapley additive explanations): Uses game theory to assign importance scores to features, ensuring consistent and interpretable results across various feature combinations.
- LIME (local interpretable model-agnostic explanations): Creates a local, interpretable model around a specific prediction by perturbing input data and analysing the changes in outputs.
- Feature sensitivity analysis: Measures the impact of altering input features on model outcomes, identifying the most influential factors in decision-making.

Counterfactual explanations: These provide "what-if" scenarios, showing how changing certain features could lead to different outcomes. For instance, in a system that predicts student's drop-out, it might indicate that less lecture attendance results in higher drop-out possibility.

Visual explanations: Techniques like saliency maps and Grad-CAM identify regions of input data (e.g., image pixels or text segments) that heavily influence predictions. Dimensionality reduction methods, such as PCA and t-SNE, simplify high-dimensional data for visualisation, helping identify key patterns in the data.

Explanations by simplification: Simplified models, like decision trees, approximate the behaviour of complex models to make their logic understandable.

Explanations by example: This involves showing real or synthetic examples that illustrate model decisions, such as displaying a subset of images classified under a specific label.

Ante-hoc explanation methods: Ante-hoc methods are inherently interpretable, with transparency designed into the model itself. These methods enable users to understand how decisions are made without requiring additional algorithms. However, the technical explanation of the model must still be represented in an understandable format to the non-technical stakeholders. Examples of well-known ante hoc XAI methods are:

Decision trees: These hierarchical models split data into branches based on feature values, offering clear, step by step explanations from root to output.

Linear and logistic regression: These models use feature coefficients to directly show how each variable contributes to predictions. For example, in a student's drop-out detection system, it might show that solving assignments late in the semester is associated with higher drop-out rates.

Generalised additive models (GAMs): GAMs allow non-linear relationships between individual features and outputs while maintaining interpretability. Contributions of each feature can be visualised, balancing complexity and transparency.

Rule based methods: These use clear "if-then" rules, such as "If income > $50,000 and age < 30, then approve loan", making decisions easily traceable.

XAI methods are not implemented in isolation but within frameworks that consider the entire system, including the model, its infrastructure and user interactions. These frameworks (Khosravi et al, 2022; Mohseni, 2019; Liao et al, 2020) emphasise human-in-the-loop (HITL) integration, ensuring human expertise is involved not just in evaluating explainability, but also in the design and development of AI systems. HITL approaches prioritise creating explanations that are contextually relevant and aligned with educational goals, enhancing decision-making by combining AI's precision with human judgment. This collaborative approach ensures AI predictions and XAI explanations are ethically sound, pedagogically aligned, and tailored to specific user needs. Dynamic, stakeholder-centred explainability models with layered explanations are crucial for meeting diverse requirements in educational contexts. Continuous co-design processes and feedback loops from users can refine these models and ensure they remain clear, usable, and relevant across varied educational settings.

2.3. Use scenarios

AI content detection tool

Emil is a 16-year-old high school student in his final year. He was working hard to maintain his grades and prepare for university. He worked part time, studied hard and engaged in extra-curricular activities. He completed a history research project that was worth a large part of his final grade. However, the week after he submitted it, he was surprised to learn he had failed the digital assessment. He asked the teacher how this could be. The teacher said an AI detection tool had flagged his research project as likely to have been generated by AI. The teacher said the tool had also flagged two of Emil's previous assignments on this basis.

Educational primer

This case highlights important concerns about fairness and the ethical use of AI in education. AI detection tools, though helpful, are not flawless. Their probabilistic models may misidentify diverse writing styles, especially for non-English speakers and students with different abilities. Currently, tools for the detection of content created by AI make a great many errors, and their use in education must be very carefully monitored by human oversight or not used at all (Perkins et al, 2024). For such AI tools to be used in education explanations of the model and especially of the limitations must be provided by AI system providers based upon extensive testing on an appropriate data set which resonates with characteristics of potential end users (students in this case).

Building AI literacy among educators and students is also crucial for understanding these tools and using them responsibly. It is important educators are aware of the limitations of AI and adopt a more human-centred approach to assessment, as AI tools should support, not replace human assessment. Educators should evaluate students' work holistically, considering the individual student's abilities, and provide personalised feedback. Future policies should safeguard human-centred approaches – including alternative assessments, appeal mechanisms and clear communication – to ensure fairness. More about assessment with AI tools can be read in the AI report by the first EDEH squad on AI in education.

Legal primer

This AI detection tool system would be classified as high-risk under the AI Act, as it directly impacts Emil's academic progression, future opportunities, and emotional well-being (Article 6(2) read together with Annex III(3)(b)). Consequently, the decision-maker at the school, e.g., school principal, is obliged to conduct a FRIA

before the system in implemented. This step ensures that risks related to bias, fairness, and transparency are identified and mitigated to protect students' fundamental rights and ensure equal opportunities (Article 27, AI Act). Following transparency principles (Article 13, AI Act), the developers of the tool must provide clear instructions to the school principal and include information about the system's capabilities limitations, and risks. Further, the developer must explain how to interpret the system's output, outline any pre-determined changes to the system, and explain how to maintain it. Instructions should describe how to collect, store and interpret data logs. (Article 13, AI Act). Moreover, to comply with the human oversight obligations (Article 14, AI Act), it is imperative that the school principal and other users, such as teachers, are adequately trained on the AI systems to understand and, if necessary, override the AI system's automated outputs. This oversight mechanism is crucial for safeguarding students from potentially flawed or unfair decisions. The school principal in this case must implement technical and organisation safeguards to ensure the system is used for its intended purpose, and is safely and fairly implemented (Article 26, AI Act).

Under the GDPR, the school's use of an AI detection tool to assess Emil's work raises significant concerns. Article 22, GDPR prohibits decisions based solely on automated processing if they significantly affect individuals, which is the case in this instance, as Emil has failed a major assessment. Since the decision relied heavily on the AI's output and did not involve any human oversight, Emil's rights were violated. Transparency obligations (Articles 12-14, GDPR) require the school to inform students about the use of AI tools, including their logic, impact and decision-making role in an academic context. Emil has the right to contest the decision, seek an explanation and request human review. In addition to these procedural failings, the school must demonstrate accountability under the GDPR by conducting a DPIA to ensure compliance with the GDPR principles of fairness, transparency and non-discrimination. To rectify the issue, the school should review Emil's case manually, disclose its AI policies, and ensure its tools are equitable and reliable for all students.

Technical primer

If the AI technique behind this tool was a "grey-box" or "white-box" (see section 1.3.), the model's explanation would be easy to explain. But current AI detector tools are based on generative AI (deep learning), making their reasoning significantly more complex and less transparent. The scenario of Emil failing a history research project after an AI detection tool flagged his work as AI-generated is a classification problem. Classification models predict discrete class labels, and in the context of the scenario, there are two possible classifications (binary classification) with four different possible results:

		Predicted condition	
		Positive (PP)	**Negative (PN)**
Actual condition	**Positive (AP)**	The model **correctly** marks the student's work as AI-generated, which is a **true positive.**	The model **incorrectly** marks the student's work as NOT AI-generated, which is a **false negative.**
	Negative (AN)	The model **incorrectly** marks the student's work as AI-generated, which is a **false positive (the case of Emil).**	The model **correctly** marks the student's work as NOT AI-generated, which is a **true negative**.

Table 7: Classification models.

The most critical error is a False Positive wrongly accusing a student of cheating, as is the case with Emil. To reduce this risk, teachers need clear information from providers about the tool's accuracy, limitations, and guidance on interpreting alerts, allowing them to treat the system as an assistant rather than the sole source of truth. False negatives, where students using AI tools are not detected, remain a significant issue, even if they are sometimes considered less critical.

Providers building AI detection models use performance metrics such as accuracy, precision, recall and F1-score to evaluate their models. Accuracy measures overall correctness, recall reflects how well the model identifies actual cases of AI-generated text, and precision indicates its ability to avoid false positives. The F1-score finds a balance between identifying AI-generated text accurately (precision) and not mistakenly flagging humanly written work (recall).

Metrics may demonstrate strong technical performance, but they do not inherently explain what these results mean in the context of education and are likely unfamiliar to educators, who need a clearer understanding of how these metrics relate to classroom application. A wrong prediction can undermine trust in the assessment process. For example, with a 99% recall, meaning 1 in 100 students may be wrongly accused of cheating, we cannot accept such an outcome for the students concerned, as it may have too much impact on their lives, both in the educational and psychological aspects.

In this case, explaining how the model reaches its results helps educators justify the tool's decisions and builds student trust by clarifying the AI's conclusions. From an educator's perspective, feature-relevance explanations, highlighting key text aspects (sentence structure, word frequency, etc.), are particularly useful when they are clear and understandable, enabling educators to interpret predictions accurately and communicate clearly with learners about why a text was flagged or not.

Challenges

Transparency is a challenge if the AI tool operates as a "black box", making its logic difficult to explain. Ensuring meaningful human oversight is resource-intensive, requiring trained staff to fairly review flagged cases and understand the system's training data and model.

Recommendations

Schools must ensure transparency in the use of AI tools in evaluations by clearly communicating their role and associated policies in a clear and direct manner. To promote fairness, these tools must be validated for accuracy and suitability across diverse student populations, with human oversight integral to the decision-making process. Before implementation, schools must conduct FRIAs for high-risk AI systems and DPIAs to comply with the GDPR principles of fairness, transparency and non-discrimination when processing personal data. Educators must be informed from their corresponding educational decision-makers (e.g., principals) that this type of AI system is high-risk, and it is not trustable. Hence, it should be recommended to avoid any kind of autonomous learning activity that could be carried out by learners with generative AI and then require a plagiarism supervision. After implementation, learners should have the ability to contest AI automated decisions, therefore, establishing an appeals mechanism is essential to safeguard learners' rights. From a technical perspective, an interdisciplinary approach is crucial to contextualise the evaluation of AI models, integrating technical, ethical and educational considerations. AI providers must supply interpretable explanations of their systems' decision-making processes, utilising post-hoc methods such as feature relevance (e.g., SHAP or LIME). Providers should ensure transparency about the features the model evaluates, offering training and documentation that explain how these features correlate with AI-generated text. Ultimately, responsibility extends beyond schools to developers, who must be held accountable for ensuring their systems are transparent, equitable and well-documented, enabling their responsible use in educational settings.

Intelligent tutoring system

Julia, a third-grade learner with mild dyslexia who is learning English as a second language, uses a new AI-powered digital textbook designed to personalise her learning in maths and English. This adaptive system tailors content to her unique challenges and strengths. The textbook highlights key terms, offers simplified phrasing, and provides visual icons for complex words. In maths, it accommodates Julia's slower pace and anticipates her confusion between numbers such as 47 and 74. Exercises are broken down into smaller

steps, and she is provided with immediate feedback and interactive examples. In English, audio prompts, translations and simplified vocabulary help her navigate reading comprehension. As Julia progresses, the textbook adapts to give her more challenging tasks, while maintaining the support tools. The digital textbook also allows Julia to monitor her progress through visual reports highlighting her strengths and areas where she can improve. By tracking patterns in her learning, the AI may not only support Julia's academic growth but also build her self-awareness as a learner. Julia's parents and teachers can access a dashboard that provides insights into her learning journey that help them understand the support she is receiving and her progress.

Educational primer

To make these tools more effective and equitable, transparency and human oversight are essential. Educators should work alongside AI to validate its recommendations and provide personalised feedback (Sağın et al, 2024). It is important educators have the option to intervene in AI recommendations and adjust them to specific needs and contexts. Inclusive AI frameworks in education need to adopt approaches such as co-creation to ensure technologies meet the diverse needs of learners and educators, the necessary ethical safeguards are in place, and inclusion and fairness are promoted. Participatory design emphasises the involvement of different stakeholders – such as learners, educators and parents – to ensure AI tools meet different cultural, linguistic and learning needs. Ethical standards, such as the UNESCO Recommendation on the Ethics of Artificial Intelligence (2021) and the EU's Ethical guidelines on the use of AI and data in teaching and learning for educators (2022), highlight the importance of inclusivity, non-discrimination and reducing inequalities in educational AI systems. Clear policies on data protection are also critical for safeguarding learner privacy.

Legal primer

Under the AI Act, this scenario would be considered a high-risk AI application (Article 6(2), read with Annex III(3)(b)), owing to the AI system's role in educational decisions, its data-driven personalisation, and the potential impact on Julia's learning outcomes. A FRIA is required, as the AI system involves high-risk profiling of a vulnerable child, impacting fundamental rights such as privacy, equality and education. The provider is obliged to design the AI system in such a manner as to enable Julia, her educators and parents to understand how the automated decisions are made, so that they can evaluate insights into the rationale behind personalised adjustments and recommendations (Article 13(1), AI Act). Furthermore,

developers must design the AI system to provide instructions for use that are concise, correct and clear, so that it is relevant, accessible and understandable by deployers in the school environment (Article 13(2), AI Act). Also, data governance (Article 10, AI Act, Recitals 66, 67 & 69) is key to ensuring the data used is relevant, representative, free of errors and complete. Regular testing, using diverse data sets that reflect the demographic and linguistic diversity of learners, is required to prevent such outcomes. Human oversight (Article 14, AI Act) remains critical, and Julia's educators should retain the ability to intervene or override AI-based recommendations. The developers must design the AI-powered system in such a way that it can be effectively overseen by the operator (in this case, the school and its educators), including by means of appropriate human-machine interface tools. Given the sensitive nature of Julia's information, including her learning challenges and progress, several GDPR provisions that operate alongside the AI Act (e.g., Article 10, Recital 69) are applicable. Only essential data, such as her reading ability, should be collected, to avoid unnecessary data collection (data minimisation).

The data must be used only for its intended purpose and must not be repurposed for unrelated activities such as marketing (purpose limitation). If the school wishes to use the data in some way, explicit consent to this from Julia's parents or guardians is required. There must be clear, transparent mechanisms to explain how the data will be used, and these must allow for consent to be withdrawn easily. The tool must safeguard privacy, using techniques such as anonymisation to protect Julia's identity. Julia and her parents retain the right to access, correct or delete her data, as needed. A DPIA must evaluate and mitigate risks, ensuring Julia's privacy and rights are upheld with transparency and accountability. The DSA may apply if the learning tool operates via an online platform. The DSA requires transparency in algorithms, particularly regarding how personalised content, such as tailored exercises or progress reports, is delivered. The online platform must protect minors from harmful content (Article 28, DSA). If Julia's learning platform deploys mechanisms such as default consent settings for extensive data processing, pre-checked opt-ins, or confusing notifications that push unnecessary upgrades, these could qualify as dark patterns (Article 23a(1), Recital 51(b), DSA), and may lead to enforcement actions.

Technical primer

Julia's AI digital textbook stresses the crucial role of XAI in maintaining transparency, trust and responsibility throughout the educational process. On the one hand, the design and content of the dashboard is a part of the decision about what to explain to Julia. The dashboard should provide clear and comprehensive visualisations, explaining Julia's progress and the reasoning behind adaptive interventions. On the other hand, protecting data privacy is essential, particularly as sensitive information about her dyslexia must adhere to regulations

such as the EU's Data Act. In this use case, ensuring the model's protection of Julia's health-related data requires not only an explanation of the model's prediction, but also a higher level of transparency extending to model hosting and the entire decision-making process. Here, we are assuming that the ITS includes an intelligent model that detects Julia's condition and generates personalised recommendations based on that condition. XAI becomes an important requirement for this use case, because models can confuse Julia's condition with other conditions. Misinterpretations, such as mistakenly attributing her learning patterns to dyslexia or another condition, require thorough explanations to justify specific adaptations and prevent biases. To explain the model's recommendations, global explanations such as feature relevance are required to understand how the model associates a learner's interaction patterns with learning content. At the same time, Julia and her educators may require a local explanation from the system providers about the predictions she received from the system, to ensure the model was corresponding to her dyslexia and not generating a prediction based on an incorrect assumption. Through the integration of HITL approaches, the system enables educators and parents to take action, ensuring Julia's learning needs are met.

Challenges

The main challenges involve ensuring compliance with the applicable laws that require transparent, explainable and privacy-compliant systems that allow human oversight and protect Julia's rights. Key priorities include robust data governance, avoidance of bias, clear consent mechanisms, and safeguarding against harmful content or manipulative practices.

Recommendations

Educational institutions must ensure transparency in the use of AI tools in teaching and learning by clearly communicating their role and associated policies to educators, parents and learners. To promote fairness, these tools must be validated for accuracy and suitability across diverse learner populations, with human oversight integral to the decision-making process. Institutions must conduct FRIAs for high-risk AI systems and DPIAs to comply with the GDPR principles of fairness, transparency and non-discrimination when processing personal data. Educators must be informed from their corresponding educational decision-makers (e.g., principals) that this type of AI system is high-risk, and it must be assessed before use with learners. From the educator's perspective, it is essential that the ITS is also assessed as effective and trustworthy regarding the teaching approaches and learning design on which it is trained, as well as the effective support that it provides to all learners with respect to their special educational needs and way of learning. It is critical that developers and ITS providers ensure embedded human oversight by design so educators can intervene with ITS decisions and "manually" assign tasks or make changes to the learning paths.

Educators must be equipped with tools to oversee, intervene or override AI recommendations to maintain accountability and prevent reliance solely on automated systems. Systems should be designed to provide clear, accessible explanations of automated decisions for Julia, her educators, and her parents, along with user-friendly instructions for the school. From a technical perspective, an interdisciplinary approach is crucial to contextualise the evaluation of AI models, integrating technical, ethical and educational considerations. AI providers must supply interpretable explanations of their systems' decision-making processes, utilising post-hoc methods such as feature relevance (e.g., SHAP or LIME). Providers should ensure transparency about the features the model evaluates, offering training and documentation that explain how these features correlate with AI-generated text. Ultimately, responsibility extends beyond institutions to developers, who must be held accountable for ensuring their systems are transparent, equitable and well-documented, enabling their responsible use in educational settings.

Automated grading

A university adopts an AI-automated grading system using an AI index to evaluate student essays. While the system quickly assesses submissions based on criteria such as structure and lexical correction, faculty members notice students from diverse linguistic backgrounds consistently receive lower scores. This trend raises concerns about systemic biases, as the algorithm appears to disadvantage those who use varied language styles or are non-native English speakers. In response to these disparities, students and parents voice their frustrations, questioning the fairness of grades assigned by an opaque system. They demand greater transparency and assurances that their academic potential will not be hindered by hidden biases.

Educational primer

This case highlights the challenges and ethical implications of deploying AI-driven grading systems in education. The use of AI systems to evaluate essays has led to unintended consequences, with learners from diverse linguistic, cultural and economic backgrounds receiving consistently lower scores. This trend raises significant concerns about systemic bias, transparency and the inclusivity of AI-based assessment tools. A key issue is the apparent bias in the grading algorithm. Learners who use varied language styles or are non-native English speakers may not conform to the patterns the AI associates with higher-quality writing (Wang, 2024). This not only affects their grades, but also potentially undermines their confidence

and academic progression. Such biases highlight the risks of relying on AI systems that lack the capacity to accommodate a diversity of linguistic and cultural expression, which is integral to a globally interconnected academic environment. The lack of transparency in the system exacerbates these concerns. Learners and parents are left in the dark about how grades are assigned, fuelling mistrust in the system. Without clear explanations of how the AI functions and what criteria it prioritises, learners cannot effectively address or challenge the grades they receive. Transparency is essential not only for fairness, but also for building trust in AI tools. An XAI system could provide meaningful feedback on why a specific score was given, helping learners understand their performance and improve.

To address these challenges, educational institutions must adopt a more human-centred approach to AI grading (Topali et al, 2024). They should play an active role in validating AI assessment algorithms and AI-generated grades, ensuring biases are identified and rectified. AI should serve as a supplementary tool, not a replacement for human judgment, with final assessments incorporating qualitative feedback that values diverse linguistic and cultural contributions. Learners should also be informed about how the system works and provided with opportunities to engage in alternative assessments or appeal processes when discrepancies arise.

Legal primer

The system is a high-risk AI system under the AI Act, specifically Annex III, which categorises systems that evaluate learning outcomes in educational settings as falling into this class. A FRIA is likely required, as the AI grading system raises concerns about bias, fairness and transparency, potentially impacting learners' fundamental rights and equal opportunities. The lack of clear explanations for grading decisions breaches transparency obligations (Article 13, AI Act), which require providers to develop AI systems that enable accessible information about how the system operates, as well as its limitations and risks. Furthermore, students and parents have a right to an explanation under Article 22 of the GDPR. The system lacks meaningful oversight (Article 14, AI Act). Further, the disadvantage to non-native English speakers demonstrates systemic bias, breaching the data governance requirements (Article 10, AI Act), which require representative and non-discriminatory training data.

The grading system may operate through an online platform, especially if the university uses a broader digital learning management system (LMS). If this is the case, the DSA's provisions on algorithmic transparency and user rights are applicable to online platforms (Article 24, DSA). The DGA is relevant if the grading system leverages shared data sets or works with trusted data intermediaries to manage data for training the AI model. The Data Act focuses on secure data sharing and interoperability. It applies if the grading system

shares data with other systems (e.g., reporting platforms or institutional databases), or relies on external data sets for its training and evaluation processes.

Technical primer

Automatic grading of free text, such as that in student essays, is a challenging task for AI systems and algorithms. This is because it relates to a complex set of accepts, including authenticity, background, and lexical styles and capabilities. An AI grading system needs to function on a semantic rather than a word level. In other words, it needs the capability to recognise the meanings of the phrases, and not simply analyse the vocabulary used in them.

Educators and education leaders who adopt an automatic grading system require the developers to provide to clear and understandable description of how the system predicts evaluation values that represent the "correctness" or "authenticity" of the essay. To that end, XAI approaches and techniques that shed light on the textual features that play a role in generating model predictions are especially useful for this use case, since they offer insights into feature importance and the overall behaviour of the model.

For example, feature-relevance approaches in XAI, such as permutation feature importance (PFI) and partial dependence plots (PDP), hold the potential to clarify for educators and education leaders if there are dominant features the system is relying on to determine the essay's grade. PFI measures the impact of individual features (e.g., grammar quality, structure, vocabulary richness) on the model's predictions by permuting (randomising) a single feature's values and observing the resulting change in model performance. Thus, it identifies whether certain features that are related to learner background are influencing the model's output, creating biased predictions. In this use case, educators should be able to request from the providers PFI explanations regarding specific features, such as lexical diversity or grammar complexity, considering that these are among the features that reflect learner-background diversity. In the same line of explanations PDPs can visualise the relation between features and the predicted output, i.e., the essay score. Their benefit is that they show the model's behaviour over a range of values of the feature. Considering the grammar complexity feature, for example, a PDP can show if increasing the complexity of the grammar in the essay always results in increasing the essay final score. This can be an indicator of a bias in the grading system.

However, the question of how to achieve a proper understanding is complex, mainly because there is a part of subjectivity on what a good explanation is or not. An understandable description can ultimately be subjective. Answering this question requires clear definitions and alignment among all stakeholders, as highlighted many times in this report.

It is also important to notice here that the bias of this model can result from the overall pipeline in which the model is trained and hosted. Training data that is mainly collected from one group of learners will result in a bias in the system prediction, favouring essays from this group. The task of finding out if this bias exists and clarifying the reasons for it and how it can be mitigated, is mainly a task of the AI system providers, who are aware of the data collection process, and required to ensure its transparency to educational stakeholders. Educational stakeholders, however, can provide important insights to providers on, e.g., criteria to evaluate and mitigate that bias, making it an interdisciplinary task in which educators and developers collaborate to detect and mitigate system bias.

Challenges

A major challenge is ensuring the training data is diverse and representative to avoid perpetuating systemic discrimination. A further challenge is that the system operates as a "black box", and therefore students, parents and faculty cannot understand or contextualise decisions, eroding trust and accountability. There is also a challenge when it comes to balancing the AI automation and human oversight function without creating operational inefficiencies.

Recommendations

The educational authority would need to cease using the system and engage stakeholders such as staff, parents and students in the review of the grading system. The educational leader would need to request global explanations to clarify how the system evaluates essays and request local explanations to aid with communicating to students why specific grades were assigned and which features influenced the outcome. Educators would need training on how to use the system to ensure robust human oversight. In addition, the university should establish safeguards to ensure alignment of the AI system with ethical and legal standards. Again, AI providers must supply interpretable explanations of their systems' decision-making processes, ensuring transparency about the features the model evaluates, offering training and documentation that explain how these features correlate with the grading. A collaborative approach involving all stakeholders would be needed to comply with specific legal provisions. It is important that educational institutions conduct comprehensive impact assessments, e.g., FRIAs (AI Act for high-risk systems) and DPIAs (GDPR for processing personal information), to evaluate risks and ensure compliance before implementation of these types of systems. In terms of data governance, developers must gather data from diverse sources to ensure it is relevant and representative and ensure data integrity to mitigate systemic bias (Article 10, AI

Act). Furthermore, developers must provide accessible mechanisms (e.g., dashboards) to provide educators and learners clear, concise explanations of grading decisions and detailed user instructions (Article 13). Developers must design systems with human-machine interfaces that allow educators to intervene or override decisions, and education leaders should actively monitor and use their judgement to ensure fairness (Article 14, AI Act). It is important that educational institutions adhere to GDPR principles, including data minimisation, when processing personal data (Article 5, GDPR). Further, institutions should anonymise data to protect personal data and limit collection only to what is strictly necessary for the intended purpose. Additionally, learners and parents should play an active role in managing consent for data usage. Developers must prevent dark patterns (Article 23a (1) and Recital 51(b), DSA) and ensure algorithmic transparency in the delivery of personalised content (Article 28, DSA). Regulators should oversee compliance with these provisions, while educational institutions play a role in prioritising platforms for the protection of learners, particularly minors, from harm.

2.4. How to implement responsibly

To provide simple and practical recommendations to the educational community with regards to implementing AI in education responsibly from a legal perspective, the following overview summarises the key ideas obtained from the previous use cases

Do not solely rely on the tool	Communicate clearly with learners	Train educators and staff	Put safeguards in place	Evaluate the tool regularly
AI tools should support, not replace, educator judgment. Always review flagged work manually before making decisions.	Tell learners how the tool works, what it does, and what it does not do. Be transparent about its role in grading or review.	Ensure staff understand the tool's purpose, limits, and how to interpret its results. Provide guidance on handling false positives.	Create an appeal process so learners can challenge unfair outcomes. Document decisions and ensure reviews are fair and unbiased.	Check if the tool is flagging work fairly across all learner groups. Ask the provider for accuracy metrics and explanation features.

Table 8: Key ideas for responsible implementation of AI in education.

2.5. Key takeaways and implementation concerns

Explanations will be critical to successful AI educational implementations, because there must be clear feedback from educators and learners using the systems about whether they are working. This feedback will partly come from explanations they construct together about the systems they are using. To mitigate risks, there must be transparent communication that allows concerns to be raised by educators and learners. It is important that AI systems that are difficult for educators to operate, and would detract from their core teaching tasks, are not introduced. There are some tensions here. If AI enables educators to provide targeted support to learners, it could do their work more effectively and efficiently. However, their time could easily be taken up by systems that are difficult to operate, or by trying to explain the system to learners. Explanations of AI that educators are trained in giving must not be for specific products (which would lead to strong dependencies on specific systems) but rather directed at the operations of systems in general. Educators will need training and support, and this must be a factor in schools' technology adoption. If there is widespread deployment of AI systems in education, this must be part of educator training, both pre- and in service. An issue with XAI is that it may lead people to believe explanations will solve problems they themselves cannot, or that explanations are more important than the problems being solved. There must be clear use cases from schools that show that educational outcomes are improved when AI enters the classroom.[17]

The following table outlines key areas that educational institutions, developers, and policy-makers should consider when adopting AI systems. These categories reflect shared priorities across different levels of the education system. Together, they offer a **realistic and balanced roadmap for responsible AI adoption in education.**

[17] See further policy recommendations from EDEH workshop on XAI in education on 17-18 October 2024 in Brussels.

Item	Category
Establish clear feedback mechanisms: • **Develop feedback channels for stakeholders to share insights on AI system performance.** • **Ensure explanations are co-constructed by stakeholders to enhance mutual understanding of system functionality.**	Feedback mechanisms
Design educator-friendly systems: • **Avoid systems that impose difficult operational requirements on educators, detracting from core teaching functions.** • **Adopt systems that are self-explanatory and do not require educators to spend an inordinate amount of time explaining how they work to learners.**	Educator-friendly systems
Provide generalisable training for educators • **Train educators to explain the scientific workings of AI systems, and avoid product-specific training to prevent dependencies on specific products (vendor lock-in).** • **Incorporate training into pre-service and in-service professional development programmes.**	Training
Focus on educational outcomes • **Evaluate and adopt AI systems based on clear, demonstrable use cases that align with improving educational outcomes.** • **Balance the importance of explanations with the actual problems AI systems are designed to solve.**	Educational outcomes
Support effective technology adoption • **Include training and support requirements in the decision-making process for adopting AI systems in schools.** • **Ensure ongoing professional development resources are available for educators to adapt to AI deployments.**	Technology adoption
Mitigate risks and address tensions • **Create channels for raising concerns about AI systems, ensuring they do not overburden educators or hinder their teaching efficacy.** • **Regularly review and adjust AI implementations to ensure they enable, rather than detract from, effective teaching and learning.**	Risk mitigation
Promote clear and transparent use cases • **Showcase successful implementations of AI systems to build trust and confidence among stakeholders.** • **Use pilot programmes to demonstrate measurable improvements in educational processes and outcomes.**	Use cases

Table 9: ACE checklist: AI Compliance for Education.

3. XAI in education from the perspective of different stakeholders

3.1. Background

Once the legal issues of XAI have been analysed in the previous chapter, and before proposing the competences required by educators to integrate it with confidence in chapter 4, it is now time to illustrate what it means to include explainability in education from a practical perspective. To this end, the current chapter faces the *understandability* of XAI established in section 1.3., meaning that different end-users need to properly comprehend the provided explanations in order to support their trust into AI systems. The stakeholders defined in section 1.5. are of key relevance in this chapter, as their individual perspectives will be the focus of the analysis.

The impact of XAI is explored through two educational AI tools, specifically **intelligent tutoring systems (ITS) and AI-driven lesson plan generators (LPG)**, which are designed to enhance personalised learning by adapting to individual learner needs and supporting educators in creating tailored instructional content. These tools aim to address the diverse needs of learners and educators, though their full potential is still evolving.

Various ITS aim to support personalised learning by adapting to individual learner needs, identifying learning gaps, and providing tailored feedback or content to suit different abilities, including struggling or advanced learners, and those with special needs. It also seeks to enable self-paced learning while offering educators insights into learner progress for targeted intervention. Similarly, AI-driven LPGs aim to assist educators by creating curriculum-aligned, differentiated instructional content taking into account learners' diverse AI competence levels and preferences. These applications are promising in their potential to save educators time, support inclusive teaching, and provide strategies tailored to the specific educational context.

Specific scenarios in the following sections illustrate how these tools operate in real-world settings, showcasing how XAI is necessary to enhance transparency, foster trust, and improve decision-making, ultimately providing meaningful support for both teaching and learning processes. But before moving to the scenarios, the next section provides a short introduction to the explanation's format, with the aim of highlighting its importance to support understanding.

3.2. Visual explanations

By implementing measures to introduce explainability, AI systems have the potential to become transparent and trustworthy tools that support educational development. The complexity of creating such transparency lies in balancing the diverse perspectives of different stakeholders each with unique concerns and expectations about how AI can effectively support education.

The format in which the explanation is provided to the end-users is a key feature. The main modalities are simple text, visualisations and verbal explanations (Minh et al, 2022; Johnson et al, 2023). The scientific evidence highlights the effectiveness of visual explanations (Sedrakyan et al, 2019; Bovek & Tversky, 2016). With them, the information is represented through a graphical interface based on pictures, graphics, schemes, and other more abstract representations of data, like diagrams, plots, charts, networks, and others (Munzner, 2014; Sahin & Ifenthaler, 2021). Obviously, visualisations can also contain text, although the goal in the scope of education is to avoid long and descriptive explanations, but to move to simpler and conceptual ones.

The advancements in the field of graphical dashboards for education over the last years have been remarkable (Sahin & Ifenthaler, 2021; Bull, 2020), from which XAI can benefit. For a deep review about visualisations on XAI, please refer to Alicioglu & Sun (2019), and in the specific case of education to Ooge (2023). Their relevance in the scope of this report will be illustrated with two real examples, one focused on learners and the other on educators. They both correspond to AI-powered learning systems, but the aim here is not the application itself, but illustrating the possibilities provided by visualisations when it comes to tailored explanations and thus to understandability.

The following images shown in figure 4 correspond to an AI-based e-learning platform for secondary school which assigns adapted exercises to the learner's level (Ooge, 2023):

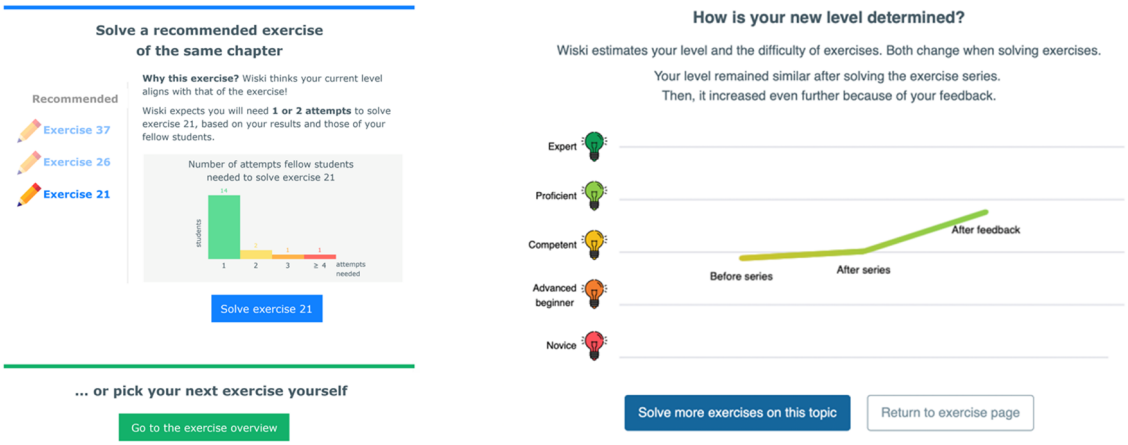


Figure 4: AI-based e-learning platform assigns exercises to the learner.
Source: Ooge, 2023.

The left one displays text at the top, which answers why a specific exercise (n. 21) was selected before providing a more detailed justification. The graph below clarifies the explanation by using group data and a histogram representation. The right image shows a visualisation of students' steering impact after an exercise series, with a text explanation about the progress at the top, and a graphical explanation at the bottom. As can be observed, the explanations are adapted to the students' age, the writing style, the colours used, etc.

Regarding educators, the following three images shown in figure 5 correspond to Santa, a multi-platform tutoring service for English learning based on AI:

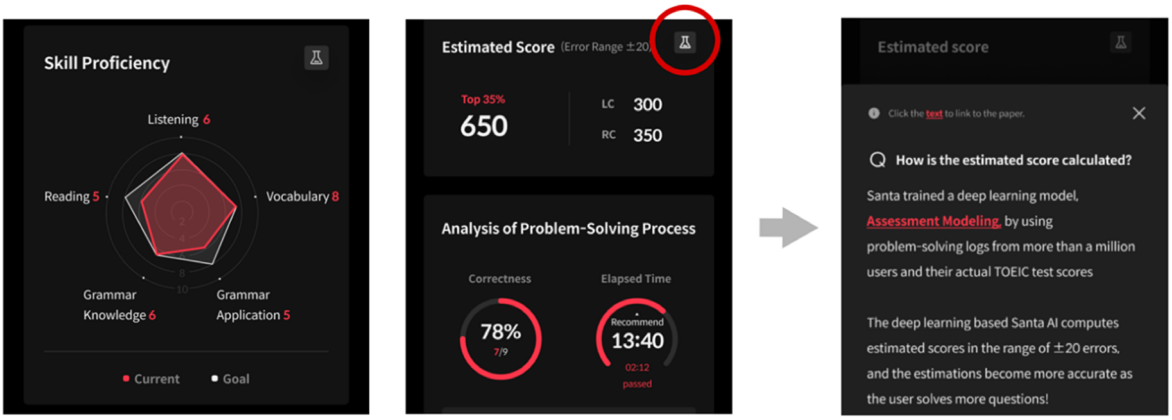


Figure 5: Explanations of estimated scores in the Santa tutoring system.
Source: Kim et al, 2020.

The left image shows a radar chart, which displays the student's proficiency in the different learning perspectives analysed by the tool. The middle image shows an estimated score obtained using a deep learning technique (Kim et al, 2020). When the teacher navigates to the top right icon, an explanation of the score calculation is provided, as can be seen on the right image. With this information, the teacher has more information about the confidence level of the AI system when predicting the score.

With these two examples, the potential of graphical visualisations in XAI is clear. However, it is also evident that much work remains to be done to achieve explanations that include all the features recommended in table 3, and that are useful for educators and learners.

3.3. Use case 1: AI-powered intelligent tutoring system

To provide users (AI deployers) with actionable insights, AI providers must integrate explainability features into ITS tools. These enhancements may empower users to validate AI system recommendations, intervene effectively, and refine the system for better outcomes, ensuring AI system fulfils its potential as a transformative educational tool.

Scenario: implementing AI-powered adaptive learning for mathematics in primary school

A primary school has recently implemented an AI-powered intelligent tutoring system (ITS) to support students' mathematics learning. The system adapts to each student's abilities, providing personalised learning materials and real-time feedback. The school aims to improve the achievement of learning outcomes, optimise teacher workload, and ensure all students—regardless of their learning needs—are receiving appropriate support. The system offers insights into student progress via a teacher dashboard and provides reports to parents and school administrators. Under the AI Act, this scenario would be considered a high-risk AI application (Article 6(2), read with Annex III(3)(b)), owing to the AI system's role in educational decisions, its data-driven personalisation, and the potential impact on student's learning outcomes.

Emma, a 10-year-old student, starts using the system for her math lessons. Initially, Emma was struggling with fractions and basic algebra, but with the ITS, she receives personalised tasks tailored to her skill level. The system identifies gaps in her understanding and adapts the difficulty of the exercises based on her performance. From the student's perspective, the AI-powered ITS in math feels like a personal tutor, presenting lessons and activities tailored to their skill level. The ITS provides real-time hints and immediate corrections to support students to reflect about mistakes and improve their content knowledge. However, it might not always be clear why certain exercises are suggested or how the system is assessing their skills.

Student perspective

By understanding why they are being assigned certain materials, the students feel a sense of ownership and trust in the system's recommendations. This aligns also with AI principles of explainability to foster student trust. To effectively engage students in their learning path and promote a self-directed learning approach, transparency in the relevance and progression of assignments is needed.

Transparency and explainability: For students to gain trust in the ITS, it is important to comprehend why certain exercises are assigned and how their performance is assessed. Transparency and tailored explainability (adjusting the tone and complexity according to the age) are core requirements for students to understand the relevance and purpose of each task and fostering a sense of control and trust in the system. It is important to know how the recommender system algorithm works, and what methods are used to keep students' attention (dark patterns risk).

Promoting engagement and ownership: When students comprehend the rationale behind their learning path, they are more likely to feel engaged and motivated. Explaining how tasks are personalised and recommended to address their specific needs helps students to take ownership of their progress and appreciate the system's value (Maity & Deroy, 2024).

Balancing support and independence: While the ITS acts as a supportive guide, it must encourage students to develop independent problem-solving skills. By gradually reducing the level of guidance as students' achievements improve, the system can help build students' confidence and self-reliance in tackling more complex challenges through adaptive scaffolding (Liu et al, 2024). How this support could be balanced and if there is opportunity for human intervention in the process, needs to be explained to students and teachers (Ogata et al, 2024).

Academic integrity: Additionally, it is important that students are well informed about appropriate and ethical use of ITS and the expectations they should have without rejecting the teacher's role (Hong et al, 2022).

Self-directed learning: An ITS should empower students to set personal goals, reflect on their progress, and customise their learning paths based on interests and needs. Incorporating opportunities for reflection, peer collaboration, and ethical awareness involve students in activities promoting metacognitive skills and critical thinking (Majumdar et al, 2023).

Teacher perspective

From the teacher's perspective, it is essential that the AI-powered ITS can be assessed as effective and trustworthy regarding the teaching approaches and learning design on which it is trained, as well as the effective support that it provides to all students with respect to their special educational needs and way of learning. The provision of quality education that follows a holistic approach and aims at the intellectual, emotional and social development of the student with respect to their rights, is of primary importance.

Human oversight: Developers and ITS providers must ensure embedded human oversight by design so teachers can intervene with ITS decisions and "manually" assign tasks or make changes to the learning paths.

Personalised learning and explanations: The ITS providers need to give explanations regarding the teaching approach and learning design embedded in the ITS. How does the ITS work to monitor each student's performance on tasks, suggestions for improvement, prompts for self-monitoring and affect-level comments.

Transparency and explainability: The cultivation of the teacher's trust in the ITS can be ensured through a comprehensive understanding of its functions, as well as the educational and learning context in which it is designed. It is also important to ensure that the tool has considered the different socio-cultural and learning characteristics of the student population, such as language, culture and cognitive age level, to ensure that the knowledge provided is relevant to all students (ethics by design). Such information must be provided in the teacher's dashboard, and teachers should promote those ITS that include it and comply with the XAI dimensions displayed in table 4.

Academic integrity: The ITS should be integrated in the educational process to provide support to the teacher and not to replace them. It is therefore important that teachers inform and train students well regarding its appropriate and ethical use and the expectations they should have of it.

Respect of children rights: Finally, it is important for the teacher to ensure that the ITS is designed with respect to the child's rights, such as the protection of personal and sensitive data, self-expression and freedom of choice. Any use of the ITS must not in any way be detrimental to their safety and well-being.

Curriculum designer perspective

Curriculum designers are critical stakeholders in the implementation of ITS in education. These AI-powered systems promise to enhance learning by personalising content and adapting to individual needs. However, their effectiveness depends on ensuring alignment with curricular objectives, providing equitable learning opportunities, and addressing diverse learner needs.

Understanding ITS decisions: Know why specific tasks are assigned and how they align with curricular standards. ITS should have a "transparency mode" that shows which input data were used (e.g., test scores, previous task performance), and logical pathways that describe how its performance metrics are linked with the task assignment. For example, opening a task may provide the following explanation: "This geometry problem was selected because the student demonstrated 80% proficiency in prerequisite algebra skills."

Detecting bias: ITS providers should enable identifying and mitigating any systemic inequities in recommendations or task assignments. Users should be supported with system functionalities which will flag potential biases, such as uneven distribution of advanced tasks across genders or socioeconomic groups. For example, a bias alert may state: "Female students receive 30% fewer advanced tasks than male peers with equivalent performance."

Supporting personalisation: ITS providers should explain how content is adapted to learners with challenges such as ADHD, dyslexia, or language barriers.

Enabling transparency: ITS providers should prepare actionable insights that foster trust among educators, parents, and administrators. A feasible way to achieve this, would be to include the XAI dimensions displayed in table 4 with the features shown in table 3.

Enabling transparency: ITS providers should prepare actionable insights that foster trust among educators, parents, and administrators. A feasible way to achieve this, would be to include the XAI dimensions displayed in table 4 with the features shown in table 3.

Educational leader perspective

The primary responsibility of an educational leader is to ensure the purposeful implementation of the ITS in alignment with the established educational objectives, ethical standards, and existing policies. This includes promoting equitable learning opportunities, fostering stakeholder trust, and ensuring compliance with national and international regulations, such as those mentioned in the previous chapter.

Alignment with institutional and policy goals: The implementation of an ITS should align with institutional priorities, such as digital citizenship, homework, and screen time policies. Explainability ensures that the system provides clear insights into how tasks are assigned in accordance with these goals. For example, the ITS should transparently show how it adjusts homework assignments to meet time limits while supporting learning objectives.

Equity and accessibility: Educational leaders should ensure that the ITS promotes equitable learning opportunities and adapts to the diverse needs of students. The ITS should explain the logic behind task recommendations, making it possible to identify and address biases or inequities. For example, reports generated by the ITS should clearly explain how tasks are personalised for students with disabilities or language barriers.

Tailored explanations: Fostering those ITS that comply with the XAI dimensions displayed in table 4, will allow that students like Emma and her teachers using the tool can receive understandable explanations.

Stakeholder training and readiness: Teachers and students should receive sufficient training from educational authorities so they may comprehend how the ITS functions and affects them. Without this, stakeholders may lack confidence in the ITS or misinterpret its functionality. This issue is addressed in more detail in the next chapter.

Data privacy and ethical use: Adhering to stringent data protection policies requires transparency in how the ITS collects, stores, and uses student data. For example, the ITS should explicitly outline what data is collected, its purpose, and how it supports personalised learning.

Monitoring and continuous improvement: Educational leaders should continuously evaluate the performance of the ITS. The tool should communicate how new algorithms improve task difficulty adjustment, enabling leaders to align these updates with institutional goals and stakeholder expectations.

Policy-maker perspective

For policy-makers, it is essential that ITS in primary schools are transparent, trustworthy, and used responsibly. These tools can significantly shape young learners' experiences and development, so clear guidelines on privacy, fairness, and accountability are vital to support their ethical deployment.

Clarity in adaptive learning choices: Policy-makers should require ITS providers to provide clear, accessible explanations for their learning path adjustments and task assignments. Transparent decision-making ensures that each student's learning path feels intentional, reducing frustration and fostering a sense of ownership over their progress. To this end, policy-makers should establish requirements for a common and comprehensive approach in terms of explainability, as suggested in the insights from the EDEH community workshop on explainable AI in education.

Data privacy and security: Given the sensitivity of young students' data, ITS systems must adhere to strict protocols for data collection, storage, and use. Policy-makers should require ITS providers to clearly outline what data is collected, its intended use, and who has access.

Fairness in task recommendations: AI-driven tutoring systems must operate equitably, avoiding biases that could favour or disadvantage certain students. Policy-makers should establish regular fairness checks within ITS systems, ensuring that learning adaptations remain impartial across diverse backgrounds and abilities. Regular fairness checks could include examining how tasks are assigned to diverse student groups, ensuring adaptations remain tailored and justified.

Accountability and human oversight: Clear accountability structures are vital for ITS that autonomously adjust learning paths. Policy-makers should specify who monitors these systems, who is accountable for potential harm and how human oversight and interventions are ensured. This includes, for example, making sure that educators or administrators are able to step in if AI-generated suggestions do not suit the needs of the students.

Encouraging AI literacy for teachers and parents: Transparency is enhanced when parents and teachers understand the AI's role in learning. Policy-makers can support training programmes that equip parents and teachers to engage thoughtfully with ITS, empowering them to question or adjust recommendations when necessary.

Developer perspective

From the developer's perspective, creating an ITS for primary school mathematics requires a focus on explainability, adaptability, and fairness across the AI lifecycle. The goal is to deliver a personalised, effective, and transparent learning tool that meets ethical standards and supports educational outcomes.

Ensuring data provenance and integrity: Developers must establish robust data pipelines to manage the accuracy and contextual relevance of inputs like student performance metrics and learning histories. Using data lineage tracking and automated validation, developers ensure data integrity while maintaining compliance with frameworks like the GDPR and the Family Educational Rights and Privacy Act (FERPA). Provenance dashboards should provide real-time insights into data flows, helping educators understand how inputs influence task recommendations and building trust in the system.

Building explainable recommendations: Developers must use XAI techniques, allowing teachers to understand why specific tasks are assigned. They must comply with the regulations established in chapter 2 in terms of transparency, trying to apply AI techniques that foster interpretability.

Facilitating real-time feedback and adaptability: Real-time adaptability requires robust decision-making systems capable of processing live data from student interactions. Event-driven architectures and sequential recommendation models enable the ITS to adjust task difficulty dynamically. Developers need to create interactive dashboards showing decision trees or adaptation flowcharts to help education stakeholders understand how changes are made and allow manual overrides when necessary.

Embedding fairness and bias detection: Developers must integrate fairness audits and bias detection mechanisms to ensure equitable learning opportunities. Techniques like demographic parity checks and adversarial debiasing (Elazar & Goldberg, 2018) can identify and mitigate biases in recommendations. Tools that visualise demographic trends, such as heatmaps of task allocations, help developers detect disparities and refine the system, aligning with ethical guidelines like UNESCO's Recommendation on the Ethics of AI for education principles.

Integrating user feedback for continuous improvement: Feedback loops are essential for refining the ITS. Developers should implement interfaces to collect education stakeholders' inputs on task relevance and effectiveness, using NLP and clustering algorithms to analyse feedback trends. Developers should use insights from this process to inform model retraining and system updates, ensuring the ITS evolves to meet real-world needs.

Key dimensions of XAI in the ITS use case

The following table exemplifies, for this specific use case, the different dimensions of XAI included in table 4. The explanations included in the table should be created according to the features shown in table 3 and adapted to the different stakeholders to be properly understood. But even with such general examples, it is important to emphasise the relevance of all dimensions. Developers should consider these dimensions when designing their ITS.

Dimension	Example
Scope	**Global:** Explaining overall trends, such as why the ITS reinforces some topics on algebra for the majority of students based on curriculum analysis. **Local:** Explaining why Emma is assigned a specific exercise on fractions, based on her previous mistakes and performance trends.
Depth	**Comprehensive:** A detailed report for teachers showing how Emma's progress in algebra has improved over time and what specific factors contributed. **Selective:** A quick explanation for Emma's parents about how the ITS identified her difficulty with fractions and adapted her tasks accordingly.
Alternatives	**Contrastive:** Explaining why Emma received fraction exercises instead of basic arithmetic by showing her performance gap in fractions. **Non-contrastive:** Showing factors that the ITS used, such as low quiz scores, without comparing alternatives.
Flow	**Conditional:** "If a student scores below 70% on fractions exercises, then recommend more tasks focusing on conceptual understanding." **Correlational:** Displaying how increased practice time correlates with improvement in Emma's fraction scores, helping educators understand her progress.

Table 10: Dimensions of XAI in the ITS use case (examples).

3.4. Use case 2: AI-powered lesson plan generator

Scenario: creating a lesson plan on fractions for middle school

At a middle school, teachers recently started using an AI-powered lesson plan generator (LPG) to support teachers in their lesson preparation. The AI tool analyses curriculum goals, expected learning achievements and some anonymised data regarding the class in general (like general description of students' background and abilities, without any specific or personal data).

The AI tool may also use and analyse student data (such as performance on recent assessments) and preferences (such as hands-on activities or visual aids) which raises the risk level of the AI system and makes additional evaluation prior to the use necessary. Based on this input, it generates a customised lesson plan

using a generative AI model that includes a variety of activities to fit the learning needs of the students. The LPG also suggests digital resources, estimated time for each activity and formative assessment options. Teachers can adapt the plan before implementation to ensure it fits their teaching approaches and classroom needs. The school aims to lessen teacher workload, adjust to students' different learning needs, and to increase the engagement of all students in the classroom. Under the AI Act, this scenario (without student data) would be considered as low or no risk because the teacher would be the one who decides about the use of the lesson plan generated by the LPG. Even in such low-risk cases, developers are obliged to ensure the system adheres to principles of XAI. This includes designing the system to provide clear explanations of how it processes inputs and generates outputs. Such explanations are necessary to verify whether the generated text is accurate and appropriate for the intended use. Additionally, developers must address potential biases in content generation to avoid perpetuating stereotypes or unfair assumptions. By ensuring transparency and enabling accuracy checks through explanations, developers help build trust and empower teachers to make informed and effective use of the tool.

Ms. Lee, a middle school teacher, uses the LPG tool to create a lesson plan on fractions for her mixed-ability maths class. She inputs anonymised class performance data from recent assessments and her preferences for interactive learning. The system quickly generates a detailed plan suggesting activities that can engage students at different levels. For students who struggle with basic math concepts, the tool suggests interactive fraction games, advanced problem-solving tasks for high achievers, and group work for collaborative learning. Before finalising the plan, she adapts a part of the game to fit her teaching strategy, and she adjusts the scope of the group work due to her workload. From the teacher's perspective, the LPG feels like a personal assistant, creating lessons and activities fitting their pedagogical approach and engaging students at different levels. It provides real support to teachers in their lesson preparation, although, the lack of clear explanations of how the tool works could arise several problems. For example, the ambiguity in the decision-making process might prevent Ms. Lee from understanding why certain activities are recommended, potentially limiting her ability to adapt the lesson to her students' specific needs.

Student perspective

By understanding how this AI system works and what they have to expect from it, the students feel a sense of trust in the teacher's recommendations which align with the principles of XAI. Thus, it is important for students to be well informed about AI system use and its functions, understanding that their own intervention is ensured as well as the protection of their rights.

Data protection: Students and parents must be informed if students' data are used in such tools, how it is used and other aspects of data protection guaranteed by the GDPR and other EU regulations. The LPG only uses anonymised data about the student's progress, but even in this case, the provision of this information is required.

Differentiation for diverse learners: The tool may differentiate instruction by suggesting activities tailored to various student needs. For example, a lesson plan may contain suggestions for more motivating activities for struggling students, advanced tasks to engage high achievers, or group work to foster collaboration between students of different levels. This targeted approach may support equitable learning opportunities, but it could also introduce undesirable effects like the widening of gaps instead of mutual aid. The strength of XAI here is that it supports the re-examination and breaking of habits in case the student detects a barrier or bias (Jauhiainen & Guerra, 2023), thus opening the door to collective and self-reflection on teaching practices and learning strategies.

Student's agency: While the AI assists the teacher in generating the plan and suggests engaging activities and resources, human oversight must be included, creating an option for students to comment on those activities and resources and give feedback if those activities are well adjusted to their needs.

Teacher perspective

From the teacher's perspective, it is essential that the LPG tool can be assessed as effective and trustworthy regarding its functions, as well as the educational and learning context in which it is designed. Especially considering the different abilities of the students, it is important to ensure that the AI tool will unbiasedly examine and assess each student's performance and preferences and respect their rights and needs through a holistic approach.

Customisation and adaptability: The AI tool may create a tailored lesson plan based on curriculum goals, Ms. Lee's teaching preferences and potentially on student performance data. This would ensure that the plan addresses the diverse needs of the mixed-ability class while aligning with her teaching style. Additionally, flexibility is also important, ensuring her ability to review, adapt, and/or modify the plan accordingly to better suit her classroom dynamics. To this end, as discussed in the next chapter, building the capacity of teachers, to know how to use generative AI while maintaining agency and control is key.

Efficiency in planning: The AI tool may streamline the lesson-planning process by analysing data and providing a structured plan, complete with resources, timing, and formative assessment options. This may save Ms. Lee's time and let her adjust the AI-generated lesson plan, making it a well-rounded lesson.

Teacher agency and control: While the AI tool assists in generating the plan and suggests engaging activities and resources, Ms. Lee must remain in control of the final decisions. This balance ensures that the tool enhances her teaching rather than undermining her professional expertise.

Personalised learning and explanations: The AI tool may provide explanations regarding the teaching approach and learning design according to each student's performance on tasks, suggestions for improvement, prompts for self-monitoring as well as affect-level comments. Additionally, it should somehow ensure the equal and active participation of all students in the group as part of the collaborative learning.

Academic integrity: The AI tool was integrated in the educational process to enhance the lesson's dynamic making it more attractive as well as to provide additional support for both Ms. Lee and the students. It is therefore important that students are well informed and trained regarding its appropriate and ethical use and the expectations they should have of it.

Transparency and explainability: In order to ensure Ms. Lee's trust in the AI tool, it is important to gain a comprehensive understanding of its functions, as well as the educational and learning context in which it is designed. Especially considering the different abilities of the students, it is important for Ms Lee to ensure that the AI system will unbiasedly examine and properly assess the performance and preferences of each student in order to encourage and motivate them to actively participate (ethics by design).

Respect of children rights: It is critical for Ms. Lee to ensure that the AI system is designed with respect to the child's rights, such as the protection of personal and sensitive data, self-expression and freedom of choice. Any use of the system cannot in any way be detrimental to their safety and well-being.

Curriculum designer perspective

The curriculum designers' role is to effectively integrate AI-driven LPGs while ensuring transparency, equity, and inclusion. To this end, generative AI must be used properly, framing its output with links, texts or documents aimed to be followed, including the specific school curriculum. By focusing on explainability, these tools can enhance lesson planning processes while preserving teacher and designer control.

Customisation for mixed-ability classes: The system suggests activities tailored to students' varying abilities, engagement levels, and learning preferences. Without clear rationales for recommendations, designers cannot understand why specific tasks are assigned to certain groups. For example, the LPG may assign basic integer addition tasks to struggling students but fail to explain the criteria, leading to mistrust.

Alignment with standards: Tools ensure lesson plans meet the requirements of educational frameworks, curriculums or other guidelines prescribed by educational authorities. Designers, however, must currently manually validate AI outputs, increasing their workload and diminishing their creative role. For example, a designer may modify a gamified task suggested for low-engagement students by adding collaborative elements to foster peer interaction.

Cultural relevance: The AI-based LPG adapts content to reflect local and cultural contexts, improving engagement and relatability. Historical data used by the system may unintentionally perpetuate stereotypes, such as disproportionately assigning simpler tasks to specific demographics.

Bias mitigation: Advanced algorithms detect and flag potential biases, promoting equitable task distribution. AI tools should flag potential inequities in task assignments and suggest equitable alternatives. For example, bias alerts may notify designers if advanced tasks are disproportionately assigned to male students.

Efficiency: The system automates repetitive tasks, allowing designers to focus on refining and personalising content. Generic recommendations lack the adaptability to align with unique classroom needs. However, AI tools must provide culturally relevant examples and ensure tasks are tailored to individual strengths without reinforcing stereotypes. For example, an AI-based LPG replaces generic integer word problems with scenarios using local temperature variations or market data for a diverse class.

Educational leader perspective

From the educational leader's perspective, the primary responsibility regarding the creation of lesson plans using an AI-driven tool is to ensure transparency and explainability. These are vital for fostering trust, enabling customisation, and ensuring alignment with educational goals. Leaders must address the unique challenges and needs of teachers, students, and administrators to maximise the potential of the tool.

Rationale behind lesson customisation: Leaders need to ensure that the AI tool provides clear explanations of how its lesson plans are generated. For example, the tool should articulate why it recommends interactive fraction games for struggling students and advanced problem-solving tasks for high achievers. Transparent reasoning ensures that teachers and stakeholders understand and trust the differentiation process, avoiding scepticism about the tool's fairness or effectiveness.

Alignment with pedagogical goals and standards: It is critical for the AI tool to demonstrate how its suggested activities align with curriculum standards and institutional priorities. For instance, leaders should be able to verify that the tool adheres to specific grade-level expectations and learning objectives for fractions. Without transparency, there is a risk of plans that diverge from institutional requirements, creating inconsistencies in classroom instruction.

Support for mixed-ability classrooms: AI-generated plans must explicitly show how they cater to varying student needs, such as providing hands-on activities for kinaesthetic learners or scaffolding for struggling students. Leaders need transparency in the criteria used for tailoring content to ensure equitable access to learning opportunities. This clarity allows for more targeted teacher interventions when plans do not meet specific classroom needs.

Adaptability to teacher preferences: Leaders must ensure that teachers can easily identify and adjust the components suggested by AI to match their teaching styles. This transparency ensures that the tool enhances rather than constrains instructional flexibility, fostering greater teacher engagement with the system.

Effectiveness of suggested resources and assessments: Leaders should assess how well the AI tool justifies its recommendations for digital resources, time estimates, and formative assessments. Transparency in these areas ensures that the outputs are actionable and contextually relevant for teachers. For instance, a plan that includes a digital game should specify its expected impact on learning outcomes, enabling leaders to evaluate whether such tools meet institutional and pedagogical goals.

Policy-maker perspective

For policy-makers, it is essential that AI-powered LPGs are transparent, equitable, and responsibly used in education. These tools shape how lessons are structured and delivered, so clear guidelines on privacy, fairness, accountability, and collaboration are necessary to support their ethical deployment.

Transparency in lesson recommendations: Policy-makers should ensure that AI LPGs provide clear, understandable explanations for their recommendations. Teachers need to know why certain activities are suggested over others, especially when they differ to meet student needs. This transparency enables teachers to trust and engage confidently with the AI's suggestions. For example, the AI might briefly explain why a group activity is recommended to support collaborative skill-building, helping teachers understand the AI's rationale.

Data privacy and security for student and teacher data: Since LPGs may use both student performance data and teacher preferences, data handling protocols are required. Policies should specify what data is collected, how it is stored, and who has access, aligned with the GDPR and similar standards. Policy-makers need to ensure the AI adheres to data protection practices, giving teachers, students, and parents confidence in the system's secure and ethical use of information, even in a low-risk system.

Ensuring fairness in lesson customisation: AI LPGs must operate impartially, avoiding biases that could favour certain teaching methods or student groups. Policy-makers should advocate for regular fairness assessments to confirm that the AI adapts lessons equitably across diverse student backgrounds and abilities. This ensures that all students benefit from suitably customised lesson plans, creating an inclusive classroom environment.

Accountability and teacher oversight: Given that LPGs guide classroom activities, it is critical to have clear accountability measures in place. Policies should define who is responsible for monitoring the system and revising recommendations when necessary. Teachers should be able to review and adapt AI-generated plans to ensure they meet classroom goals. For example, if the AI suggests advanced tasks for a mixed-ability group, teachers should be empowered to adjust the plan to suit all students.

Supporting AI literacy for teachers in lesson planning: Policy-makers can enhance transparency by promoting AI literacy programmes that help teachers understand and interact effectively with LPGs. Training that explains how the AI makes recommendations empowers teachers to engage critically with the tool, allowing them to adapt lesson plans as needed to match their unique classroom needs. This literacy fosters a collaborative approach, where AI serves as a supportive resource rather than a directive.

Developer perspective

From the developer's perspective, developing a LPG for middle school educators like Ms. Lee, the focus lies on creating a tool that supports adaptable, explainable, and equitable lesson planning across the AI lifecycle. The goal is to ensure that the AI tool is transparent, adaptive, and fair, providing educators the possibility to deliver engaging, real-world lessons aligned with regulatory and ethical standards.

Explainable lesson plan recommendations: Educators need to trust AI-generated recommendations. Tools like SHAP or LIME, can clarify why specific activities are suggested, such as linking a drought-related activity to Ms. Lee's request for environmental applications in math. This transparency aligns with the EU's "right to explanation" and IEEE ethical standards, ensuring educators understand and trust the AI's reasoning.

Contextually relevant data collection and processing: Developers must ensure the tool integrates curriculum standards, prior student performance, and real-world examples (e.g., climate change scenarios) to craft resonant lesson plans. For instance, using drought data in a math lesson on fractions helps contextualise abstract concepts. However, balancing data integrity and relevance is a challenge. Adhering to the GDPR and ethical standards ensures transparency in data sourcing, while provenance tracking systems visualise how contextual data impacts lesson suggestions.

Real-time feedback and context-based adjustments: The generator must allow on-the-fly adjustments based on engagement metrics. For example, if static exercises lose student interest, the AI might suggest switching to an interactive flood simulation. Transparency by application design in how engagement data informs these adjustments supports the OECD's AI Principles, enabling educators to manage lessons dynamically and effectively.

Bias detection and fairness auditing: Ensuring equitable learning requires addressing biases in data or algorithms. Regular fairness audits can highlight disparities, such as overemphasis on urban or rural examples. Aligning with UNESCO's ethical AI principles, developers can implement tools to monitor and adjust content distribution, ensuring inclusivity across diverse student backgrounds.

User feedback and iterative improvement: Continuous refinement based on educator feedback is crucial. A simple feedback mechanism lets teachers rate AI-recommended activities, such as a flood-related math lesson's relevance. Insights from this process guide developers in enhancing the tool's adaptability and contextual alignment, embodying a "human-in-the-loop" approach per European Commission guidelines.

Key dimensions of XAI in the LPG use case

The following table exemplifies for this second use case the different dimensions of XAI included in table 4. Again, the explanations shown in the table should be created according to the features displayed in table 3 and adapted to the different stakeholders to be properly understood. Developers should consider these dimensions when designing their LPG.

Dimension	Example
Scope	**Global:** Explaining why the LPG prioritises hands-on activities for certain topics based on broad curriculum goals. **Local:** Explaining why a specific visual aid was recommended for Ms. Lee's fractions lesson, given her students' performance data.
Depth	**Comprehensive:** Providing Ms. Lee with a detailed explanation of how the LPG combines curriculum standards, assessment data, and preferences to generate plans. **Selective:** A quick note explaining why the tool suggested group work for struggling students.
Alternatives	**Contrastive:** Highlighting why the LPG suggested a visual fractions game instead of a lecture format, based on Ms. Lee's interactive learning preference. **Non-contrastive:** Listing the main factors (e.g., student engagement scores) considered without comparing alternatives.
Flow	**Conditional:** "If students perform below the expected level in fractions, then include a review activity before introducing new concepts."

Table 11: Dimensions of XAI in the LPG use case (examples)

3.5. Stakeholder's intervention level and points of attention

Following the two previous scenarios, it becomes clear that explainability plays a crucial role in ensuring AI tools like ITS or AI-driven LPGs are effective, trusted, and actionable in real-world educational settings. While these tools aim to enhance learning experiences, tackle challenges, and personalise instruction, their full capabilities depend on providing clear explanations that ensure both accuracy and alignment with educational goals. The following table summarises the intervention levels and key points of attention for the main stakeholders in education obtained from the analysis of the two previous use cases, with the aim of being useful for readers in other similar AI-driven tools in education.

Dimension	Example	Points of attention
Students	• Direct engagement with personalised content, activities, and task assignments. • Active participation and immediate feedback fostering self-directed learning and agency.	• Clear and personalised explanations with accessible dashboards. • Transparency in task assignment and AI decision-making. • Ethical handling of data and support for learner agency.
Teachers	• Oversight and validation of AI-generated recommendations, lesson plans, and activities. • Active involvement in adapting AI outputs to suit classroom contexts and detect biases.	• Ensure fairness, inclusivity, and alignment with learning objectives. • Maintain ethical oversight with editable outputs and clear standards. • Promote AI literacy and mitigate biases through manual intervention.
Curriculum designers	• Alignment and quality control of AI outputs with established curricular standards. • Monitoring and correcting AI-generated content for bias and ensuring consistency with policy guidelines.	• Emphasise transparency in content creation and curriculum alignment. • Address biases and promote equity, diversity and inclusion. • Support interactive feedback mechanisms and build AI literacy into curriculum design.
Educational leaders	• Oversee institutional implementation of AI systems while ensuring compliance with educational policies. • Facilitate professional development and monitor overall AI performance.	• Ensure transparency and ethical compliance through institutional dashboards and audit trails. • Maintain equity and accessibility. • Support continuous teacher training and stakeholder feedback loops.
Policy-makers	• Provide regulatory oversight and enforce fairness in AI decision-making. • Ensure that AI systems operate transparently and are aligned with broader public policy goals.	• Safeguard data privacy and protect citizen rights. • Mandate accountability, fairness audits, and risk management in AI implementations. • Promote citizen AI literacy and ethical standards across systems.
Developers	• Develop AI systems that deliver personalised recommendations by integrating XAI techniques. • Build robust data pipelines with lineage tracking and real-time dashboards, enabling dynamic adaptations. • Facilitate real-time feedback loops and manual overrides to refine recommendations based on live student interactions and educator inputs.	• Ensure accurate data provenance and integrity while complying with GDPR, FERPA, and other relevant frameworks. • Embed regular fairness audits and bias detection mechanisms supported by visual tools. • Maintain clear, interpretable outputs and contextual explanations that build trust among all education stakeholders. • Use stakeholder feedback, analysed to guide continuous system improvements.

Table 12: Key points of attention and interactions with the AI system for the main stakeholders in education.

3.6. Ensuring human-centred explainability in AI for education: roles, responsibilities, and the need for oversight

Ensuring explainability in education goes beyond algorithms; it requires active engagement, ethical participation, and shared responsibility. Educators can interpret and contextualise AI outputs to align with individual learning goals, while learners must take ownership of their AI usage, demonstrating academic integrity by articulating how they engage with AI tools. Parents and school authorities benefit from explainable dashboards that provide transparency while respecting data privacy. Policy-makers depend on fairness audits and bias assessments to ensure equity, and developers must incorporate continuous feedback to build tools that remain transparent, adaptable, and aligned with stakeholder needs.

The incorporation of explainability by design into AI systems for education does not rest solely on the developers. It relies on companies, with their corporate and institutional choices. It is also a responsibility of suppliers and distributors, which must align the systems with market regulations the end-users' needs. Of course, it also depends on policy-makers who must facilitate the integration of the required educational perspectives to developers.

While improving the technical aspects of explainability is essential, focused on transparency and interpretability of the models, human oversight remains critical to ensure AI tools are meaningful, reliable, and actionable. Specific XAI algorithms alone cannot fully address the complexity of real-world educational contexts, as they often lack adaptability and depth. Human judgment is indispensable in validating, interpreting, and contextualising AI outputs, ensuring they are transparent, ethical, and aligned with educational needs.

Achieving meaningful AI explainability in education requires a collaborative ecosystem where educators, learners, parents, developers, and policy-makers actively engage in continuous dialogue. This effort must go beyond technical solutions, focussing on human-centric design and shared responsibility. In such a scenario each stakeholder plays a critical role with the common goal to create an adaptive, reflective system that prioritises human judgment, keeping academic integrity, and ensuring that AI tools remain supportive instruments that could enhance educational understanding. By fostering open communication, regular feedback mechanisms, and a commitment to ethical innovation, the community of stakeholders can guide the development of AI technologies that are transparent, contextually responsive, and aligned with diverse educational needs.

4. Defining educators' competences for and towards XAI

4.1. Background

The introduction chapter outlined how the techniques and procedures developed in the field of XAI are oriented towards the end user, with the aim of achieving a proper trust level in AI systems. End users can be various stakeholders in the scope of AI, from developers to the general public, as all of them require a certain level of understanding of the AI system's output. However, in the realm of education, the main target are the learners. They must be trained to live in a world surrounded by AI, and more specifically, to use AI systems for learning, prioritising their agency and fostering critical thinking. Consequently, the first and most relevant stakeholder group when talking about XAI in education are the educators, as they are the ones responsible training learners accordingly.

In UNESCO's Recommendation on the Ethics of Artificial Intelligence, we can read that 'AI literacy and awareness are essential for all citizens to navigate and engage responsibly with AI systems' (UNESCO, 2022, p. 36). This means equipping learners and educators not only with technological knowledge or technical skills, but also with the capacity to constantly question and critique these systems in order to adopt these powerful tools as opportunities, but at the same time understanding the ethical implications, limitations and threats of AI systems.

As explained in chapter 2, the AI Act emphasises the importance of transparency and accountability, ensuring that AI systems are designed and deployed with clear guidelines for their use and potential risks (Article 3, AI Act). When talking about educating stakeholders on AI, it is crucial to include an understanding of how these systems can sometimes reinforce biases, create ethical dilemmas, and have far-reaching societal impacts, as explained in the DigComp framework. Without this deeper comprehension, learners and educators may lack the tools to navigate the complexities of AI, potentially leading to misuse or uncritical acceptance of technology.

This last chapter is focused on the competences that educators need for AI regarding XAI, both to use AI-based tools in their teaching and to teach the technical foundations and ethical implications of the technology. Using the UNESCO's AI Competency Framework for students (Fengchun & Kelly, 2024) as a reference, section 4.3. highlights the core educator competences directly related to XAI, according to the ISCED levels of education. A new scenario, however, must be faced: Regardless the level of education, educators will need to select the appropriate AI tool for their learners, the specific context, and the learning objectives. Therefore, in addition to the core AI competences for XAI, the educator requires specific competences, illustrated in section 4.4., to be able to evaluate the explainability features of the tools based on the key dimensions defined in table 4.

If educators put all these competences into action, they will be able to analyse and evaluate every XAI tool, use it and take advantage of its capabilities, creating learning experiences adapted to their specific case. In addition, they will have the capacity for providing learners with the fundamental skills for critical thinking towards AI systems, increasing their agency and safety. Representative examples of possible XAI implementation at different ISCED levels are presented in section 4.5., to conclude this chapter with a set of recommendations for the main stakeholders in the realm of AI literacy.

4.2 Foundations for AI regarding XAI

Critical thinking as the goal

In an AI-shaped era, ethical considerations and dilemmas require critical thinking more than ever. Therefore, learners, along with their educators, must be safeguarded against cognitive atrophy or manipulation, like those posed by generative AI, and are expected to be equipped with critical thinking attitudes, such as intellectual depth, reasoning based on data, information, and evidence, as well as confidence in reason as outlined in the Paul-Elder Critical Thinking Framework (Paul & Elder, 2006). Transparency and explainability of AI systems are prerequisites for practicing critical thinking. Without access to data, models, and algorithms our control over them and human agency is compromised.

The UNESCO AI competency frameworks place a strong emphasis on critical thinking as an essential skill for educators and learners in the context of AI integration in education. The frameworks advocate for a robust approach to ensure that AI systems used in education are not only effective but also transparent and accountable, fostering trust among educators and learners. Other remarkable initiatives, such as the *Transparency Index Framework* (Chaudhry et al, 2022), must be pointed out. This framework is deeply related to critical thinking by enhancing understanding and informed decision-making among stakeholders through data and algorithmic transparency. This clarity allows users to critically assess the implications of AI tools in educational settings and encourages a culture of inquiry by prompting stakeholders to ask questions about the systems they use.

An illuminating example on how it is so urgent to focus on the immense risks related to the opacity of some AI systems is The MIT AI Risk Repository which points out how algorithms, whose decision-making processes are opaque, can potentially lead to unintended biases or discriminatory outcomes. This could even jeopardize the very foundations of educational systems and their mission — namely, the mission of *educere*, which is to draw out learners' talents and enhance them through various forms of literacy, all rooted in the universally nurtured concept of critical thinking. Therefore, it is crucial to intervene in educational systems promptly,

enabling both learners and educators to adopt AI as a genuine opportunity to further develop critical thinking. This, in turn, should lead to research-driven and problem-solving approaches rather than passive acceptance of AI-generated outputs from simple prompts, which increasingly often lack clarity and a traceable origin.

AI literacy as the way

One of the most frequently cited definitions of AI literacy is 'a set of competences that enables individuals to critically evaluate AI technologies, communicate and collaborate effectively with AI, and use AI as a tool online, at home, and in the workplace' (Long & Magerko, 2020, p. 2). In other words, being AI literate means to be able to effectively use, monitor and critically reflect on the AI tools in personal, professional, or educational contexts. Consequently, it is a way to achieve the critical thinking goal established above.

For the specific, but very relevant, case of generative AI, literacy has been seen as an essential skill, next to traditional digital literacy skills for learners in order to assist and personalise their learning (AAIN Generative AI working group, 2023). First, it requires understanding the different types of generative AI tools, that can create content like text and images, learning how to formulate effective prompts to get the desired outputs and using them to enhance learning and work. Secondly, the skill of evaluating the accuracy and trustworthiness of the results obtained includes identifying potential biases and fabricated information as well as necessitates verifying the information provided by AI with reliable resources. Third, it encompasses using AI ethically and responsibly, which means protecting sensitive data, recognising issues like data privacy and acknowledging the use of AI in academic work (Pretorius, 2023).

AI literacy is crucial for XAI for several reasons. As it relies on understanding AI-made decisions, it requires from the users the ability to grasp the "why" and "how" behind an AI system's behaviour. In addition to that, the ability to evaluate the trustworthiness of the information provided by AI can be greatly enhanced by the introduction about transparent processes that reveal potential biases or errors. Furthermore, XAI facilitates questioning ethical concerns towards AI by exposing biases, constraints, and limitations in decision-making, enabling users to critically assess the technology. Finally, XAI aligns with AI literacy by making AI more accessible to individuals without deep technical expertise, empowering a broader audience to responsibly engage with AI technologies.

4.3. Core competences and principles for integrating XAI in education

With the aim of providing a clear path towards the integration of XAI in education, this section will focus on the core competences for teachers in this scope, which will be built upon the *UNESCO's AI Competency Framework for students* (Fengchun & Kelly, 2024). This framework has been chosen as a reference because it includes competences to learn with AI and also to learn about AI (European Commission, 2023), while the teacher's framework is focused on the competences to teach with AI. Hence, from the student's framework, those competences related to XAI have been selected, and the ones required for educators have been defined according to them. The competences have been grouped based on their importance for different ISCED level educators. They progress across levels, starting with foundational, basic skills for lower ISCED levels and building up to more specific advanced competences for higher levels. It is assumed that *higher levels include the previously acquired ones, so no repetition is required.*

ISCED levels 1-3 (primary education, lower secondary education, upper secondary education)

The following competences bring together principles and skills that equip learners at these initial levels to understand, evaluate, and optimise AI systems, ensuring they operate transparently and responsibly—an essential aspect of XAI.

Ethics of AI: At the levels of understanding, application, and design, the importance of integrating ethical principles at every stage of the AI lifecycle, from conceptualisation to implementation, is emphasised. Transparency and explainability play a fundamental role in this process, empowering an informed public to actively participate in the regulation and ethical use of AI. In this context, XAI is essential, as it enables the breakdown and justification of how and why AI makes certain decisions. This ensures that these decisions align with ethical standards and prevent discriminatory practices or biases. Furthermore, an XAI facilitates public understanding of its processes, enabling citizens to adopt a critical perspective and make informed decisions regarding its adoption and responsible use.

Human-centred mindset at the application level: This competency involves awareness that people are responsible for decisions generated by AI, especially in high-impact contexts. XAI provides a basis for justifying these decisions and enables both designers and users to assume the associated legal and ethical responsibilities. This mindset is closely related to ISCED levels 1, 2 and 3, as it builds a basic awareness of human responsibility in AI, moving from an exploration of real-world applications of AI and getting to the importance of explainability in ensuring ethical and legal accountability.

Techniques and applications of AI at the levels of understanding and design: These competences encompass the ability to determine when and how AI should be appropriately used. XAI allows to evaluate whether the selected models and architectures are the ones most suitable for the problem at hand, offering a detailed view of the models' limitations and capabilities. Thanks to this approach, students at ISCED level 1 are introduced to basic AI applications to understand where and how AI is used in daily life. At ISCED level 2, they explore when AI should be applied and analyse its suitability for specific tasks. And at ISCED level 3, they evaluate and design AI systems using XAI to assess the appropriateness, limitations, and capabilities of different models and architectures.

AI system design at the creation level: XAI facilitates the continuous improvement of models by providing detailed information about their functioning, which enables the identification of optimisation areas, correction of potential errors, and minimisation of biases in each iteration. Framing this level through ISCED, students at ISCED level 1 gain foundational exposure to how AI operates through simple, creative activities. At ISCED level 2, they begin to explore iterative improvement and bias correction in AI models through guided experiments. And at ISCED level 3, they engage deeply with AI tools including explainability to critically analyse, optimise, and minimise biases in AI system design.

ISCED levels 3-5 (vocational education and training (VET))

VET is subject and occupation based and therefore uses AI in many ways. AI can be used as a teaching and learning tool, but it is also a subject on the curriculum, and these subjects need to change to foster explainability. Critical digital literacy is central to AI in VET. In this realm, the following teacher competences for XAI are required:

Collaboration with AI: In VET, AI literacy should be more than technical understanding—it may involve learning how to collaborate with AI systems, interpret their outputs, and use AI tools to enhance decision-making and problem-solving within a particular trade or profession. For example, medical assistants will have to work with and critically interpret AI-powered triage systems.

Hands-on skills: AI literacy should extend to the development of hands-on skills. For many vocational learners, AI literacy may include acquiring the technical skills to operate, maintain, and troubleshoot AI-enabled machinery or software, understanding their outputs and responses. Electricians for example may need to work with XAI-powered smart home systems.

Ethical use of AI tools in the industry: AI literacy in VET must include a strong emphasis on the ethical and safety considerations of using AI technologies. In industries where AI plays a critical role in decision-

making — such as healthcare, transportation, and manufacturing — workers must understand the ethical implications of AI, such as the potential for bias in AI algorithms, the impact of automation on job roles, and the importance of data privacy and security. In VET, this means preparing learners to critically assess the outputs and decisions made by AI systems, especially in contexts where human safety is at stake, such as in automated manufacturing or AI-driven medical diagnostics.

Lifelong learning: AI literacy in VET should be also aligned with the growing need for lifelong learning in vocational education. As AI technologies rapidly evolve, workers will need to continuously update their skills and knowledge to stay competitive. In vocational education, fostering AI literacy means encouraging learners to take ownership of their ongoing education, continually updating and improving their knowledge and skills in alignment with industry needs, trends and technological advancements. As XAI will play a key role in all AI tools in the future, it must be included in recycling programs for professionals too.

Consequently, VET programmes must incorporate AI literacy into their upskilling and reskilling initiatives, helping workers transition to new roles that involve overseeing AI systems, managing AI-integrated processes, or developing AI-powered solutions within their industries. This ensures that workers remain relevant and adaptable as AI technologies continue to evolve. For instance, the forward looking Erasmus+ project *AI Pioneers* (Bekiaridis & Atwell, 2023) has developed a proposed extension to the *European Framework for the Digital Competence of Educators: DigCompEdu* (Redecker & Punie, 2017) for vocational teachers and trainers in Europe, detailing the following competences for teaching and learning with AI. The levels are consistent with DigCompEdu.

ISCED levels 6-8 (higher education)

When addressing XAI competences in higher education, it is important to differentiate between those required by researchers and those needed by educators of non-technical and technical degree courses. In any case, it should be recommended to educators at this level to previously acquire the core competences of ISCED levels 1 to 3. The ones here are defined by the advanced, autonomous, and research-driven nature of teaching and learning in higher education. Furthermore, explainability in research becomes essential as it is important for researchers to be able to explain not only the way they use AI but also make transparent how AI works.

Competences for researchers using AI

Promoting good practices in open science and transparency in AI research: Researchers utilising AI, must demonstrate a strong commitment to open science principles and transparency. This includes adopting practices that ensure their research is reproducible, accessible, and ethically disseminated. Researchers should prioritise sharing models, datasets, and code under open licenses, enabling scrutiny, validation, and further development by the scientific community. They must also transparently document methodologies, assumptions, and limitations to foster trust and inclusivity, aligning with global frameworks like the UNESCO Recommendation on Open Science. Finally, in line with the Living guidelines on the responsible use of generative AI in research by the European Commission, researchers should adopt responsible and transparent approaches when developing and using generative AI tools.

Understanding XAI techniques: Knowing when to apply ante-hoc or post-hoc explainable methods according to the interpretability of the AI system and the research requirements is essential. When researchers understand the reasoning behind AI predictions, they can make more confident and well-founded decisions based on these predictions. Moreover, differentiating between the key dimension of XAI included on table 4 is also important. For instance, distinguishing between explanations that focus on specific predictions (local) and those that provide insights into overall model behaviour (global) allows researchers to detect potential biases in AI models, such as unequal treatment of demographic groups and ensure fairness.

Knowledge of human-centric design: Closely related to XAI competences is the knowledge of human centric design that researchers need to have when involving AI into their work. Transparency in research cannot be reached if the researcher does not have the ability to tailor explanations for different stakeholders ensuring the explanations are understandable and actionable. But this transparency can only be reached if the AI model is at its turn transparent. Researchers who can explain model behaviour to both technical and non-technical audiences make their work more accessible, fostering greater acceptance and understanding from a broader audience.

Awareness of societal and environmental impact of AI: This competence refers to the broader implication of developing and deploying AI systems by understanding how AI systems shape social norms, what ethical dilemmas they pose, and what environmental challenges they bring. Researchers who critically reflect on how AI might prioritise certain cultural and/or linguistic narratives while under-representing others adopt research practices that enhance reliability, equity, and trust in their work.

Competences for non-technical degree teachers

Critical literacy in XAI for informed decision-making in specific areas: Non-technical university educators should be able to critically assess and interpret how AI-based tools and applications generate outputs or recommendations relevant to their subject area. This involves being aware of the basic mechanisms that render AI systems explainable, the common sources of bias, and the limitations that might arise from opaque or proprietary AI models. Crucially, it also includes understanding how to communicate these concerns to students in a way that fosters responsible, evidence-based usage of AI across different academic disciplines, encouraging this critical attitude in their professional future.

Pedagogical integration of XAI principles: Beyond critically consuming AI outputs, non-technical educators can benefit from the ability to design or adapt learning activities that highlight the principles and implications of XAI for their specific domain. This might include creating assignments in which students reflect on AI-based decision-making in real-world scenarios, or guiding students to analyse and compare AI-driven results with human-driven reasoning. It also involves framing AI's explainability (or lack thereof) in broader discussions around ethics, equity, and the societal impact of algorithmic decision-making.

Competences for technical degree teachers

Comprehensive knowledge of XAI techniques and algorithms: This competence involves providing students with a thorough understanding of various XAI techniques and algorithms needed to ensure data, model, process, outcome, and purpose of explanations. Educators should help students gain in-depth knowledge of the principles, applications, and limitations of these techniques, enabling them to develop AI systems that are both effective and transparent.

Incorporating explainability in AI system design: This competence focuses on teaching students how to integrate explainability into the AI system design process from the very beginning. Educators should highlight the importance of designing AI models that are not only technically well-developed but also follow ethical principles and provide clear and understandable explanations for their decisions. This involves guiding students to select algorithms and design approaches that balance high performance with interpretability, ensuring that technical and non-technical users can understand the AI system's logic and outputs.

4.4. Competences for the key dimensions of XAI

Incorporating specific topics related to XAI into educational curricula is a new scenario in AI literacy that must be faced by the educational decision-makers from the EU's Member States. It would foster learners and educators alike to question how AI systems arrive at their decisions, encouraging a more participatory and reflective approach to technology use. Teaching with AI could further enhance learning by using XAI tools to personalise education, fostering better engagement, and enabling learners to interact with AI systems in real-time while understanding the reasoning behind the system's choices. Thus, XAI holds transformative potential in making AI understandable and actionable for all learners, regardless of their educational level, and it must be incorporated to literacies.

The key dimensions of XAI defined in table 4 establish four main types of explanations, which will be present, to some extent, in all AI systems in the near future. Every AI user should comprehend their differences and usefulness, so it is necessary to include them in AI literacy. Consequently, any educator should be aware of and understand these explainability dimensions to train their learners about them. Moreover, with such a knowledge, educators could promote those AI tools that cope more appropriately with XAI principles and correlate the key dimensions of XAI with the learning goals, their personal/institutional and their learners' needs.

The UNESCO's AI competency framework for teachers (Cukurova & Miao, 2024), establishes the AI competences required by educators to empower them to use AI-based technological tools in their teaching practices in a safe, effective and ethical manner. As the competences required for the key dimensions of XAI have the same goal, they must be obtained from the framework. They are displayed in the following table.

Key dimension	Recommended competences
Scope: global vs local	**AI foundations and applications:** Educators need to understand both global and local explanations to assess the appropriateness of AI tools and their implications for specific educational contexts. This includes evaluating how AI tools function across various scenarios and understanding specific instances of AI behaviour.
Depth: comprehensive vs selective	**AI pedagogy:** Educators should be able to interpret both comprehensive and selective explanations. This competency allows them to provide detailed evaluations for in-depth system reviews and simplified insights for immediate feedback, which is crucial for effective teaching and learning. **AI for professional development:** Understanding the level of selectiveness in explanations can also aid educators in their professional growth, as they learn to adapt their instructional strategies based on the insights provided by AI tools.

Alternatives: contrastive vs non-contrastive	**Human-centred mindset and ethics of AI:** Educators must be equipped to provide contrastive explanations, helping learners and stakeholders understand differences in outcomes. This involves ethical considerations and accountability in using AI outputs, ensuring that educators can defend their decisions and maintain educational integrity. **AI foundations and applications:** Understanding non-contrastive explanations is also essential, as educators need to grasp the significant factors influencing AI decisions without necessarily comparing outcomes.
Flow: conditional vs correlational	**AI pedagogy:** Educators should be able to convey explanations effectively, using conditional formats for clarity and correlational insights for broader understanding. This competency is vital for integrating AI tools into pedagogical strategies. **AI for professional development:** The ability to communicate explanations clearly and effectively is crucial for educators 'ongoing professional development, enabling them to share insights with colleagues and adapt to new AI technologies.

Table 13: Recommended competences for educators towards the key dimensions of XAI.

4.5. Practical implementations

As different competences are needed by educators teaching at different ISCED levels, the examples of possible XAI implementation in education can vary from one context to another. Starting with an introduction to AI literacy through unplugged activities at ISCED level 1, continuing with building advanced digital AI literacies at ISCED levels 2 to 4, or having a specialised focus on VET/higher education, the following pages provide a few examples of activities in which XAI plays a key role.

It must be pointed out that the final implementation of these activities depends on how and by whom in the Member States the curricula are designed, or whether their designing is part of ad hoc projects, as it was done for the introduction of programming and computational thinking.

ISCED level 1 (primary education, age 6-12)

Introducing AI literacy

At the primary level, AI literacy can begin with simple concepts such as how machines learn and make decisions. Interactive educational tools, like simple AI-based games or storytelling apps, could be used to demonstrate how AI reacts to different inputs (e.g., voice commands, facial recognition). For example, a classroom might use an AI-enabled virtual assistant to help students with reading comprehension. Teachers can use this opportunity to explain how the AI understands their words and reacts accordingly, providing a basic foundation in how AI processes information.

At ISCED level 1, the tendency is to try to minimise screen time. But even at this early age some XAI activities can be implemented. They could be unplugged activities, familiarising young students with computational

thinking through gamification and usage of everyday elements they are familiar with. These activities are developed in the school context rather than taking place in the student's daily life, at home. At this stage, XAI can be incorporated through basic, child-friendly tools that explain why the AI suggests specific outcomes.

A primary school teacher could introduce an AI-based spelling or math app that not only marks answers but also explains the logic behind correcting the mistakes. Voice explanations, colour coding, sound-based feedback could be used rather than text-based explanations. This would foster early critical thinking about AI, making students aware that machines follow specific rules and data, and encouraging them to question those processes.

The Data Science Fiction Scratch project, paired with an explanatory text (Tzampazi, n.d.), is designed to look into transparency where critical thinking, informatics and AI literacy meet. This project serves as both an interactive resource and a coding lesson, highlighting that biases can arise even without AI — let alone when AI is involved — because AI is fundamentally shaped by human input. Basic coding skills help demystify these processes, debunk myths, and enhance system transparency. While many user-friendly machine learning platforms and children's projects focus on input-output relationships and transparency in data collection, they often leave the inner workings of models unexplained. Unlike an AI algorithm, this project emphasises the algorithmic aspects of data processing and decision-making, offering a clear example and concise explanation of why AI literacy — an essential part of informatics literacy — is deeply interconnected with coding, data, and mathematical literacy, the latter forming a critical foundation. At its core, this ties back to critical literacy: Why, how, and who can manipulate data? This reinforces the need for transparency, making opaque models — whether black boxes or locked boxes — unacceptable..

The AI Unplugged initiative provides introductory AI activities that do not require a computer and are adequate for this level. For instance, in the Good-Monkey-Bad-Monkey game, students create a classification model using a decision tree. They analyse a set of paper cards with monkeys showing different expressions, and they have to assign them to two possible categories: good or bad. Then, they create the decision tree using the selected expressions as the feature for the branch creation. The goal is to test it with new cards, figuring out how their algorithm is not perfect, as the categories are subjective, so teachers can introduce the idea of accuracy or probability of success. In the scope of XAI, teachers could adapt this activity to include an explanation regarding the feature that was used to select one branch or the other.

ISCED level 2 (lower secondary education, age 12-15)

Building AI literacy

As students' progress, AI literacy can delve deeper into how AI is used. Middle school students can start learning about AI's applications in various industries like healthcare, transportation, and entertainment. Teachers can introduce discussions about how AI impacts their lives, such as in social media, recommendation systems, and smart devices.

Students can be challenged with a project in which they investigate an AI system (like a music recommendation algorithm) and explain how it works. Using tools like the UNESCO AI Ethics framework, they can also assess the potential biases in these systems.

This framework provides clear guidelines on the ethical development and application of AI, emphasising principles like fairness, transparency, and accountability. By using this tool, teachers can critically evaluate AI-based platforms (like grading tools or adaptive learning systems) to identify potential biases rooted in data, design, or decision-making processes. For instance, teachers can analyse how student performance prediction tools make decisions, ensuring that no student is unfairly disadvantaged due to factors like gender, ethnicity, or socio-economic background. The framework also encourages educators to advocate for more transparent AI models, pushing developers to provide explainable insights into how AI predictions are made. This process empowers teachers to act as responsible mediators between AI systems and students, ensuring ethical and equitable educational practices.

Students and teachers could use XAI in educational platforms that personalise learning. In order to be able to select the most appropriate platform for their context and learners, teachers would need to develop the competences mentioned above in this chapter. AI-based systems can provide feedback on assignments, explaining how the student's work compares to others and why it generates specific recommendations for improvement. This could encourage students to critically assess these systems and their accuracy.

A few ideas for implementation

To deepen XAI competency in secondary schools, educators can implement various hands-on activities that combine technical exploration with ethical analysis. For instance, in a mathematics or computer science class, students could build basic supervised learning models like linear regression or decision trees to grasp how algorithms "learn" from data inputs. By experimenting with different datasets — such as demographic or environmental data — they can observe how altering input variables affects model outcomes, making the influence of data selection transparent.

In social studies, teachers could guide students through case studies of AI applications in real-world scenarios, like predictive policing, where algorithms assess risk based on historical data. Students could analyse how biases in data (e.g., biased arrest records) can perpetuate social inequities, fostering discussions about the ethical dimensions of AI use. Additionally, students could use web-based platforms to create AI models that visually represent predictions, enabling them to test assumptions and visualise decision boundaries.

In the context of learning English as a second language (ESL), students could use XAI tools to understand and evaluate language models and grammar-checking applications. For instance, by experimenting with AI-powered grammar tools, students can see how the tool suggests changes based on specific language rules. Teachers might guide students to analyse why the AI makes certain recommendations, prompting students to reflect on syntax, word choice, and context.

Fostering a classroom culture that encourages questioning AI outputs is essential. Teachers could run debate sessions where students discuss the ethical implications of AI applications they've studied, considering perspectives like data privacy, transparency, and accountability. By blending technical exercises, real-world case studies, and ethical considerations, educators prepare students not only to understand AI mechanics but to engage critically and responsibly with AI systems they encounter in the future.

ISCED level 3 (upper secondary education, age 15-18)

Deepening AI literacy

At the high school level, students could delve into more advanced aspects of AI, including its ethical implications and societal impacts. These explorations should encourage critical thinking and foster discussions about AI's role in fields like law enforcement, hiring, or healthcare. Students can use examples from the AI Act, which emphasises transparency and fairness in AI systems, to understand how regulations aim to prevent harm and ensure accountability.

A valuable activity could involve students analysing real-world AI case studies where biases have caused unintended consequences. For instance, facial recognition technology. Students could study cases where facial recognition algorithms have exhibited biases, such as misidentifying individuals of certain ethnic groups or genders more frequently. This activity would highlight the importance of fairness and accountability in AI design.

Students could use open-source web-based platforms to create and train simple machine learning models. Through their adoption they could so demonstrate how changes in training data influence model accuracy and bias. For instance, they are particularly suited for educational purposes because they simplify complex AI concepts and provide immediate visual feedback.

With the aim of preparing students for tertiary education, more realistic situations could be faced by teachers, so XAI requirements could be clearer and more specific.

A few ideas for implementation

To help students understand AI bias, teachers can guide them through a simple project involving the creation, training, and testing of a machine learning model. Students learn that AI systems "learn" from training data, and the quality and balance of this data directly impact the model's accuracy and fairness. The proposed activity can be developed in four steps:

1. **Set up a model:** Students build a basic image classification model (e.g., identifying fruits or objects).
2. **Train the model:** Students upload and label images for different categories (e.g., apples, bananas, oranges).
3. **Experiment with data:** Students compare two training scenarios — one with balanced data and one with biased data (e.g., 100 apple images but only 10 for bananas and oranges).
4. **Test the model:** Students test the model with new images and observe how balanced datasets lead to fairer and more accurate predictions.

Another application of AI technology to foster critical thinking and teaching high schoolers the importance of AI transparency and accountability could be related to the development of a creative project combining literature analysis and AI technology.

For such a didactic proposal, students are invited to adopt an AI art generator: They would create visual representations of key themes, settings, or symbols from a chosen literary work, such as George Orwell's 1984. This activity encourages interdisciplinary learning, blending English literature with art and technology, while fostering critical thinking about AI's capabilities and limitations.

The project, as per the previous example related to open-source web-based platforms to create and train simple machine learning models is divided into several phases with precise objectives:

1. Preparation

- Literature analysis: Students identify 1984's key themes and symbols (e.g., Big Brother, the telescreen) and brainstorm descriptive prompts for the AI generator.
- AI introduction: Teachers explain how AI art tools create visual outputs from prompts and introduce concepts like training data and style emulation.

2. Execution

- Create AI art: Students input prompts (e.g., "dystopian city under surveillance") into the AI generator to visualise themes. They create multiple iterations to refine the outputs.
- Human enhancement: Students enhance AI-generated images using traditional or digital techniques, adjusting details, colours, and composition to better reflect the story's mood and key symbols.

3. Critical evaluation

- AI analysis: Students assess how well the AI captures 1984's themes, identifying any missing elements or overlooked nuances.
- AI vs human creativity: They discuss whether AI can achieve the emotional depth of human artists, reflecting on how metaphors, cultural context, and symbolic meaning are interpreted differently by humans and machines.

4. Teacher-led discussion

Teachers facilitate a discussion on AI's creative limitations as AI relies on training data, patterns, and algorithms, but lacks the ability to interpret deeper symbolic meaning or emotional layers. This discussion can be focused then on the human role in art, emphasising human insight's unique role in creativity and giving meaning, which machines cannot replicate.

In terms of learning outcomes, students will acquire further critical thinking skills as they will be able to evaluate AI's ability to understand and represent human concepts, being aware of how important interdisciplinary learning is fundamental to understand how AI-driven creativity works. Finally, this hands-on activity could help students reflect on the balance between human and machine contributions in the creative process.

ISCED levels 4-8 (post secondary non-tertiary education, short-cycle tertiary education, Bachelor's level or equivalent, Master's level or equivalent, doctoral level or equivalent)

In higher education, a differentiation must be carried out between students in technical degrees that will be AI developers or expert users, and those in non-technical degrees that will standard users.

Advanced AI literacy

Technical degree students will be exposed to the intricacies of AI development and implementation. Courses could focus on building, evaluating, and ethically deploying AI systems. Students could analyse the risks and benefits of AI systems in fields such as medicine, law, or business. They could also study how to make AI explainable and accountable, a critical skill as they prepare for professional roles.

Higher education students in computer science or ethics classes could be tasked with creating their own AI models and using XAI tools to explain the decisions made by their systems. This would give them hands-on experience in both building AI and ensuring its transparency for users. They can implement and develop further the educational bridge between higher education and lower school levels through the creation of XAI tools, such as chatbots, for primary and secondary school levels by higher education students. This would even provide them with the unique opportunity to create an even tighter and more productive knowledge and learning loop throughout the entire school system, with the consequence of better solutions for education thanks to an increasingly adopted XAI approach to AI.

Higher education students could use XAI systems to aid their research or assignments, but with a focus on understanding the implications of relying on AI. For instance, if they use AI to analyse large data sets or assist in writing reports, XAI techniques should help them track how the AI reached certain conclusions. This makes them more aware of potential biases or inaccuracies.

A few ideas for implementation

In research-heavy fields like sociology or data science, students can use SHAP or LIME to break down how AI models analyse and interpret social data. This would allow them to critique the AI's assumptions and understand whether the conclusions drawn are valid or require further scrutiny.

Deepening AI literacy

For students in non-technical degrees, their literacy requirements are similar to those of ISCED level 3, but more specific. As these students will be professionals in specific areas, they need competences for properly using AI tools in their scope, understanding the technical features, and maintaining a critical view of the responses provided.

A few ideas for implementation

University students in medicine should learn how to use an image diagnosis tool based on AI, as they will use it in their future occupations. But before that specific training, an activity to practice developing could be training an artificial neural network using a simple application for image classification. The students would collect medical images from the internet or specific databases and try to train a model to predict the probability of suffering some kind of medical issue or disease. Students will realise the difficulty to get to a high level of accuracy, but they must try to reach the highest and include an explanation of their final success rate. This way, students will understand that the level of trustworthiness in sensitive fields such as medicine is hard to obtain. They must be aware of this when selecting and using similar tools in their professional life.

In teacher training programmes, learners should not only develop AI competences and learn to deploy AI as a system or tool to support teaching and learning but also develop their understanding of the role of XAI in education. This involves fostering critical thinking skills to evaluate and use AI tools effectively for educational content generation and raising their awareness of the potentially biased AI-generated outputs. Such competences might be explicitly relevant to future foreign language educators, who need to be aware of cultural and linguistic biases in AI-generated outputs.

Higher education students enrolled in teacher training programmes for foreign language teaching could use AI tools to generate three key outputs for course planning and teaching: a course syllabus, a lesson plan, and a learning activity. For this task, students could use AI tools to generate course curricula, assessments, assignments, quiz questions, etc. After completing each output, they upload their work into AI tools to receive feedback and suggestions for improvement. The educator encourages learners to test and refine their prompts to enhance the AI-generated feedback. The educator then facilitates discussions about AI risks in content generation, emphasising the importance of understanding how AI outputs might reflect specific cultural or linguistic contexts (e.g., a variety of Spanish dialects). Throughout the process, students are encouraged to explore and critically assess the cultural and linguistic orientation of the AI tools used and discuss how these orientations may affect the suitability of AI-generated outputs for diverse learner groups.

At the end of the activity, students could write a reflection based on open questions such as: How did you succeed in completing the tasks? Which prompts led to the best outputs and why do you feel they were the best? How did the tool improve your syllabus or lesson plan and in what way? Which AI-suggested activities did you adopt or reject and why? What suggestions were helpful or unhelpful and why do you think so? Did you consider the cultural and linguistic context of the AI tool? If so, how do you think this influenced the outputs, particularly in the context of foreign language teaching? This use case refers to XAI competences of understanding cultural and linguistic implications (knowledge), contextualising/ adapting AI tools to align with learning objectives and cultural or linguistic needs of students (skills). I also emphasises the ethical responsibility of future educators to question AI systems and demand better transparency, inclusivity, and fairness from developers as well as critically reflect on how AI systems might prioritise certain cultural and/or linguistic narratives while under-representing others and seek to foster a more balanced and inclusive approach in teaching practices (values).

4.6. Recommendations for different stakeholders

The following table summarises the main actions that could be taken by different educational stakeholders to properly integrate XAI into education. They can be complemented with those extracted from the 2024 EDEH XAI community workshop policy recommendations to get to a formal roadmap supporting the integration of XAI.

Stakeholders	Key actions	Description
Educators	**Develop a general understanding of how AI systems work and adopt a critical approach.**	Be alert to AI-generated content that is inaccurate or biased and try to understand its outputs.
	Ensure the constructive alignment of educational goals with AI tools.	AI tools should support learning outcomes and align with teaching and assessment strategies.
	Work with AI.	Combine AI strengths with your judgment and experience to create a balance.
	Participate in professional development to enhance AI literacy skills.	Keep abreast with developments in AI and its application in teaching practices.
Educational leaders	**Select AI tools that follow XAI principles.**	Ensure AI applications are transparent, accountable, and align with institutional goals.
	Provide training opportunities for educators and staff.	Support educators in guiding learners on how AI functions in general.
	Prioritise on adopting AI solutions that support educators and align with pedagogical goals.	Select tools that have proven to add value in education and that have been validated by experts.
	Encourage clear communication about AI decision-making processes.	Ensure stakeholders understand the rationale behind AI-driven recommendations.
	Adopt human-centric design principles when integrating AI into your institutions.	Ensure that AI systems are used in ways that prioritise well-being, needs and educational goals of learners and educators.
Policy-makers	**Define a set of competences at the European level.**	Develop a European AI Literacy Framework. Standardise the inclusion of XAI in curricula across Member States and promote extensive educator training to ensure inclusive AI education.
	Provide access to processing power for educational institutions.	Establish regional AI hubs equipped with shared computational infrastructure. Foster partnerships with tech companies for subsidised AI access and fund the creation of user-friendly platforms with integrated XAI tools.
	Promote the use of open educational resources (OERs).	Invest in creating and translating OERs and offer training for educators on using and creating OERs. Ensure these resources are multilingual and culturally inclusive to meet the needs of diverse learners.
	Increase funding for AI in education.	Launch dedicated funds to support AI infrastructure, training and curriculum development. Provide grants for innovative AI projects and fund research into XAI technologies.

Table 14: Main actions for different stakeholders to integrate XAI into education.

4.7. Summary and final considerations

Once AI literacy is properly implemented, learners and educators will become more adept at critically interacting with AI systems, fostering XAI. In daily educational life:

- Learners will learn to view AI as a tool that requires questioning. Whether they are using AI for research, projects, or learning, they will approach it with a mindset looking for transparency and fairness.
- Educators will incorporate AI tools into their teaching practices but with an awareness of how to explain these tools' inner workings to learners. They will facilitate discussions on the ethical implications of AI, using XAI to show how AI makes decisions and why this is important to understand in the context of education — and beyond.

Ultimately, integrating AI literacy at all levels helps foster a generation that is not only skilled in using AI but is also capable of critically assessing its role in society, making informed decisions about its ethical implications, and contributing to its responsible development.

5. Conclusion

XAI in education goes beyond solely being a technical enhancement – it is a foundational requirement for fostering human agency, trust, and transparency in educational environments that increasingly rely on AI tools for teaching and learning. As this report has demonstrated, XAI plays a critical role in aligning AI systems with educational values, legal requirements, and pedagogical goals.

The report has explored the evolving legal landscape, the implications of the AI Act and GDPR, and how these regulations intersect with real-world educational practices. Practical scenarios have highlighted the complexity of ensuring compliance while maintaining usability for diverse educational stakeholders. From educators and learners to developers and policy-makers, each stakeholder has distinct responsibilities and expectations, which must be addressed through tailored and meaningful AI explanations.

Moreover, the integration of XAI into educational contexts calls for a rethinking of digital competences, in particular AI literacy. Educators must be equipped not only with the skills to use AI tools, but also with the critical thinking needed to interpret their outputs and ensure ethical deployment. Learners, in turn, require support in developing their understanding of AI decisions, fostering agency and self-regulated learning.

Ultimately, XAI is not an endpoint but a shared process of co-creation. Building transparent, fair, and human-centred AI for education means creating systems that are not only technically sound but also contextually appropriate, legally compliant, and pedagogically effective. This calls for continued collaboration across disciplines, sectors and stakeholder groups. As AI continues to evolve, so must the collective efforts to ensure its use in education supports, not replaces, human judgment, values, and oversight.



© European Union, 2025 – all rights reserved.