ASIA PACIFIC JOURNAL
OF EDUCATORS AND EDUCATION

Volume 41, Issue 1, 2026
https://doi.org/10.21315/apjee2026.41.1.19

Research Article:

XAI as Compliance and Design Strategy for Automated Grading Systems

Hasan Abu-Rasheed[1] and Lezel Roddeck[2]*

[1]Goethe University Frankfurt, Computer Science and Mathematics, Theodor-W.-Adorno-Platz, 60629 Frankfurt am Main, Germany
[2]Bucerius Law School, Foreign Language Communication, Jungiusstrasse 6, 20355, Hamburg, Germany

*Corresponding author: lezel.roddeck@law-school.de

ABSTRACT

Ensuring explainability in Artificial Intelligence (AI) systems requires the careful integration of legal compliance and technical design. This paper proposes a compliance and design strategy for operationalising explainability obligations under the European Union's Artificial Intelligence Act (AI Act), and the General Data Protection Regulation (GDPR), using automated grading systems (AGSs) in education as a case study. We argue that legal requirements and technical approaches should be coordinated from the earliest stages of system development to ensure trustworthy AI deployment. Our analysis combines legal analysis on the provisions of the AI Act's relating to "high-risk" educational AI and the GDPR's provisions on solely automated decision-making with a technical assessment of explainability methods including SHAP, LIME, feature sensitivity analysis, and interpretable models, and evaluate their appropriateness for educational contexts. An illustrative analysis of automated grading systems (AGSs) as a representative educational AI use case highlights how risks can emerge in automated grading and how targeted explainability approaches may mitigate these risks. Legal analysis reveals a duty to provide meaningful, fair and transparent explanations. A compliance and design strategy is proposed that integrates human-in-the-loop (HITL) mechanisms, collaborative design processes with educators, and stakeholder-centric explanation techniques. This interdisciplinary strategy demonstrates how technical methods and legal standards can work together to reduce risks, improve accountability, and strengthen trust in AGSs.

Keywords: AI act, GDPR, Explainable AI (XAI), educational stakeholders, automated grading systems

Published: 9 June 2026

To cite this article: Abu-Rasheed, H., & Roddeck, L. (2026). XAI as compliance and design strategy for automated grading systems. *Asia Pacific Journal of Educators and Education, 41*(1), 449–476. https://doi.org/10.21315/apjee2026.41.1.19

INTRODUCTION

Artificial intelligence (AI) is having an increasingly significant impact on educational environments. One of the most high-risk applications is automated grading systems (AGSs), which uses machine learning and natural language processing (NLP) to evaluate learners' performance in exams or admissions processes. While AGSs offer efficiency and scalability, their opacity raises significant concerns regarding transparency, fairness and accountability. Explainability or XAI, requires AI-generated outputs to be understandable, traceable, and challengeable by both users and affected individuals. It is becoming a crucial requirement in educational AI, as it supports efforts to operationalise legal compliance and enables stakeholders to receive contextually relevant and legally sound explanations for AI-outputs.

The European Union's Artificial Intelligence Act (AI Act) and General Data Protection Regulation (GDPR) establish obligations that embed principles of explainability, transparency, and human oversight to safeguard fundamental rights. However, operationalising these requirements in educational contexts remains challenging. Institutions must navigate overlapping legal duties, technical constraints, and pedagogical considerations while maintaining trust and accountability. Consequently, this paper aims to bridge the gap between legal requirements and real-world technical implementation by developing an integrated compliance and design strategy for AGSs linking XAI technical approaches and legal obligations under the AI Act and GDPR to pedagogical practice.

First, the article outlines the research methodology and key research questions. It then sets out the technical foundations of XAI, defining the core concepts, methods and stakeholder requirements involved. Subsequently, the paper examines the explainability requirements under the AI Act and the relevant GDPR provisions, as well as the growing importance of AI literacy in education. Finally, it presents a case-based analysis of AGSs, demonstrating how risks arise and how targeted explainability approaches can mitigate them.

To address these challenges, a compliance and design strategy is proposed. This strategy should integrate human-in-the-loop (HITL) mechanisms, collaborative design processes with educators, and stakeholder-centric explanation techniques. It advocates for an integrated approach that aligns legal obligations with technical XAI approaches from the earliest stages of AI development.

RESEARCH METHODOLOGY

This research adopts an interdisciplinary, conceptual approach, combining doctrinal legal analysis with a technical analysis of XAI methods and their implementation in AI-supported grading systems (AGSs) through a scenario-based, literature-informed framework. The technical component examines XAI methods in terms of their

interpretability, stability, and suitability for different stakeholder contexts, as well as their integration into system design. The legal component analyses explainability-related obligations under the AI Act and the GDPR through textual interpretation of primary legal sources, supported by policy documents and secondary academic literature (McConville & Chui, 2017). While the analysis is limited to EU law and a single educational use case, the framework is intended to be transferable to other jurisdictions that adopt similar ethical principles. In particular, the technical and design-oriented insights can be applied across different regulatory contexts.

To investigate the risks and implications of AI-supported systems in education from both technical and legal perspectives, the paper adopts a theoretical case study based on documented functionalities of AGSs described in academic literature, technical reports, and publicly available system documentation.

The research is guided by three questions:

1. Which technical methods of explainable AI are suitable for educational contexts, and how can they support legal compliance while addressing stakeholder needs?
2. What explainability requirements are imposed by the AI Act and GDPR on educational high-risk AI systems, and how do they apply to AGSs?
3. What strategies and design approaches are needed to embed explainability across the AI system lifecycle?

TECHNICAL FOUNDATIONS OF EXPLAINABILITY IN AI SYSTEMS

XAI has emerged as a key approach to addressing the "black box" nature of many modern AI systems (Fiok et al., 2022). Table 1 outlines the key technical definitions that are referenced throughout this article.

Table 1. Conceptual distinctions between explainability, interpretability, transparency, and understandability in AI

Concept	Definition (with source)	Key distinction
Explainability	The ability of AI systems to provide clear, human-understandable explanations for specific decisions or predictions (Arrieta et al., 2020).	Active: focuses on explaining outputs.
Interpretability	The degree to which a model's internal logic can be understood by humans, often requiring technical knowledge (Russell & Norvig, 2022, p. 728).	Passive: depends on human ability to understand the model.

(Continued on next page)

Table 1. *(Continued)*

Concept	Definition (with source)	Key distinction
Transparency	The availability of information about AI systems, including how and why decisions are made; legally requires systems to be understandable to users (Bellas et al., 2025; Art. 13, AI Act).	System-level: focuses on openness and disclosure.
Understandability	The extent to which explanations are meaningful and aligned with the user's knowledge and context (Saeed & Omlin, 2023).	Human-centric: focuses on user comprehension.

Ante-hoc vs. Post-hoc Methods: A Fundamental Trade-off

The first key dimension in XAI design concerns whether interpretability is built into the model from the start (ante-hoc) or applied retrospectively to an opaque model (post-hoc). This is not merely a technical preference: in high-risk AI system, it directly affects the trustworthiness of explanations. Ante-hoc models provide inherently faithful explanations, whereas post-hoc methods rely on approximations that may reduce reliability.

Ante-hoc methods produce inherently interpretable models. Examples include decision trees (traceable rules), rule-based systems (explicit logic), and generalised additive models (GAMs), which decompose predictions into interpretable feature contributions (Caruana et al., 2015; Wood, 2017). Their main advantage is that the model itself is transparent. This is an important consideration where explanation faithfulness is legally required.

However, ante-hoc interpretability comes at a cost. In tasks such as automated essay grading, which require capturing complex semantic features like argumentation and coherence, simpler interpretable models tend to underperform compared to more complex approaches, such as transformer-based architectures (Dimari et al., 2024). Prioritising interpretability may therefore reduce grading accuracy, raising concerns about fairness and validity.

Post-hoc methods address this trade-off by generating explanations for high-performing opaque models. However, these explanations are approximations rather than a direct account of how the model works. Common approaches include SHAP, LIME, and counterfactual explanations. SHAP assigns feature importance scores, but it assumes features are independent. In educational contexts, where elements such as argument quality and evidence are closely related, this can lead to misleading results. LIME provides simple, local explanations for individual predictions, however, since it relies on random sampling, results may vary across runs (Burger et al., 2023). This instability can be problematic when decisions are challenged. Counterfactual explanations show how outcomes could change (Wachter et al., 2017), but they may encourage superficial improvements, rather than improving underlying reasoning.

The ante-hoc vs. post-hoc choice therefore reflects a core trade-off: the most accurate grading models are the least interpretable, while post-hoc explanation methods introduce their own fidelity gaps. Rather than resolving this solely at the design stage, this paper frames it as an ongoing governance challenge which should be managed through explicit documentation, explanation auditing, and human oversight mechanisms throughout the system's operational lifetime.

Global vs. Local Explanations: Different Questions and Accountability Functions

Alongside the ante-hoc/post-hoc distinction, XAI methods differ in scope: global explanations describe the overall model behaviour across all predictions, while local explanations address its behaviour in a specific instance. This distinction matters because they serve different accountability functions.

Global explanations characterise model behaviour across the full dataset, clarifying which features drive predictions overall, whether the model treats demographic groups consistently, or whether grading logic has drifted over time. They support systemic oversight. For example, a global SHAP summary plot may reveal that a grading model weights surface features such as sentence length more heavily than substantive reasoning quality, an insight essential for the impact assessment of the model.

Local explanations, by contrast, address individual predictions, answering questions such as "why did this learner's essay receive this grade?". They therefore support individual rights, including the right to contest (Article 86, AI Act) and understand decisions (Article 13, AI Act; Article 22, GDPR).

A critical implication is that a system that is globally fair in aggregate can still produce locally biased decisions for individual learners, meaning that while the system as a whole may be fair, certain learners may still be disadvantaged. Therefore, relying on only one type of explanation thus represents partial compliance, both legally and ethically. The tiered framework introduced in the case study section therefore assigns different XAI methods to different stakeholder levels, since global and local explanations serve distinct legal and epistemic functions.

Model-agnostic vs. Model-specific Methods: Flexibility and Its Limits

A third-dimension concerns whether an explanation method can be applied to any model (model-agnostic) or is tied to a specific architecture (model-specific). Model-agnostic methods, including SHAP and LIME, treat the model as a black box and analyse its input-output behaviour without requiring access to its internal logic, making them applicable across different system types and vendors. Model-specific methods exploit a particular model's internal architecture, often yielding richer and more faithful insights. In AGSs, both approaches play complementary roles. Model-specific methods are better suited to

developer-level validation and bias auditing, whereas model-agnostic methods are more suitable for generating the accessible, stakeholder-facing explanations for learners and educators. Therefore, they are not interchangeable; selecting the wrong approach could result in either unintelligible feedback or insufficiently robust oversight.

The XAI methods outlined above address diverse stakeholder needs and evolving legal requirements on explainability. Rather than offering a single solution, they form an interconnected decision matrix. This paper therefore suggests a layered, stakeholder-specific approach: learners require actionable local explanations, educators need combined local and global insights for oversight, and administrators depend on system-level analytics, audit trails, and compliance evidence. The following sections examine the legal implications before illustrating how these requirements can be operationalised.

LEGAL FOUNDATIONS OF EXPLAINABILITY IN EDUCATIONAL AI

The AI Act complements existing instruments, such as the GDPR, and forms part of a broader digital regulatory framework aimed at aligning technological innovation with the protection of fundamental rights (Roddeck et al., 2025).

Explainability under the AI Act

To be explainable, high-risk AI system should provide users and affected individuals with accessible, meaningful and actionable information (Articles 13, AI Act), while also ensuring sufficient understandability to enable effective human oversight and intervention (Article 14, AI Act). Furthermore, affected individuals have the right to obtain explanations that enable them to challenge decisions and seek redress (Article 86, AI Act). However, the terms "transparency" and "explainability" are not specified as to their required level or form. It is unclear whether compliance is achieved through technical interpretability alone (Article 13, AI Act) or if it also requires meaningful, user-oriented explanations to support contestability and understanding (Article 86, AI Act). This ambiguity may lead to uneven implementation, with institutions either adopting minimal compliance or disclosing excessive information without improving understanding (Wachter et al., 2017; Edwards & Veale, 2017).

Furthermore, explanations do not necessarily lead to genuine understanding or effective oversight. Behavioural research suggests that users may experience an "illusion of explanatory depth", whereby they perceive themselves to understand a system more fully when provided with a coherent explanation, even if their actual comprehension remains limited (Rozenblit & Keil, 2002). For example, heatmaps in AGSs may highlight influential text segments, yet reflect statistical correlations rather than pedagogical reasoning.

Explainability under GDPR

The GDPR complements the AI Act by safeguarding individuals' personal data but adopts a more limited and procedurally oriented approach to explainability. Specifically, it restricts decisions based solely on automated processing, where such decisions produce legal or similarly significant effects, unless specific exceptions apply (e.g. consent) (Article 22, GDPR). Even in such cases, appropriate safeguards must be in place, including the right to obtain "meaningful information about the logic involved" in automated decision-making (Article 12, GDPR). Such information should be communicated in a concise, intelligible, and accessible form (Articles 12 to 15, GDPR).

However, the narrow focus on decisions taken "solely" by automated means limits the provision's scope, as most systems in practice involve some degree of human input. For example, where an AGS generates a recommended score but the final grade is formally confirmed by an educator, the processing will generally fall outside the provision's scope. This places considerable reliance on human-decision-makers despite risks of automation bias and limited capacity to critically assess outputs. Where Article 22, GDPR does apply, paragraph (3) requires safeguards, including the right to obtain human intervention, to express one's point of view, and to contest the decision. As a result, the GDPR framework may offer only partial protection in practice.

Impact Assessments under AI Act and GDPR

Deployers of high-risk systems are obliged to conduct a Fundamental Rights Impact Assessment (FRIA) prior to deployment (Article 27, AI Act). This structured evaluation ensures that any potential risks, such as discrimination, data misuse or privacy violations, are identified and mitigated before deployment. In parallel, Article 35, GDPR requires a Data Protection Impact Assessments (DPIAs) where processing of personal information is likely to pose high risks to individuals' rights and freedoms. While the AI Act positions DPIAs as complementary to FRIAs (Article 27[4]), this overlap risks creating procedural duplication without necessarily enhancing substantive protection (Fülöp & Poindl, 2025).

Furthermore, this may create administrative burden, particularly for educational institutions with limited resources, without necessarily improving the level of protection. This underscores the need for sector-specific frameworks tailored to educational contexts, as well as sufficient institutional capacity to ensure that impact assessments function as effective risk mitigation tools rather than merely procedural formalities.

Accountability under the AI Act

The AI Act distinguishes between providers (e.g., edtech developers) and deployers (e.g., educational institutions), assigning responsibilities across the AI lifecycle (Article 3(3), AI Act). Providers are responsible for model design, documentation, and compliance (Articles 16–28), while deployers must ensure appropriate use, monitoring and oversight.

Article 6(3), AI Act provides that certain AI systems listed as high-risk (Annex III) may be excluded from that classification where providers can demonstrate through self-assessment that the system does not pose a significant risk to fundamental rights or materially influence decision-making. While self-assessment reduces the regulatory burden, it raises particular concerns in education, where AI tools may influence grading, admissions, and learner profiling, often involving vulnerable populations. Allowing providers to classify systems as low-risk without external scrutiny could create conflicts of interest and undermine the safeguards designed to protect learners' rights (Wachter, 2024).

Integrating explainability throughout the AI lifecycle can mitigate the risk of misclassification by enabling transparent justification of system categorisation and identifying potential biases (Mittelstadt et al., 2019; Bellas et al., 2025). This approach also supports educational providers, particularly those with limited technical expertise, in assessing alignment with legal and pedagogical standards (Information Commissioner's Office & The Alan Turing Institute, 2020).

Furthermore, explainability enables learners and educators to contest decisions while embedding documentation, auditability and governance within institutional processes (Hacker et al., 2020; Mittelstadt et al., 2019). However, as noted by Edwards and Veale (2017), individual explanations alone are insufficient without broader accountability mechanisms, and as Selbst (2021) argues, such mechanisms must be designed in a way that is operationally workable. Together, these perspectives position explainability as a safeguard for individuals and a structural mechanism for institutional accountability.

Against this backdrop, fostering explainability across technical systems and institutional practices emerges as a prerequisite for trustworthy, human-centric innovation in education (Bellas et al., 2025). However, its effectiveness depends on complementary factors, particularly AI literacy. These factors are necessary to ensure that explanations can be interpreted and acted on meaningfully.

AI Literacy: Enabling Meaningful Explainability and Human Oversight in Education

While explainability is key to enabling AI systems to produce outputs that can be interpreted, its effectiveness depends on users' ability to understand and act on them. This highlights the importance of AI literacy, which is recognised in Article 4, AI Act as forming the basis for meaningful human oversight and responsible use. In educational contexts where AI

informs decisions relating to assessment, learning and institutional processes, it is crucial that explanations are not only provided, but also understood. Promoting AI literacy fosters critical thinking and personal agency, empowering stakeholders to engage with the ethical and pedagogical implications of AI (UNESCO, 2021). Explainability and AI literacy are therefore mutually dependent: without AI literacy, explanations lack practical value, and without explainability, oversight is undermined. This interplay is particularly significant in high-risk use cases, such as AI-supported grading systems.

USE CASE: THE CASE FOR AUTOMATED GRADING IN EDUCATION

To illustrate these dynamics analytically, this paper uses AGSs as a representative educational AI hypothetical scenario through which legal, technical, and governance issues may be examined.

Definition and Relevance

Technically, modern AGSs rely on transformer-based NLP models (e.g. BERT, RoBERTa, GPT-4) to capture semantic nuance beyond keyword matching (Dimari et al., 2024). While these architectures improve accuracy, they pose explainability challenges under Article 13, AI Act. Scoring pipelines align model outputs to rubrics (Tseng et al., 2024), raising questions of fairness and non-discrimination when applied inconsistently. Quality assurance mechanisms, such as HITL reviews, directly relate to Article 14, AI Act and Article 22, GDPR safeguards by ensuring that educators retain ultimate authority over grading decisions.

From an institutional perspective, implementing AGSs requires more than technical infrastructure. While software and computing resources are prerequisites, the real challenge lies in governance, training, and pedagogical alignment. Institutions must build capacity through staff training and change management, as confidence in AI tools remains limited (Onatere-Ubrurhe & Ubrurhe, 2025). Policies must address privacy, academic integrity, appeal rights, and oversight responsibilities (Figueras et al., 2024).

Since AGSs touch every layer of the educational ecosystem, from learners and teachers to administrators and regulators, this section explores the positive and negative impact they pose for different stakeholders.

Table 2. Positive and negative impact of AGSs on different educational stakeholders

Stakeholder	Positive impact	Negative impact
Learners	Rapid feedback supports learning adjustments and progress tracking. Objective grading can reduce bias and improve fairness. Data analytics highlight learning patterns and competencies (Fiok et al., 2022).	AGSs may fail to capture creativity or nuance. Opaque criteria can lead to percieved unfairness. Limited personalised feedback reduces formative learning opportunities (Saeed & Omlin, 2023).
Educators	Reduces grading workload and supports consistency. Provides insights into class performance for targeted teaching (Arrieta et al., 2020).	May reduce professional autonomy, especially with "black box" systems. Misalignment between pedagogy and algorithms can undermine assessment quality. Limited ability to assess higher-order skills.
Administrators	Enhances scalability, efficiency, and standardisation. Supports quality assurance, reporting, and compliance through aggregated data.	Risks of over-reliance, accountability gaps, and reputational harm. Lack of transparency complicates dispute resolution and may conceal bias at scale (Bellas et al., 2025).
Developers	Growing demand drives innovation and standard-setting. Real-world use improves system design and robustness.	Challenges in aligning technical, pedagogical, and legal requirements (Arrieta et al., 2020). Risks of non-compliance, bias, and limited explainability. Resource constraints for XAI integration.
Regulators and policy makers	Supports monitoring, standardisation, and evidence-based policymaking through structured data.	Opacity limits oversight. Rapid innovation creates regulatory gaps. Bias risks and inconsistent regulation may undermine fairness and trust.

Addressing Trust and Accountability Gaps

The challenges and stakeholder concerns outlined above highlight the need for explainable AI in educational assessment systems. Key risks, such as limited transparency, potential model bias, and limited stakeholder understanding, may be mitigated by robust explainability frameworks that meet the needs of different stakeholders (Bellas et al., 2025). Improving transparency in grading processes can also increase acceptance and reduce resistance. Explainability also enables meaningful human oversight, allowing educators to validate system decisions and intervene when necessary, thereby preserving professional autonomy.

Explainability further supports accountability and fairness. Transparent AGSs make it easier to detect and address bias, promoting equitable outcomes across diverse learner groups. Moreover, explainability mechanisms provide a basis for auditing and regulatory compliance, aligning technological innovation with institutional responsibility and the protection of learners' fundamental rights.

However, as discussed above users are more likely to trust AI systems that provide plausible or visually appealing explanations, even where those explanations are incomplete or misleading. For example, a grading system that highlights keywords or presents a confidence score may appear transparent, while its underlying assessment logic remains opaque. Such features may inadvertently amplify over-trust, reducing users' willingness to critically question system outputs (Miller, 2023). This raises concerns about the relationship between explainability and educational validity. Even technically accurate explanations may not align with pedagogically sound assessment criteria. In AGSs, explanations that emphasise surface-level features, such as structure or keyword frequency, may obscure deficiencies in reasoning and critical engagement, thereby reinforcing a narrow conception of assessment quality. Furthermore, explainability does not eliminate institutional power imbalances: learners remain dependent on institutional systems to interpret explanations, which could render the practical effectiveness of contestation rights ineffective.

Stakeholder-Specific Explainability Requirements

The XAI dimensions outlined earlier do not converge on a single optimal approach. Accordingly, an XAI design space is mapped onto three stakeholder groups, distinguished by their expertise (AI literacy), their decision-making authority, and the legal rights they hold in relation to AGS outputs.

Learners (Tier 3 - AI and pedagogy novices)

Explanation requirements: Learners require simple, actionable explanations that support learning and self-reflection. These include local explanations clarifying why a specific assignment received a particular grade, contrastive explanations indicating how outcomes

could be improved, plain-language summaries free of technical jargon, and feedback that explicitly links results to defined learning objectives.

Example explanation:

> Your essay received a B grade (85/100) due to strong argument structure (35/40 points) but weaker evidence support (10/20 points). To improve, include more specific examples and citations, particularly in paragraphs 3 and 5. Adding two relevant references to your central claim would increase your evidence score. This suggestion points to one concrete step, but your teacher provides further guidance on the overall evaluation.

In addition to text-based feedback, explanations may include visual elements. For example, a LIME visualisation could present a simple bar chart displaying the primary features that impacted a learner's grade (e.g. argument coherence or evidence relevance) employing user-friendly, non-technical labels. These approaches are effective for learner-facing feedback but require careful design. LIME outputs may vary across runs, which can lead to inconsistent explanations. This means that a learner who requests a re-explanation may receive a slightly different one for the same grade.

Figure 1 shows how a LIME explanation can be presented to learners in the form of a bar chart, indicating which features had the greatest influence on their individual grade. The chart uses simple, rubric-aligned labels such as "argument coherence" and "evidence quality" rather than statistical notation.

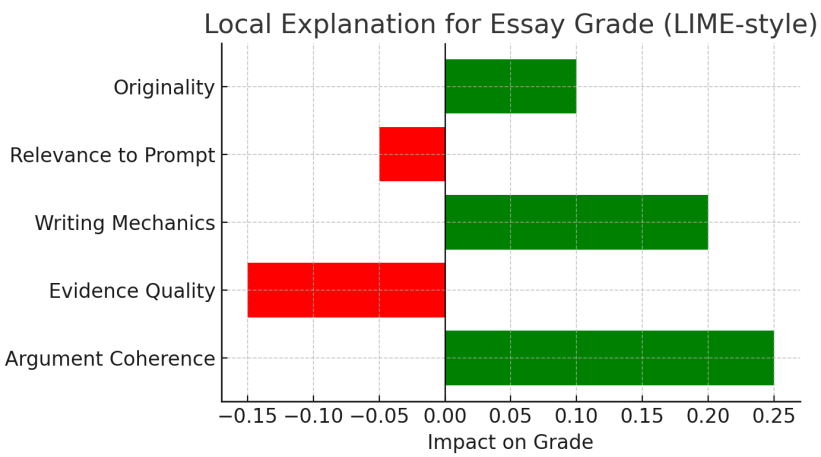

Figure 1. An example of LIME visualisation for the main features considered to grade an individual essay

While such explanations may enhance usability, they also risk promoting strategic behaviour, where learners optimise for model-recognised features rather than developing deeper analytical or critical skills. This raises concerns about the long-term pedagogical impact of explainable AI systems.

Educators (Tier 2 - AI users with pedagogy knowledge)

Explanation requirements: Teachers require explanations that support pedagogical decision-making and enable oversight. These include global explanations of cohort-level patterns alongside local explanations for individual cases, model confidence indicators to flag uncertain outputs, pedagogically aligned rationales linking grades to learning outcomes, and alerts identifying potential bias.

Example explanation:

> Grade summary: argument coherence (weighted 40%, 8.5/10), evidence quality (weighted 30%, 7.2/10), writing mechanics (weighted 30%, 9.1/10). Model confidence in this grade: 92%. No individual review flag raised for this submission. Class-level summary: across this cohort, language and expression accounts for 47% of average grade variance this semester, compared to its 30% rubric weighting.

This suggests the model may be placing more weight on how learners write than on what they are arguing, which does not reflect the intended assessment criteria. A review of borderline grades in this class would be recommended.

Explanations may include complementary tools. For example, SHAP summary plot can show the average importance of features across a cohort, helping educators assess whether the model's weighting criteria align with the intended rubric. These tools serve different functions: global views reveal systemic patterns, while local explanations assess individual cases. However, their effectiveness depends on educators having sufficient AI literacy; otherwise, such tools risk creating only an illusion of oversight rather than meaningful control.

Figure 2 shows a SHAP summary plot at the class level, showing the average contribution of grading features across a cohort. Unlike the LIME visualisation in Figure 1, this provides a global explanation of model behaviour rather than an individual case. For educators, it serves as a validation tool against the assessment rubric—if features such as language and expression appear disproportionately influential relative to their intended weighting, this may indicate a misalignment between the model's behaviour and the grading criteria.

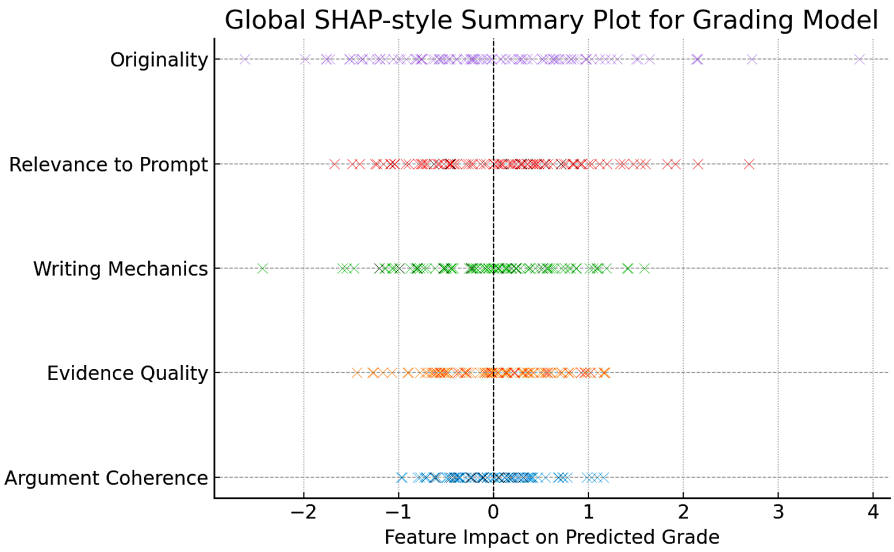

Figure 2: SHAP global summary plot of grading model feature impacts for a class

Reliance on explanation interfaces, while efficient, may inadvertently shift educators' roles from evaluators to reviewers of machine-generated outputs, potentially weakening professional autonomy over time.

Administrators (Tier 1 - System-Level Oversight)

Explanation requirements: Educational leaders and administrators require system-level transparency to support governance and compliance. This includes global explanations of system behaviour across cohorts, performance analytics across courses and instructors, audit trails of grading decisions and interventions, compliance reporting, and insights for resource allocation.

Example explanation:

> System performance dashboard showing that 4,847 essays have been graded this semester, with 91% agreement with human validators, dropping to 74% in Philosophy and 78% in History. No significant bias was detected across demographic groups ($p > 0.05$). Educator intervention rate: 6.1% (up from 3.8% last semester), concentrated in Philosophy and History. Appeals success rate: 12% of contested grades, primarily in creative writing assessments. Of these 8% involved essays where the writing style differed substantially from the training data distribution.

These explanations may combine tools such as aggregated SHAP reports to track feature importance over time, bias trend analyses across demographic groups, and compliance dashboards integrating model logs, explanation coverage, and intervention rates.

These represent the most technically demanding explainability requirements within the tiered framework, and, in practice, the those most frequently neglected. Sustaining them requires not only technical infrastructure but also designated institutional roles, data governance procedures, and scheduled review cycles. For many institutions, particularly those that are smaller or under-resourced, this constitutes a continuous investment rather than a one-time implementation cost. This feasibility constraint must be addressed directly in any deployment strategy. Treating explainability at administrator-level as a purely technical deliverable, without accounting for the organisational capacity required to maintain it, may result in the failure to operationalise AI governance.

Figure 3 shows an administrator-level explainability dashboard displaying agreement rates with human reviewers and bias flags over time. Unlike Figures 1 and 2, this output is not tied to individual cases or cohorts, but rather captures system-wide behaviour throughout its operational history. This supports audit trail and drift monitoring requirements. Rather than interpreting its own outputs, the dashboard surfaces trends, such as a decline in agreement rates in a specific department or an increasing proportion of flagged submissions in a demographic group.

While these tiered explanation requirements illustrate the need to tailor explainability to different stakeholders, they also reveal a broader regulatory dimension. The following section therefore outlines the legal framework governing AGSs.

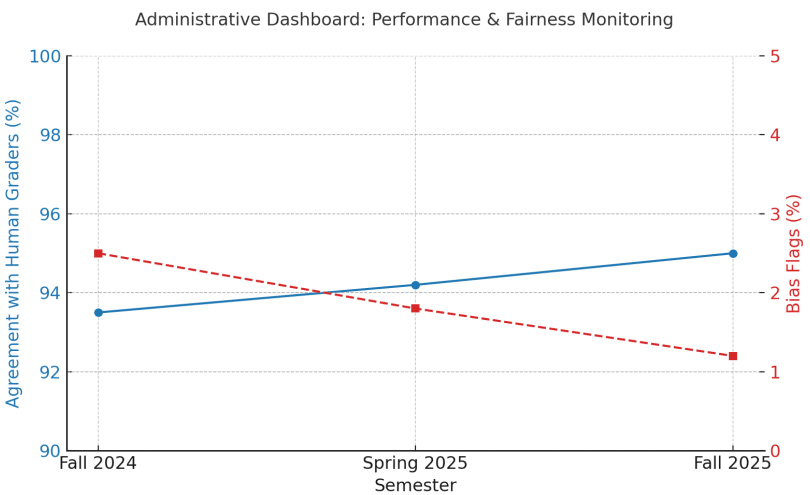

Figure 3. An example of agreement rate with human graders and potential bias for supporting human oversight

Legal Obligations and Compliance Risks for AGSs

AGSs are typically classified as high-risk under the AI Act, given their impact on educational outcomes and opportunities.

Transparency and Rights to Explanation

As discussed above, the AI Act and GDPR establish a framework of transparency, contestation, and information rights for high-risk AI systems. In the context of AGSs, these obligations require that deployers can understand and use systems appropriately (Article 13, AI act), while affected individuals receive accessible explanations of grading logic (Article 86, AI Act). In practice, this translates into differentiated explanation formats, such as simplified feedback for learners and more detailed outputs for educators.

Where personal data is processed, GDPR obligations apply, including lawful processing (Article 6, GDPR), and additional safeguards for sensitive data (Article 9, GDPR). Learners must be informed about the use and logic of AI systems (Articles 12–14), and they have the right to contest decisions and request a human review (Article 22(3)).

Human Oversight

Article 14, AI Act requires that deployers implement mechanisms for effective human oversight and intervention. In the context of AGSs, this entails enabling educators to critically assess and, where necessary, override automated grades. Explainability is essential, but insufficient without adequate AI literacy and institutional support. Where these are lacking, the risk of automation bias increases, with educators deferring to system outputs and potentially compromising fairness and academic integrity (Laux & Ruschemeier, 2025). As a result, formal oversight mechanisms may exist without ensuring meaningful intervention in practice.

Accountability

In practice, the roles of providers and deployers may overlap, particularly when institutions co-develop or fine-tune systems. This creates uncertainty around who is responsible for explainability and user rights. This may result in compliance gaps, particularly regarding access to explanations and the enforcement of user rights.

Traceability requirements under Article 12, AI act and accountability requirements under Article 5(2), GDPR further require institutions to demonstrate compliance. However, many institutions lack the governance structures to operationalise these obligations, which limits their effectiveness in practice (Kilian et al., 2025).

Impact Assessments

As discussed above, the AI act and GDPR establish a dual impact assessment framework though FRIAs and DPIAs (Mantelero et al., 2025). In the context of AGSs, these requirements translate into specific implementation challenges, particularly in assessing pedagogical fairness, bias in grading outcomes, and the effectiveness of human oversight mechanisms. However, their practical value depends on institutional capacity to operationalise these assessments (Kilian et al., 2025).

Impact on Key Stakeholders

These legal obligations outlined apply to all stakeholders.

1. Learners have the right to be informed of automated grading, to understand the reasoning behind outcomes and to challenge decisions.

2. Institutions must safeguard these rights by establishing clear procedures and ensuring that communication is accessible.

3. Educators act as the primary layer of human oversight. They must interpret and validate automated grades, and override them where necessary. This requires usable interfaces, as well as institutional support through training and governance structures.

4. Edtech providers must embed explainability in their designs from the outset. This includes providing clear documentation, system logs and user guidance that are accessible to non-technical users. They must also ensure that their systems can be applied meaningfully across different educational contexts.

5. Institutional leaders and administrators, as deployers, coordinate compliance through procurement, oversight, staff training, and the maintenance of audit mechanisms.

6. Finally, policymakers and regulators, such as Data Protection Authorities and national AI oversight bodies, must provide interpretative guidance and ensure consistent enforcement. By promoting harmonised standards across EU Member States, regulators can reduce the risk of fragmented implementation and strengthen trust in automated assessment practices.

Taken together, these roles highlight that explainability in AGSs must be approached as a shared responsibility across the educational ecosystem.

Technical Implications

The stakeholder analysis shows that explainability in AGSs is not a single technical feature, but rather a layered framework that varies according to the user, the decision-making context and legal obligations. Translating these requirements into systems that comply with the regulatory requirements involves four interconnected technical challenges, each with institutional and engineering implications.

First, model architecture must balance accuracy with interpretability. Where grading decisions have significant consequences, complex, opaque models, such as deep transformers, should be complemented by post-hoc explanation methods. Techniques such as SHAP and LIME should be integrated into the model pipeline in a way that enables real-time generation of local and global explanations without degrading system performance, and with awareness of the limitations each method carries for different stakeholder groups.

Second, the scope, complexity and terminology of outputs should adapt to different users (learners, educators and administrators), drawing on the same underlying data but being tailored to varying levels of expertise.

Third, explainability must be supported by robust logging, version control, and model monitoring. Each decision should be traceable to the relevant model version, data, and features, enabling audits and bias detection. This also introduces significant data management demands that must be addressed at the procurement stage.

Fourth, XAI systems should allow educators to review, override, and annotate decisions, feeding corrections back into the model to improve accuracy and pedagogical alignment over time.

Table 3 maps the relationship between key technical requirements, corresponding legal obligations under the AI Act and GDPR, and their implications for stakeholders.

These requirements demonstrate that explainability cannot be retrofitted after deployment. As shown in Table 3, each technical element requires early design decisions and sustained institutional resources throughout the system lifecycle. The following section advances a proactive, lifecycle-integrated strategy that translates legal obligations into practical safeguards, recognising that retrofitting explainability is significantly more costly than embedding it from the outset.

Table 3. Operationalising explainability: A technical-legal stakeholder mapping for AGSs

Technical requirement	Legal basis (AI Act/GDPR)	Stakeholder implications
Model architecture (accuracy and interpretability)	Transparency and interpretability (Article 13, AI Act).	Learners: Understandable grading outcomes. Educators: interpretable outputs for validation. Administrators: assurance of system reliability.
Role-aware explanation delivery	Clear and accessible information (Article 13, AI Act). Transparent communication (Articles 12-14, GDPR).	Learners: simple, actionable feedback. Educators: detailed explanations for oversight. Administrators: aggregated insights for governance.
Logging, versioning, and monitoring	Record-keeping, traceability (Article 12, AI Act). Accountability (Article 5(2), GDPR).	Educators: traceable decisions for review. Administrators: audit trails and compliance reporting. Regulators: verifiable oversight.
HITL mechanisms	Human oversight (Article 14, AI Act). Right to human review (Article 22(3), GDPR).	Learners: ability to challenge decisions. Educators: authority to override and validate outputs. Institutions: safeguards for fairness and accountability.

OPERATIONALISING EXPLAINABILITY: A COMPLIANCE AND DESIGN STRATEGY

Rationale for a Strategic Approach

Explainability should be treated as a core design and governance objective. This entails a shift from viewing explainability as a purely technical feature to recognising it as a socio-technical governance mechanism. In this context, the proposed strategy addresses three interrelated challenges: (i) translating legal requirements into practice; (ii) clarifying stakeholder roles; and (iii) embedding explainability across the AI lifecycle.

Lifecycle Phases of AI Deployment in Education

To structure this approach, explainability obligations are mapped onto the following key lifecycle stages, see Table 4.

Table 4. Development lifecycle phases

Lifecycle phase	Explainability objective
Design and development	Embed interpretability features and model documentation
Procurement and Selection	Assess vendor claims and compliance with legal explainability norms
Implementation and use	Tailor outputs to users (e.g. teachers, learners); enable oversight
Monitoring and evaluation	Audit outcomes, maintain logs, and flag anomalies
Review and redress	Support contestability, user feedback, and system updates

Design and development

The design of explainable AI systems begins with a co-design phase in which stakeholders define what should be explained, to whom and in what form. This stage serves both pedagogical and legal functions, linking use cases (e.g. grade rationales, feedback and auditability) to regulatory requirements, such as transparency requirements (Article 13, AI Act) and information obligations (Articles 12-14, GDPR). Collaboration is required between developers, educators, data protection officers and, where, possible, learner representatives. However, stakeholder participation in co-design is often constrained in practice, particularly for learners, raising questions about whose perspective shapes explainability requirements (Selwyn, 2019).

Explainability must be embedded into the system architecture from the outset. Developers should prioritise interpretable models, document system logic and training data, align assessment rubrics with algorithmic scoring and integrate HITL mechanisms to support oversight and reduce automation bias (Article 14, AI Act). However, as discussed earlier, prioritising interpretability may introduce trade-offs with model performance and scalability. Close collaboration with educators should help mitigate these issues and ensure that explanations are both technically robust and pedagogically meaningful (Ravi et al., 2025).

Procurement and selection

At the procurement stage, institutions must evaluate whether AI systems meet regulatory and pedagogical requirements. This includes assessing data structures, model architectures and explanation methods to ensure that outputs are meaningful for various stakeholders. Where feasible, interpretability should be embedded directly into the model ante hoc (e.g.

through the use of a rubric-constrained scoring system or monotonic sub-scores) (Article 13, AI Act). For more complex models, post hoc techniques such as SHAP, LIME or counterfactual explanations can be employed, provided their outputs are validated for stability and accuracy.

Explanation requirements differ across stakeholders, from local, actionable feedback for learners to system level analytics for administrators. While this differentiation is necessary, it also introduces challenges in ensuring consistency and coherence across explanation types (Miller, 2019).

In practice, institutions often rely on vendor-provided documentation and claims, which may limit their ability to independently assess explainability and compliance. This creates a risk of procuring opaque systems that are difficult to govern and may embed dependencies on external providers, thereby reducing institutional control over explainability and oversight (Pasquale, 2015).

Implementation and use

This phase operationalises explainability in practice through multi-level explanation frameworks that support both usability and compliance. Table 5 illustrates this framework.

Table 5. Multi-level explanation framework

Dimension	Options
Scope	Global (system logic) vs. Local (individual decision)
Depth	Comprehensive (detailed rationale) vs. Selective (focused summary)
Alternatives	Contrastive ("Why B, not A?") vs. Non-contrastive ("B because…")
Flow	Conditional ("If X, then Y") vs. Correlational (patterns across time or learners)

The implementation phase must focus on integrating these explanation mechanisms into the user experience in a way that is both usable and legally meaningful (Miller, 2019). For learners, the interface might present a grade breakdown with the two most influential factors, a short counterfactual suggestion aligned with the rubric, and a direct link to appeal procedures, reflecting GDPR requirements on transparency and data subject rights (Articles 12-15).

Educators should be equipped with diagnostic tools that extend beyond grading outputs. These may include global feature importance visualisations, detailed rubric-aligned rationales, and mechanisms for overriding grades with annotated justifications (human oversight obligations, Article 14, AI Act). Administrative/ educational leaders as users, by

contrast, require access to fairness dashboards, drift detection outputs, and full audit trails exportable to support compliance and auditability (Article 12, AI Act).

Where AGSs process personal data and generate grading decisions that may have legal or similarly significant effects, GDPR obligations are triggered. In such cases, explanations must be provided in concise and intelligible form, with clear avenues for contestation (Articles 13, 15, and 22 GDPR). These duties are particularly salient in education, where minors are often affected and varying levels of digital literacy must be accommodated.

Before deployment, comprehensive quality assurance is necessary to confirm that explanations are faithful to the underlying model logic, pedagogically plausible, and fair across learner groups. Validation should combine technical tests, such as perturbation analysis to verify that explanations respond appropriately to features that are altered. Educators can assess plausibility by rating whether the explanations match rubric changes, with human-centred evaluation, in which educators assess plausibility against rubric explanations. This pre-deployment phase should also involve red-teaming exercises to probe for vulnerabilities, dry-runs of rights-based requests, and the finalisation of FRIA (Article 29, AI Act) and DPIA (Article 35, GDPR) documentation.

Once deployed, continuous monitoring ensures ongoing compliance. Key performance indicators include explanation coverage, override rates, bias across groups and appeal outcomes. Ongoing governance processes, such as regular reviews and training, support accountability obligations under Article 5(2), GDPR and ensure that systems remain aligned with AI Act requirements on risk management and oversight. Clear appeal procedures must guarantee human review, in line with Article 22(3), GDPR.

By embedding these processes into the lifecycle of AGS development and deployment, institutions can ensure that explainability serves its dual role: meeting the stringent requirements of the AI Act and GDPR while providing pedagogically meaningful transparency that fosters trust, fairness, and accountability in educational assessment. However, the effectiveness remains contingent on institutional capacity, governance structures, and AI literacy, without which explainability risks functioning as a formal requirement rather than a substantive safeguard (OECD, 2024).

Practical Feasibility of the Compliance and Design Strategy

While the proposed compliance and design strategy provides a structured approach to embedding XIA as a design and governance mechanism, the feasibility of implementation varies significantly depending on institutional capacity, technological maturity, and the evolving regulatory landscape (Selwyn, 2019; Holmes et al., 2022).

A central constraint is the financial implications. Costs arise not only during the initial development stage, when systems may need to be redesigned, but also throughout the AI deployment lifecycle, including the monitoring, logging, auditing and documentation processes required for DPIAs and FRIAs. Moreover, institutions may incur costs relating to legal consultations, procurement processes and the integration of third-party explainability tools. Larger or well-resourced institutions with established digital infrastructures may be better positioned to absorb the financial burden, while smaller or under-resourced institutions may face disproportionate challenges. As observed by Selwyn (2019) and Holmes et al. (2022), the uneven distribution of digital resources in education continues to influence the adoption and implementation of emerging technologies, including AI.

From a technical perspective, integrating explainability mechanisms poses significant challenges, particularly regarding legacy systems and complex, opaque models. Many educational institutions rely on pre-existing digital infrastructures that may not easily accommodate an explainability-by-design approach. Retrofitting such systems to accommodate transparency, traceability, and auditability may require substantial system redesign or replacement.

Institutional governance structures are key to successful implementation. This may entail establishing new roles, such as AI compliance officers, data stewards, and FRIA coordinators. However, as Williamson and Eynon (2020) and the OECD (2021) point out, many educational institutions lack the governance maturity needed to manage complex digital systems and the risks associated with them. In such contexts, there is a risk of compliance mechanisms becoming fragmented or applied inconsistently.

Effective implementation presupposes a baseline level of AI literacy among educators, administrators, and other stakeholders. As emphasised by both UNESCO (2021) and the European Commission (2026), AI literacy is essential for the responsible deployment of AI. However, current levels of understanding within the education sector remain uneven. Without targeted training, there is a risk of both misusing AI and becoming overly reliant on its outputs.

The interaction between the AI Act and the GDPR gives rise to legal ambiguities and potential conflicts. For instance, the transparency and traceability requirements of the AI Act may conflict with the data minimisation and purpose limitation principles of the GDPR, particularly when detailed logging for auditability increases data collection and retention. The lack of clear standards for 'meaningful explanation' (Article 14, AI Act) makes implementation even more challenging. Further, the technical limitations of explainability restrict the practical implementation of legal transparency requirements, particularly in opaque models. While post hoc methods can provide some insight, they cannot guarantee reliable or meaningful explanations, leaving an 'explainability gap' between legal expectations and technical capabilities.

Despite these challenges, there are several pathways that may aid the feasibility of the compliance and design strategy.

First, the development of standardised tools (e.g. FRIA templates) can reduce complexity and support consistent implementation across institutions. This aligns with emerging efforts at both national and EU levels to provide practical guidance for AI governance.

Second, a phased approach to implementation may allow institutions to prioritise high-risk systems while gradually expanding compliance measures to other areas. This risk-based approach is consistent with the logic of the AI Act and may help to balance regulatory requirements with resource constraints.

Third, investment in targeted capacity-building initiatives will be essential for equipping stakeholders with the skills need to engage with AI system. Such initiatives should be tailored to different user groups and integrated into broader institutional strategies for digital transformation.

Finally, greater collaboration between institutions, regulators and technology providers could encourage the exchange of best practices and the development of interoperable solutions. This would reduce duplication and promote economies of scale (OECD, 2019; UNESCO, 2021; European Commission, 2021). This is further supported by established AI governance frameworks, including the National Institute of Standards and Technology AI Risk Management Framework (National Institute of Standards and Technology, 2023) and emerging standards such as International Organization for Standardization/IEC 42001, which emphasise lifecycle-based risk management, organisational accountability, and continuous monitoring, and conceptualise AI governance as an iterative, institutionally embedded process.

Taken together, these measures aim to bridge feasibility gaps and support more consistent and equitable implementation across educational contexts.

RECOMMENDATIONS

Building on this analysis five recommendations are proposed:

First, explainability should be treated as a core design objective from the outset, rather than being added after deployment. Moreover, compliance should be embedded in system architecture, documentation, and institutional policy, and periodically reviewed by ethics or compliance committees in light of evolving legal and pedagogical standards.

Second, explanations should be tailored to the needs of stakeholders to support understanding, human oversight and contestation. The accessibility and intelligibility of

formats should be empirically tested, particularly with regard to minors and those with limited digital literacy.

Third, participatory design processes should be promoted to ensure that relevant stakeholders are involved in shaping how AGSs operate and communicate. Institutions should adopt practical toolkits, training programmes, and internal review protocols, both to meet their regulatory obligations and build a culture of responsible AI use.

Fourth, the non-empirical nature of this study limits the ability to evaluate how the proposed framework operates in real-world institutional settings. Future research should therefore explore empirical validation through pilot implementations and user studies across diverse educational environments.

Finally, further research is needed to advance on explainability in education. Future work should explore innovative approaches to explainability, including the exploration of innovative approaches such as gamified explanations for learners, open-source auditing tools, and standardised explainable AI practices within educational technologies. Longitudinal studies should investigate the effect of explanation practices on trust, perceptions of fairness and learning outcomes over time.

CONCLUSION

In education, where AI-supported grading systems (AGSs) shape performance evaluation, learner self-perception, and access to opportunities, the ability to understand and challenge AI outputs is essential. Explainability thus should be embedded within broader governance frameworks that integrate human oversight, capacity-building, and regulatory compliance.

As this paper has shown, the AI Act and the GDPR require AI systems in education to be explainable, traceable, and contestable. Operationalising these requirements demands the integration of technical design, deployment practices, and institutional governance. Technical tools, such as interpretable models, logging, and post hoc explanations, should be supported by organisational measures, including staff training, FRIA/DPIA processes, and AI literacy initiatives. These efforts require careful management of trade-offs: highly complex models may undermine comprehensibility, while overly detailed explanations may reduce usability. In educational contexts, where trust, fairness, and feedback are central, explainability should therefore be prioritised alongside meaningful human oversight. As models, curricula, and assessment practices evolve, explanation mechanisms must be regularly reviewed to maintain alignment and avoid technical drift. Cross-institutional collaboration, including the sharing of tools and practices, can support more consistent standards. Ethical oversight, through institutional review bodies or governance committees, further ensures that explanation practices meet both legal and pedagogical expectations.

Only through the combination of legal foresight, technical clarity, and educational commitment can explainability move beyond procedural compliance exercise to become a transformative principle that strengthens trust, fairness, and learner empowerment in the age of AI.

REFERENCES

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bellas, F., Ooge, J., Roddeck, L., Rashheed, H., Prifti Skenduli, M., Masdoum, F., Zainuddin, N., Niewint, J., Costello, E., Kralj, L., Dcosta, D., Katsamori, D., Neethling, D., Maat, S., Saurabh, R., Alasgarova, R., Radaelli, E., Stamatescuù, A., Blazic, A., . . . Obae, C. (2025). *Explainable AI in education: Fostering human oversight and shared responsibility* (Technical Report). European Digital Education Hub; Publications Office of the European Union. https://data.europa.eu/doi/10.2797/6780469

Burger, C., Chen, L., & Le, T. (2023). Are your explanations reliable? Investigating the stability of LIME in explaining text classifiers by marrying XAI and adversarial attack. *arXiv.* https://doi.org/10.48550/arXiv.2305.12351

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). ACM. https://doi.org/10.1145/2783258.2788613

Dimari, A., Tyagi, N., Davanageri, M., Kukreti, R., Yadav, R., & Dimari, H. (2024). AI-based automated grading systems for open book examination system: Implications for assessment in higher education. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)* (pp. 1–7). IEEE. https://doi.org/10.1109/ICKECS61492.2024.10616490

Edwards, L., & Veale, M. (2017). Slave to the algorithm? Why a "right to an explanation" is probably not the remedy you are looking for. *Duke Law & Technology Review, 16*(1), 18–84. https://doi.org/10.2139/ssrn.2972855

European Commission. (2026). *Guidelines on the ethical use of artificial intelligence and data in teaching and learning for educators.* Publications Office of the European Union. https://op.europa.eu/s/Aeo7

Figueras, C., Rossitto, C., & Cerratto Pargman, T. (2024). Doing responsibilities with automated grading systems: An empirical multi-stakeholder exploration. In *Proceedings of the 13th Nordic Conference on Human-Computer Interaction* (pp. 1–13). ACM. https://doi.org/10.1145/3679318.3685378

Fiok, K., Farahani, F. V., Karwowski, W., & Ahram, T. (2022). Explainable artificial intelligence for education and training. *The Journal of Defense Modeling and Simulation, 19*(2), 133–144. https://doi.org/10.1177/15485129211028651

Fülöp, T., & Poindl, P. (2025). Article 27: Fundamental rights impact assessment for high-risk AI systems. In C. N. Pehlivan, N. Forgo, & P. Valcke (Eds.), *The EU Artificial Intelligence (AI) Act* (pp. 571–590). Wolters Kluwer.

Hacker, P., Krestel, R., Grundmann, S., & Naumann, F. (2020). Explainable AI under contract and tort law: Legal incentives and technical challenges. *Artificial Intelligence and Law, 28*(4), 415–439. https://doi.org/10.1007/s10506-020-09260-6

Holmes, W., Bialik, M., & Fadel, C. (2022). Artificial intelligence in education: Promises and implications for teaching and learning. Center for Curriculum Redesign.

Information Commissioner's Office, & The Alan Turing Institute. (2020). *Explaining decisions made with AI.* https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/explaining-decisions-made-with-artificial-intelligence/

Kilian, R., Jäck, L., & Ebel, D. (2025). European AI standards: Technical standardization and implementation challenges under the EU AI Act. *SSRN*. https://doi.org/10.2139/ssrn.5155591

Laux, J., & Ruschemeier, H. (2025). Automation bias in the AI Act: On the legal implications of attempting to de-bias human oversight of AI. *SSRN*. https://ssrn.com/abstract=5117560

Mantelero, A., Guzmán, C., García, E., Ortiz, R., & Mor, M. A. (2025). *FRIA model: Guide and use cases. FRIA methodology for AI design and development*. Catalan Data Protection Authority.

McConville, M., & Chui, W. H. (2017). *Research methods for law* (2nd ed.). Edinburgh University Press.

Miller, T. (2019) Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T. (2023). Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 333–342). ACM. https://doi.org/10.1145/3593013.3594001

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT 2019)* (pp. 279–288). ACM. https://doi.org/10.1145/3287560.3287574

National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). https://doi.org/10.6028/NIST.AI.100-1

OECD. (2019). Recommendation of the Council on Artificial Intelligence. OECD/LEGAL/0449. OECD Legal Instruments. https://oecd.ai/en/assets/files/OECD-LEGAL-0449-en.pdf

OECD. (2024). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449, amended 2024). OECD Legal Instruments. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Onatere-Ubrurhe, J. O., & Ubrurhe, O. (2025). Leveraging artificial intelligence to redesign TVET assessment systems for enhancing creativity and innovation in technical education. *International Journal of Vocational and Technical Education Research, 11*(2), 1–20.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information.* Harvard University Press.

Ravi, P., Masla, J., Kakoti, G., Lin, G. C., Anderson, E., Taylor, M., Ostrowski, A. K., Breazeal, C., Klopfer, E., & Abelson, H. (2025). Co-designing large language model tools for project-based learning with K-12 educators. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)* (pp. 138:1–138:25). ACM. https://doi.org/10.1145/3706598.3713971

Roddeck, L., Abu-Rasheed, H., Ooge, J., & Bellas, F. (2025). *Educational stakeholders need a legal guide on explainable AI* [Paper presentation]. XAI-Ed: Pedagogy-Founded Explainable AI for Transparent, User-Centred AI in Education — Workshop at the 26th International Conference on Artificial Intelligence in Education (AIED 2025), Palermo, Italy.

Rozenblit, L., & Keil, F (2002) The misunderstood limits of folk science: An illusion of explanatory depth. *Cognitive Science, 26*, 521–562. https://doi.org/10.1207/s15516709cog2605_1

Russell, S., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (4th ed.). Pearson.

Saeed, W., & Omlin, C. (2023). Explainable AI (XAI): A systematic meta-survey of current challenges and future opportunities. *Knowledge-Based Systems, 263*, 110273. https://doi.org/10.1016/j.knosys.2023.110273

Selbst, A. D. (2021). An institutional view of algorithmic impact assessments. *Harvard Journal of Law & Technology, 35*(1), 117–194. https://jolt.law.harvard.edu/assets/articlePDFs/v35/Selbst-An-Institutional-View-of-Algorithmic-Impact-Assessments.pdf

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education.* Polity Press.

Tseng, E.-Q., Huang, P.-C., Hsu, C., Wu, P.-Y., Ku, C.-T., & Kang, Y. (2024). CodEv: An automated grading framework leveraging large language models for consistent and constructive feedback. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 5442–5449). IEEE. https://doi.org/10.1109/BigData62323.2024.10825949

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence.* https://unesdoc.unesco.org/ark:/48223/pf0000381137

Wachter, S. (2024). Limitations and loopholes in the EU AI Act and AI liability directives: What this means for the European Union, the United States, and beyond. *Yale Journal of Law & Technology, 26*(3), 671–718.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology, 31*(2), 841–887.

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology, 45*(3), 223–235. https://doi.org/10.1080/17439884.2020.1798995

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). Chapman & Hall/CRC. https://doi.org/10.1201/9781315370279