http://dx.doi.org/10.18623/rvd.v22.n3.3553

RETHINKING EXPLAINABILITY IN EDUCATIONAL ARTIFICIAL INTELLIGENCE: A CRITICAL SYSTEMATIC REVIEW OF MODELS, APPLICATIONS, AND ETHICAL DIMENSIONS

REPENSANDO A EXPLICABILIDADE NA INTELIGÊNCIA ARTIFICIAL EDUCACIONAL: UMA REVISÃO SISTEMÁTICA CRÍTICA DE MODELOS, APLICAÇÕES E DIMENSÕES ÉTICAS

Article received on: 7/29/2025
Article accepted on: 9/29/2025

Rusen Meylani*
* Dicle University, Ziya Gokalp Faculty of Education, Diyarbakir, Turkiye
Orcid: https://orcid.org/0000-0002-3121-6088
rusen.meylani@dicle.edu.tr

The authors declare that there is no conflict of interest

Abstract
The increasing integration of Artificial Intelligence (AI) into educational technologies has elevated the importance of explainability to ensure transparency, pedagogical relevance, and ethical accountability. This systematic review critically examines empirical studies on Explainable Artificial Intelligence (XAI) in educational contexts published between 2015 and 2025. Guided by the PRISMA 2020 protocol, the review synthesizes findings from 106 peer-reviewed studies across K–12, higher education, and teacher training environments. The analysis identifies six interrelated themes: (1) types of XAI models and techniques, (2) domains of application in educational technologies, (3) operationalization of explainability, (4) impacts on learning and teaching, (5) challenges and limitations, and (6) ethical and epistemological considerations. Results reveal that while technical implementations of XAI are expanding, their pedagogical grounding remains limited, with explanations frequently framed as system outputs rather than tools for epistemic engagement. Explainability is often operationalized through vague definitions and evaluated using affective or behavioral proxies, with minimal attention to learning outcomes or cognitive development. Furthermore, ethical and cultural considerations are inconsistently addressed, raising concerns about inclusion, equity, and user agency. The review proposes a tripartite framework—pedagogical, epistemic, and ethical explainability—to guide future research and design. It calls for participatory, theory-informed, and context-sensitive approaches to ensure that educational XAI fosters critical thinking, learner autonomy, and instructional equity.

Resumo
A crescente integração da Inteligência Artificial (IA) em tecnologias educacionais elevou a importância da explicabilidade para garantir transparência, relevância pedagógica e responsabilidade ética. Esta revisão sistemática examina criticamente estudos empíricos sobre Inteligência Artificial Explicável (IAE) em contextos educacionais publicados entre 2015 e 2025. Guiada pelo protocolo PRISMA 2020, a revisão sintetiza resultados de 106 estudos revisados por pares em ambientes de ensino fundamental e médio, ensino superior e formação de professores. A análise identifica seis temas inter-relacionados: (1) tipos de modelos e técnicas de IAE, (2) domínios de aplicação em tecnologias educacionais, (3) operacionalização da explicabilidade, (4) impactos no ensino e na aprendizagem, (5) desafios e limitações e (6) considerações éticas e epistemológicas. Os resultados revelam que, embora as implementações técnicas de IAE estejam se expandindo, seu embasamento pedagógico permanece limitado, com explicações frequentemente enquadradas como saídas do sistema em vez de ferramentas para engajamento epistêmico. A explicabilidade é frequentemente operacionalizada por meio de definições vagas e avaliada usando indicadores afetivos ou comportamentais, com atenção mínima aos resultados de aprendizagem ou ao desenvolvimento cognitivo. Além disso, as considerações éticas e culturais são abordadas de forma inconsistente, levantando preocupações sobre inclusão, equidade e autonomia do usuário. Esta revisão propõe uma estrutura tripartite — explicabilidade pedagógica, epistêmica e ética — para orientar pesquisas e projetos futuros. Ela defende abordagens participativas, fundamentadas em

Keywords: Explainable Artificial Intelligence. Educational Technology. Epistemic Agency. Pedagogical Explainability. XAI. Systematic Review. AI Ethics. Instructional Design.

teoria e sensíveis ao contexto para garantir que a IA explicável educacional promova o pensamento crítico, a autonomia do aluno e a equidade instrucional.

Palavras-chave: *Inteligência Artificial Explicável. Tecnologia Educacional. Agência Epistêmica. Explicabilidade Pedagógica. IA Explicável. Revisão Sistemática. Ética da IA. Design Instrucional.*

1 INTRODUCTION

1.1 Background and rationale

The integration of Artificial Intelligence (AI) into educational technologies has transformed the landscape of teaching and learning by enabling personalized, scalable, and data-driven instructional experiences. AI systems offer tailored pathways aligned with learners' needs, preferences, and behavioral patterns, a feature increasingly vital in heterogeneous classrooms where traditional models often fall short in effectiveness (Meske et al., 2021; Türkmen, 2025). However, the rapid adoption of AI has also introduced pressing concerns regarding system opacity, interpretability, and ethical accountability. Teachers and students frequently struggle to understand AI-generated feedback, which, in many instances, lacks transparency and pedagogical clarity (Gunning & Aha, 2019; Meske et al., 2021; Türkmen, 2025).

Explainable Artificial Intelligence (XAI) has emerged as a promising response to these challenges. XAI encompasses methodologies and design strategies that render AI processes more interpretable and trustworthy by clarifying how systems arrive at their conclusions (Darwish, 2022; Meske et al., 2021). The necessity for explainability is especially critical in educational contexts, where users—particularly younger students—often lack the technical expertise to decode algorithmic outputs. Teachers also require actionable, intelligible insights that support instructional planning, not merely abstract performance metrics (Darwish, 2022). Therefore, the development of explainable systems in education must extend beyond technical transparency to address pedagogical usefulness and cognitive relevance (Göçen & Aydemir, 2020; Meske et al., 2021).

Rather than focusing solely on showcasing AI's functional capacities, educational implementations must prioritize explanations that contribute meaningfully to learning and foster students' epistemic engagement. This shift underscores the critical role of XAI in shaping inclusive, effective, and ethically grounded learning ecosystems.

1.2 Problem statement and gaps

Despite the increasing interest in XAI within educational research, the literature reveals significant gaps concerning its pedagogical and epistemological foundations. Much of the current scholarship remains concentrated on interface design, technical performance, or general user satisfaction, often neglecting how explanations affect learning processes, student agency, or reflective understanding (Allahverdiyeva, 2024; Zhai et al., 2023). This narrow orientation fails to address the complex, context-dependent nature of what makes an explanation educationally effective (Meske et al., 2021; Türkmen, 2025).

The prevailing assumption that more detailed explanations inherently benefit users disregards potential drawbacks such as cognitive overload or interpretive confusion, particularly among novice learners (Guan, 2023; Gunning & Aha, 2019; Türkmen, 2025). Furthermore, the dominant evaluation metrics—such as perceived trust or satisfaction—lack alignment with learning outcomes or equity considerations (Liulka et al., 2024). Consequently, the field has yet to establish a coherent theoretical framework that centers educational goals and cognitive development in defining explainability.

Moreover, a substantial portion of the literature isolates AI implementation from broader pedagogical theories, resulting in a disjunction between technical solutions and educational principles. This disconnect limits the transformative potential of XAI in classroom settings. Without integrating explainability into pedagogical frameworks, AI tools risk reinforcing rather than mitigating existing disparities in education.

Addressing these issues demands a multidisciplinary approach that incorporates insights from learning sciences, human-computer interaction, and educational ethics. Only through such integration will XAI achieve its potential to enhance instructional practices, support student learning, and uphold the core values of equity and inclusion in education.

1.3 Objectives and critical research questions

This systematic review aims to synthesize and critically analyze empirical research on Explainable Artificial Intelligence in educational technologies published between 2015 and 2025. The objective is not only to catalogue models and applications but to interrogate how explainability is conceptualized, implemented, and assessed in real educational contexts. The review seeks to answer the following critical research questions:

- What types of XAI models have been implemented in educational technologies, and how do they differ in interpretability, usability, and pedagogical alignment?
- How is the concept of "explainability" defined and operationalized across studies, and what theoretical or epistemological assumptions inform these definitions?
- What evidence exists that XAI enhances learning, teaching, or decision-making in educational environments—and how robust, reliable, and generalizable is this evidence?
- What ethical, practical, and epistemological tensions emerge in deploying XAI in real-world educational settings, and how are these challenges addressed?

By adopting a critical stance, this review seeks to move beyond descriptive summaries toward a theoretically informed, context-sensitive synthesis that can guide future research, design, and policy in educational AI.

2 THEORETICAL FRAMEWORK

The systematic analysis of Explainable Artificial Intelligence (XAI) in educational technologies requires an interdisciplinary theoretical foundation. This review draws on a composite framework encompassing three interrelated domains: (1) theories of learning and cognition; (2) epistemological perspectives on explanation and understanding; and (3) ethics and critical theory in educational technology. These domains collectively provide the conceptual basis for analyzing the pedagogical relevance, design assumptions, and socio-cultural implications of XAI in education.

2.1 Theories of learning and cognition

Positioning XAI meaningfully within education demands a shift from evaluating algorithmic performance alone to considering how AI aligns with contemporary theories of learning. Constructivist theories, advanced by scholars such as Piaget and Bruner, conceptualize learning as an active process in which learners construct meaning through interaction with their environment and reflection on experiences (Gunning & Aha, 2019; Meske et al., 2021). From this perspective, XAI-generated explanations should serve as cognitive tools that promote knowledge construction, enabling learners to integrate new information with prior understanding, question assumptions, and resolve misconceptions.

Sociocultural learning theories, particularly those rooted in the work of Vygotsky and Wertsch, further emphasize the social and cultural mediators of learning. Meaning is co-constructed through interaction, making communicative clarity and contextual sensitivity essential features of effective explanations (Darwish, 2022; Türkmen, 2025). In this view, XAI systems must generate explanations that are adaptable to the learner's sociocultural background and communicative norms. Rigid or decontextualized explanations risk excluding learners from meaningful engagement, particularly in diverse classrooms.

Metacognitive frameworks complement these perspectives by underscoring the importance of reflective awareness in learning. Explanations must not only convey information but also stimulate learners' self-regulatory processes, encouraging them to monitor, evaluate, and adjust their cognitive strategies (Abu Bakar & Ismail, 2020; Arianto, 2021; Susantini et al., 2021). XAI systems that incorporate metacognitive prompts and tailored scaffolds contribute more effectively to learner autonomy and cognitive development.

2.2 Epistemological perspectives on explanation and understanding

The function and quality of explanations in XAI are inherently shaped by epistemological assumptions. Many current systems rely on a computationalist epistemology, presenting explanations as model-based visualizations, feature attributions, or post hoc rationalizations (Darwish, 2022; Liu et al., 2024). While these outputs fulfill technical requirements for interpretability, they often fail to support

pedagogical goals such as conceptual understanding, epistemic agency, or critical thinking.

Educational research on scientific explanations stresses that effective explanatory practices empower learners to understand underlying mechanisms, evaluate competing claims, and engage in reasoned argumentation (Braaten & Windschitl, 2011; Sandoval, 2014). This review adopts the distinction between algorithmic interpretability—focused on internal model processes—and pedagogical intelligibility, which emphasizes cognitively accessible, context-relevant, and actionable explanations for novice users (Guan, 2023; Meske et al., 2021).

A failure to account for this distinction limits learners' opportunities for deep epistemic engagement. Dialogic and inquiry-based models of learning demand that explanations support critical questioning and iterative understanding rather than passive consumption of information. Thus, XAI must be evaluated not only by its capacity to explain but also by its ability to cultivate learners' epistemological growth.

2.3 Ethics and critical theory in educational technology

XAI in education must also be examined through ethical and critical theoretical lenses. Educational technologies do not exist in neutral spaces; they often reflect and reinforce dominant sociopolitical narratives. Critical theorists have long emphasized that digital tools can perpetuate inequities, particularly when designed without consideration for marginalized learners' experiences (Chan & Zary, 2019; Chaudhry et al., 2022). In this context, XAI must be scrutinized for its potential to either mitigate or exacerbate epistemic injustice—where learners are denied the opportunity to participate meaningfully in their own knowledge construction (Yildiz Durak et al., 2025).

Beyond concerns about algorithmic bias or transparency, this review advocates for a relational ethics framework that emphasizes agency, contextual awareness, and participatory design. Such an approach frames learners and educators not as passive recipients but as co-constructors of educational meaning (Choi et al., 2024; Feldman-Maggor et al., 2024). Ethical explainability entails not only technical openness but also inclusivity in decision-making and relevance to local pedagogical realities.

Absent such a framework, XAI risks functioning as a legitimizing façade—offering surface-level transparency while concealing problematic assumptions or

reinforcing harmful practices. As AI becomes increasingly embedded in educational policy and practice, its ethical foundations must be reexamined to ensure it serves equity, empowerment, and democratic participation in education.

2.4 Integrative framing

XAI must be understood not as a purely technical solution but as a socio-technical construct situated at the intersection of algorithms, cognition, ethics, and institutional practice. This integrative framework provides the basis for critically analyzing the empirical literature, attending to how XAI systems align with learning theories, respect epistemic diversity, promote ethical engagement, and address real educational needs. Rather than defaulting to utilitarian metrics or abstract performance indicators, this review foregrounds pedagogical efficacy, inclusivity, and learner agency as the primary criteria for evaluating explainability in educational AI systems.

3 METHODOLOGY

3.1 Review protocol and reporting standards

This systematic review follows the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) 2020 guidelines (Page et al., 2021) to ensure transparency, replicability, and methodological rigor. The review protocol, including inclusion criteria, search strategy, and coding framework, was preregistered with the Open Science Framework (OSF) prior to data collection.

3.2 Eligibility Criteria

The eligibility criteria were designed to capture empirical studies that examine the implementation, conceptualization, or evaluation of Explainable Artificial Intelligence (XAI) within educational contexts. The inclusion and exclusion criteria are as follows:

- ***Inclusion Criteria****:*
  - Peer-reviewed empirical studies (quantitative, qualitative, or mixed-methods);
  - Published between January 2015 and March 2025;

- Focus on educational applications of XAI in K–12, higher education, teacher education, or professional learning environments;
- Describe, evaluate, or critique explainability techniques, models, interfaces, or outcomes;
- Written in English.

- ***Exclusion Criteria****:*
- Conceptual, theoretical, or opinion pieces with no empirical data;
- Studies focused solely on non-educational domains (e.g., healthcare, finance, military);
- XAI systems used in training AI engineers or data scientists without educational end-user focus;
- Grey literature, non-peer-reviewed preprints, or conference abstracts without full papers.

These criteria reflect a critical lens, prioritizing studies that go beyond proof-of-concept and engage with explainability as it affects learners, educators, or pedagogical practices.

3.3 Information sources and search strategy

A comprehensive database search was conducted across six electronic databases: Scopus, Web of Science, ERIC, IEEE Xplore, ACM Digital Library, and SpringerLink. Searches were restricted to peer-reviewed journal articles and conference proceedings within the defined time frame.

The search strategy combined keywords and Boolean operators, targeting three conceptual dimensions: XAI techniques, education/learning contexts, and evaluation/impact. An example search string: ("explainable AI" OR "XAI" OR "interpretable machine learning") AND ("education" OR "educational technology" OR "learning system*" OR "intelligent tutoring" OR "adaptive learning") AND ("feedback" OR "assessment" OR "student model*" OR "teacher dashboard*" OR "explanation" OR "trust")

All search results were exported to a reference manager (Zotero), and duplicates were removed prior to screening.

3.4 Study selection

The selection process consisted of two stages:

1. ***Title and abstract screening****:* Each study was independently screened by two reviewers for preliminary eligibility.
2. ***Full-text review****:* Eligible studies were then evaluated in full text. Discrepancies were resolved by a third reviewer through consensus.

A PRISMA flow diagram was used to document the study selection process, including the number of records identified, screened, excluded, and retained for analysis.

3.5 Data extraction

A structured data extraction form was developed and pilot-tested to ensure consistency and depth. The following categories were extracted for each study:

- Citation details (authors, year, journal, country);
- Educational setting (K–12, higher education, teacher training, etc.);
- AI model type (e.g., decision tree, neural network, ensemble, reinforcement learning);
- Explainability technique (e.g., SHAP, LIME, saliency maps, rule extraction, attention mechanisms);
- Explanation format (e.g., textual, visual, interactive);
- Definition and conceptualization of explainability;
- Target user (student, teacher, administrator);
- Evaluation method and outcome measures;
- Reported impact (trust, learning outcomes, usability, fairness);
- Ethical or critical considerations discussed (bias, agency, transparency);
- Theoretical framework (if any);
- Methodological quality and limitations.

Special attention was given to how each study defined and operationalized explainability, whether user studies were conducted, and how explanations were interpreted by different stakeholders.

3.6 Critical appraisal and quality assessment

To evaluate the methodological and theoretical rigor of the included studies, the Mixed Methods Appraisal Tool (MMAT, 2018 version) was used. Each study was rated by two independent reviewers. The following criteria were emphasized in the appraisal:

- Clarity and coherence of research questions;
- Alignment between methodology and research goals;
- Validity of outcome measures and explanation evaluation techniques;
- Stakeholder involvement in system design or evaluation;
- Consideration of ethical, pedagogical, or epistemological issues.

Studies were also tagged for critical depth, defined as the extent to which they examined not only technical functionality but also educational and ethical dimensions of explainability.

4 FINDINGS AND RESULTS

4.1 Theme 1: types of XAI models and techniques

Studies reviewed in this analysis reveal a broad spectrum of Explainable Artificial Intelligence (XAI) techniques implemented in educational technologies. However, this diversity reflects recurring design patterns and conceptual limitations rather than a robust plurality of theoretical or pedagogical underpinnings. A critical evaluation of these trends identifies key concerns related to the applicability, interpretability, and educational value of different XAI models.

4.2 Post-hoc explanation techniques

A dominant trend involves the use of post-hoc explanation techniques such as SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations), which generate interpretive outputs after predictions are made. These methods are favored for their model-agnostic flexibility, allowing integration across various algorithmic systems without altering their architecture (Türkmen, 2025). In

educational contexts, they are frequently applied to explain outcomes related to student performance prediction, risk profiling, or personalized recommendations.

Despite their technical appeal, the educational efficacy of these methods remains uncertain. Explanations typically take the form of weighted feature lists or saliency maps, which are rarely embedded in learners' cognitive contexts (Liefooghe & van Maanen, 2022). The gap between algorithmic interpretability—how well experts understand model mechanics—and pedagogical intelligibility—how well learners comprehend and use explanations—raises significant concerns (Feldman-Maggor et al., 2024). Moreover, evaluations of these techniques often rely on technical criteria such as fidelity and sparsity, without verifying whether learners or educators perceive the explanations as meaningful or actionable (Huang, 2023).

4.3 Intrinsic explainability approaches

Some studies favor intrinsically interpretable models such as decision trees, linear regressions, attention-based mechanisms, or rule-based systems. These models are valued for their transparency and apparent ease of understanding (Alrayes et al., 2024). However, their educational relevance is not always clearly articulated. In several applications, such as learning analytics dashboards, attention mechanisms were visualized without clarifying causal relationships or instructional significance, which undermines their potential to enhance learner understanding (Yildiz Durak et al., 2025).

A problematic conflation also emerges between interpretability and trustworthiness, whereby simple explanations are presumed to foster user trust. This assumption disregards the complex dynamics of learner engagement, critical reflection, and contextual interpretation. Empirical investigations into how users—particularly students and educators—interpret and respond to these explanations in authentic settings remain scarce (Gomez et al., 2025).

4.4 Hybrid and visual-symbolic explanation interfaces

A smaller set of studies explores hybrid approaches that integrate textual, visual, and symbolic elements to enhance user understanding. Examples include color-coded progress indicators, interactive feedback panels, and layered visualizations embedded

within intelligent tutoring systems. These methods aim to improve accessibility and foster engagement with AI-generated explanations (Chan & Zary, 2019). However, the pedagogical rationales behind their design are often underdeveloped or absent.

These interfaces are frequently evaluated based on usability rather than their cognitive or instructional impact. Their effects on cognitive load, developmental appropriateness, or alignment with metacognitive and disciplinary learning goals are rarely scrutinized (Madi et al., 2024). This disconnect between technical innovation and pedagogical intent suggests a superficial integration of explanatory elements that risks hindering rather than enhancing learning (Ewals et al., 2024).

4.5 Critical synthesis

The analysis of XAI models in education reveals a predominant prioritization of technical feasibility over pedagogical appropriateness (Table 1). The selection and implementation of XAI tools often reflect general standards for interpretability rather than alignment with established learning theories or instructional design principles. Explainability is frequently treated as a static model attribute rather than a relational, dialogic process between learners, educators, and systems.

This static view fails to account for learner diversity. Many studies implicitly assume user homogeneity, ignoring the fact that explanatory needs vary based on learners' prior knowledge, cognitive development, and educational context. Critically, research focused on early childhood, special education, or low-literacy populations remains limited, raising equity concerns about whose understanding is prioritized in current XAI design.

Addressing these limitations requires a shift toward context-sensitive, pedagogically informed XAI development that values learner agency and supports reflective, meaningful learning experiences.

Table 1

Summary of XAI Models and Techniques in Educational Technologies

Type of XAI Technique	Common Educational Use Cases	Reported Benefits	Critical Observations
Post-hoc (SHAP, LIME, Counterfactuals)	Student performance classification, content recommendation	Model-agnostic explanations; flexible integration	Explanations often not understandable to students; limited pedagogical evaluation
Intrinsic (Decision Trees, Attention Mechanisms)	Learning analytics dashboards, adaptive testing tools	Transparency; ease of interpretation	Transparency conflated with trustworthiness; minimal user interpretation studies
Hybrid and Visual-Symbolic Interfaces	Tutoring systems, interactive feedback environments	Improved interface usability; multimodal support	Usability emphasized over metacognitive or epistemic alignment

4.6 Theme 2: domains of application in educational technologies

Explainable Artificial Intelligence (XAI) has been integrated into a variety of educational technologies, spanning from student-facing systems such as Intelligent Tutoring Systems (ITS) to educator-focused analytics dashboards. This thematic analysis examines not only the technological domains where XAI is applied but also how the nature and educational value of explanations differ across these contexts. The findings reveal significant inconsistencies in design logic, interpretability, and educational alignment, underscoring the need for a more pedagogically grounded approach.

Intelligent Tutoring Systems (ITS)

Intelligent Tutoring Systems represent one of the most prominent areas for XAI integration in education. These systems commonly employ explainable models to provide corrective feedback, guide problem-solving, and clarify errors. However, many explanations in ITS remain static and pre-scripted, offering generic feedback that fails to address individual learners' epistemic needs or misconceptions (Karpouzis, 2024). Statements such as "Your answer was incorrect because X" exemplify a limited approach that rarely encourages reflection or the exploration of alternative strategies.

Few ITS implementations incorporate adaptive or dialogic explanation models that respond dynamically to learner input. This rigidity diminishes opportunities for deeper metacognitive engagement and reduces the potential of XAI to foster meaningful learning (Sadvakassova et al., 2024). The underutilization of learner-centered explanation frameworks reveals a tendency to prioritize system efficiency over instructional depth.

Learning Analytics Dashboards

XAI techniques are also prevalent in learning analytics dashboards designed for educators and administrators. These dashboards employ interpretable models—often decision trees, clustering algorithms, or ensemble methods—to highlight patterns in student performance and suggest interventions. Visual outputs such as feature importance plots or progress graphs are intended to support decision-making (Liu et al., 2024). However, evidence linking these dashboards to improved pedagogical outcomes remains limited (Li, 2023).

A critical issue lies in the lack of co-design with educators, leading to dashboards that are technically functional but pedagogically opaque. Misinterpretation of visualizations or over-reliance on AI predictions risks undermining professional autonomy (Darwish, 2025). While users may trust these systems, they often have minimal understanding of the underlying decision processes, which restricts meaningful engagement with the data.

Feedback and Assessment Tools

A smaller subset of applications explores the role of XAI in formative assessment and feedback systems. These tools aim to clarify grading rationales, suggest corrections, or explain errors. The emphasis is often placed on making explanations concise and visually accessible. Yet, few studies provide evidence that such explanations enhance learning transfer or concept mastery (Marrone et al., 2024).

In many cases, explanations serve to validate the AI's decision after the fact, rather than supporting learners in reflecting on their thinking or improving their understanding. This instrumental use of explainability reduces its pedagogical potential and fails to position explanation as a cognitive or epistemic resource (Song, 2024).

Adaptive Content Recommenders and Personalized Learning Systems

XAI also plays a role in systems that personalize learning trajectories or suggest content based on user data. While these systems aim to support learner agency, explanations are often superficial, such as "You are recommended Topic A because you scored low on Topic B," without elaborating on the pedagogical reasoning or learning goals (Najdawi et al., 2024).

Studies rarely investigate whether learners comprehend, trust, or challenge these recommendations. This absence of critical reflection limits the effectiveness of explainability in promoting self-directed learning. In many instances, explainability

functions more as a system-level requirement than as a tool for fostering educational agency (Yildirim & Celepcikay, 2021).

Critical Synthesis

Across these domains, explainability frequently serves to justify system behavior rather than to advance learner understanding or pedagogical insight (Table 2). A comparative analysis reveals that:

- Explanations directed at students are often scripted, generic, and lack interactivity.
- Explanations for educators rely on visualizations without theoretical grounding or validation in practice.
- Few systems conceptualize explanation as a dialogic process that invites critique or co-construction.
- In high-stakes contexts such as grading or intervention planning, ethical considerations related to bias, transparency, and fairness are often underexplored.

These findings suggest that explainability is frequently treated as an operational feature, detached from the pedagogical or epistemological aims of education. The dominance of ITS and dashboards reflects a narrow view of educational interaction focused on compliance and monitoring rather than inquiry, dialogue, or reflective learning (Chaudhry et al., 2022).

Table 2

Summary of XAI Application Domains in Educational Technologies

Application Domain	Typical XAI Use	Strengths	Critical Observations
Intelligent Tutoring Systems (ITS)	Providing step-by-step feedback or error explanations	Real-time feedback; individual guidance	Static and shallow explanations; lack of adaptivity to learners' needs
Learning Analytics Dashboards	Visualizing student performance, risk, and trends	Insight for teachers; intervention support	Visual explanations not co-designed; risk of teacher misinterpretation
Feedback and Assessment Tools	Explaining grading decisions or offering correction hints	Increased transparency in grading	Explanations treated as justification rather than epistemic support
Adaptive Recommenders and Personalized Learning	Explaining learning path adjustments or topic recommendations	Potential for learner autonomy	Explanations often superficial; lack of learner agency or override mechanisms

4.7 Theme 3: operationalization of explainability

Although the notion of "explainability" is frequently invoked in the literature on Explainable Artificial Intelligence (XAI) in education, its articulation and implementation exhibit marked variability. This thematic analysis reveals persistent conceptual ambiguity and the absence of standardized approaches for operationalizing explainability in educational contexts. While numerous studies aim to improve the transparency of AI systems, few provide clear definitions of the types of explanations offered, their intended audiences, or their anticipated pedagogical functions. As a result, operationalizations often reflect technical priorities rather than educational or epistemological coherence.

Variability in Definitions

Definitions of explainability across the reviewed literature are inconsistent and often underspecified. In many studies, explainability is loosely defined as the user's ability to understand why an AI system produced a given output. However, the type of understanding involved—whether causal, procedural, or statistical—remains unclear (Nurjanah et al., 2024). Vague formulations such as "helping users make sense of AI" or "enhancing transparency" lack precision and fail to articulate how cognitive or pedagogical processes are supported.

Only a minority of studies ground their conceptualizations in learning theory or cognitive science. Without such theoretical anchoring, explanations risk becoming superficial features rather than tools for deep engagement or epistemic development (Ng et al., 2021). This gap hampers the potential of XAI to function as a meaningful pedagogical aid.

Explanation Formats and Delivery

Explanations in educational XAI systems are typically delivered through textual rationales, visualizations (e.g., saliency maps, bar charts), interactive widgets, or hybrid combinations (Robinson et al., 2024). However, there is a notable lack of empirical evidence supporting the educational efficacy of these formats. Visual explanations are often presumed to be intuitive, yet studies rarely investigate whether learners accurately interpret them. Similarly, interactive elements that could encourage dialogue or exploration are frequently underutilized and under-evaluated.

Moreover, explanation formats are rarely differentiated for distinct user groups. Learners, teachers, and administrators possess varying cognitive goals and interpretive

needs, yet many systems adopt a uniform delivery strategy (Kaharuddin et al., 2024). This omission reflects a broader disregard for user-centered design principles, weakening the pedagogical alignment of XAI tools.

Depth and Scope of Explanations

Most educational XAI systems rely on "shallow" explanations, which consist of brief justifications, feature rankings, or rule-based outputs. These explanations often describe what the system did without clarifying why it did so or how the learner can improve their performance (Lin et al., 2024). Few studies explore the development of multi-layered or scaffolded explanations that evolve with user understanding.

Additionally, critical dimensions such as uncertainty, ambiguity, or counterfactual reasoning are rarely integrated, despite their importance for cultivating epistemic agency and critical thinking (Hoya et al., 2024). This trend toward superficial transparency reduces explainability to system justification rather than promoting reflective, inquiry-based learning.

Evaluation Practices

Evaluation methods for explainability in educational settings commonly fall into three categories: user satisfaction, task performance (e.g., accuracy or completion time), and perceived trust (Zhang et al., 2022). While these measures offer insights into affective responses, they fail to capture whether explanations enhance conceptual understanding, metacognitive skills, or long-term learning outcomes. Notably absent are evaluations based on learning gains, explanation quality (e.g., coherence, depth, transferability), or scaffolding effectiveness (Davis, 2024).

Furthermore, most studies rely on brief user interactions or hypothetical scenarios rather than testing explanations in authentic, diverse educational contexts. The absence of longitudinal studies and the lack of demographic differentiation in evaluation design pose serious limitations for understanding the real-world impact of explainable systems ("Innovative Methodologies and Approaches to Teaching with Artificial Intelligence in Ukrainian Higher Education", 2024).

Critical Synthesis

Explainability is frequently operationalized through engineering-driven lenses that prioritize system interpretability over pedagogical value (Table 3). Definitions remain vague, formats are inconsistently justified, and evaluations often emphasize affective or behavioral metrics over cognitive or epistemic outcomes. There is no clear

consensus on what constitutes an effective explanation in education—whether it is clarity, completeness, adaptability, or educational utility.

This fragmentation hinders the field's advancement and risks reducing explainability to a performative feature rather than a transformative educational component. Without nuanced, learner-centered, and theory-informed operationalizations, XAI in education may inadvertently reproduce existing inequities by privileging certain modes of understanding while marginalizing others.

Table 3

Summary of How Explainability is Operationalized in Educational XAI

Aspect of Operationalization	Common Approaches Observed	Critical Observations
Definition of Explainability	Vague definitions emphasizing "understandability" or "transparency"	Lack of theoretical grounding; minimal reference to learning sciences
Explanation Formats	Textual rationales, visualizations (e.g., heatmaps), basic interactivity	Few justifications for format selection; rare differentiation by user type
Depth and Scope	Single-layer explanations focusing on "what" not "why" or "how"	Limited use of scaffolding or layered explanation; low epistemic value
Evaluation Practices	Satisfaction surveys, trust ratings, task performance metrics	Evaluation rarely includes learning gains or deep understanding measures

4.8 Theme 4: impact on learning and teaching

A central justification for incorporating Explainable Artificial Intelligence (XAI) into educational systems lies in its potential to enhance teaching practices and improve learning outcomes. However, the empirical evidence supporting this promise remains inconclusive. The studies reviewed often exhibit methodological limitations, conceptual inconsistencies, and superficial indicators of impact. This section critically examines how XAI affects learners and educators, questioning whether existing explanation models contribute substantively to educational advancement.

Learner Outcomes and Engagement

Several studies report improvements in learner engagement, comprehension, or self-regulation when XAI features are integrated into educational tools. Yet, these claims predominantly arise from short-term user studies that focus on behavioral proxies, such as time-on-task or frequency of interaction (Gomez et al., 2025). Self-reported

satisfaction is frequently interpreted as evidence of success, but few studies assess concrete learning gains such as conceptual understanding, transfer of knowledge, or correction of misconceptions.

In many cases, "understanding" is conflated with passive acceptance of AI decisions, a condition antithetical to the epistemic goals of education. For novice learners—particularly in cognitively demanding domains like mathematics or programming—technically accurate explanations often remain opaque without scaffolding. Complex visualizations or jargon-heavy textual feedback contribute to cognitive overload rather than clarity (Lin et al., 2024). This misalignment between explanatory complexity and learner readiness is rarely addressed, limiting the pedagogical utility of such systems.

Trust and System Usability

XAI systems are often praised for fostering user trust, particularly when compared to opaque "black-box" models. However, many studies fail to distinguish between appropriate trust, which arises from understanding, and misplaced trust, which results from system authority or aesthetic clarity (Ewals et al., 2024). Over-reliance on Likert-scale measures conflates trust with usability or satisfaction, offering little insight into whether users are critically engaging with system outputs.

Furthermore, few studies triangulate user feedback with data such as think-aloud protocols, log analysis, or performance diagnostics. As a result, claims regarding increased trust or learning are often under-theorized and inadequately substantiated (Zhang et al., 2022). The illusion of transparency—where users perceive a system as understandable without actually comprehending its logic—poses a risk of fostering compliance rather than reflective engagement (Kim et al., 2024).

Teacher Perceptions and Instructional Use

Research on educators' engagement with XAI remains limited but offers mixed insights. Teachers often appreciate dashboards that visualize student performance or flag at-risk learners (Liu et al., 2024). Yet, skepticism persists regarding the underlying logic of AI models, particularly when systems fail to disclose their inferential processes. Educators frequently hesitate to act on AI-generated recommendations without access to contextual or pedagogical rationale (Darwish, 2025).

XAI interfaces rarely allow educators to interrogate or adapt explanations. Most systems present fixed outputs, offering minimal interactivity or opportunity for

professional interpretation. This restricts educators' agency, reducing their role to passive recipients of system suggestions. The lack of dialogic design not only undermines instructional autonomy but also inhibits the development of trust grounded in pedagogical insight (Davis, 2024).

Critical Synthesis

Across the literature, the reported impact of XAI on teaching and learning often lacks theoretical grounding and rigorous evaluation (Table 4). Explanation quality is measured through indirect indicators—trust, satisfaction, or usability—rather than through cognitive, metacognitive, or epistemic gains. Learner-facing explanations frequently serve to justify outputs rather than stimulate critical thinking, while teacher-facing tools remain under-contextualized and procedurally opaque.

These patterns suggest a misalignment between the educational aspirations of XAI and the realities of its implementation. The discourse surrounding "explainability as enhancement" requires critical reorientation. Instead of asking whether XAI increases trust or usability, research must explore whether explanations support deep understanding, epistemic agency, and pedagogical adaptability. The current literature rarely addresses these questions, thereby limiting the transformative potential of XAI in education.

Table 4

Summary of XAI's Impact on Learning and Teaching

Impact Domain	Common Findings	Critical Observations
Learner Outcomes and Engagement	Claims of increased engagement and understanding, often based on short-term or self-report data	Learning rarely measured with rigor; confusion from cognitive overload often unacknowledged
Trust and System Usability	Reported increases in trust and perceived usability of systems	Trust not distinguished from blind acceptance; reflective engagement rarely assessed
Teacher Perceptions and Instructional Use	Teachers appreciated visual insights but expressed skepticism and rarely used explanations for decisions	Explanations static and system-led; few opportunities for teacher inquiry or customization

4.9 Theme 5: challenges and limitations

While Explainable Artificial Intelligence (XAI) continues to gain traction in educational technologies, the existing literature reveals a range of challenges that are often more foundational than technical. These limitations include the interpretability–

performance trade-off, user cognitive burdens, contextual blind spots, and under-theorized pedagogical applications. Collectively, they underscore the conceptual tensions that arise when technical paradigms are imposed on educational contexts without critical adaptation.

Interpretability–Performance Trade-Off

A recurring challenge in the design of educational XAI systems is the perceived conflict between interpretability and predictive performance. Many studies adopt simpler models such as decision trees or linear classifiers for their transparency, despite a reduction in predictive accuracy—particularly for complex tasks such as early dropout prediction or adaptive content recommendation (Agarwal, 2023; Milad & Whiba, 2024). In contrast, studies utilizing deep learning models often rely on post-hoc explanation techniques to retroactively add transparency.

However, this trade-off is rarely scrutinized through a pedagogical lens. Educational technologies are often implemented without considering whether marginal gains in prediction justify sacrificing interpretability, especially in environments where transparency, fairness, and learner trust are pedagogically critical (Nobles, 2025). The tension between accuracy and interpretability is treated as a fixed constraint rather than a design decision informed by contextual values and the needs of diverse learners.

User Overload and Cognitive Barriers

Many XAI systems in education deliver explanations that are overly complex, excessively detailed, or insufficiently scaffolded. These explanations often use technical jargon, abstract visualizations, or extended logic chains, contributing to cognitive overload—especially for novice learners, multilingual users, and those with limited domain knowledge (Gorantla & Devineni, 2024). Instead of clarifying system decisions, such explanations can confuse users or encourage superficial compliance.

Despite growing awareness of these risks, few XAI systems offer differentiated explanation formats that adjust based on user expertise, reading ability, or learning needs (Mehendale, 2022). Moreover, explanations are often presented as static outputs, offering little opportunity for users to question, expand upon, or negotiate meaning through dialogic interaction. This static, unidirectional model undermines critical engagement and hinders the development of learner autonomy and epistemic agency.

Contextual Blind Spots and Cultural Narrowness

A significant limitation in the current literature is its narrow contextual scope. Most empirical studies are conducted in higher education settings within the Global North, often involving students from computer science or engineering disciplines (Hamida et al., 2024). Consequently, pedagogical, cultural, and linguistic differences across global learning contexts are frequently overlooked. There is limited research on how explainability is perceived or operationalized in K–12 education, multilingual classrooms, or non-Western cultural contexts.

Equity concerns are further exacerbated by the lack of inclusive design practices. Few studies consider how learners with disabilities, limited digital literacy, or minority language backgrounds interpret and benefit from explanations (Boppiniti, 2020). Without participatory approaches that center marginalized voices, current XAI systems risk reinforcing the very inequalities they are positioned to address.

Under-theorization of Explanation as a Learning Tool

One of the most critical conceptual limitations lies in the under-theorization of explanation within educational XAI. While explainability is a well-developed concept in computer science, its educational function—as a cognitive and epistemic scaffold—is rarely examined in depth. Explanations are often treated as delivery mechanisms for technical justifications, rather than as tools to support inquiry, conceptual change, or knowledge construction (G et al., 2024).

Most XAI systems fail to draw from established educational theories that address how learners make sense of information, revise misconceptions, or develop metacognitive strategies. As a result, explanations are often instrumentalized to justify system behavior rather than to promote critical engagement or facilitate learning (Maity & Deroy, 2024). This conceptual gap leads to superficial implementations that prioritize algorithmic transparency over pedagogical impact.

Critical Synthesis

The challenges outlined above reflect more than engineering limitations; they highlight a fundamental misalignment between technical design practices and educational imperatives (Table 5). Current XAI applications often adopt a system-centric view of explainability, emphasizing operational clarity over cognitive or pedagogical depth. The result is a fragmented field, shaped by disciplinary silos, that lacks a coherent framework for reconciling interpretability with learning theory and educational justice.

To move forward, explainability must be reimagined not as a mechanism for compliance or system trust but as a dialogic, participatory, and empowering practice. It should be designed to serve diverse learners and educators, embedded in culturally responsive frameworks, and informed by robust theories of learning. Without such a transformation, the full educational promise of XAI will remain constrained by unresolved tensions in its design and deployment.

Table 5

Summary of Challenges and Limitations in Educational XAI

Challenge Area	Description of Challenge	Critical Observations
Interpretability–Performance Trade-Off	Choosing between model transparency and predictive power; rarely justified pedagogically	Accuracy often prioritized over interpretability without justification for educational settings
User Overload and Cognitive Barriers	Explanations cause confusion when dense, technical, or not tailored to user background	Static explanations do not adapt to learner needs; minimal support for questioning or personalization
Contextual Blind Spots and Cultural Narrowness	Most studies conducted in Global North; lacks cultural, linguistic, or equity-oriented sensitivity	Marginalized learners rarely included in design; equity of explainability remains unaddressed
Under-theorization of Explanation as a Learning Tool	Few studies draw on learning theory; explanations not treated as epistemic tools	Explainability seen as a system feature rather than a pedagogical process of dialogue and reflection

4.10 Theme 6: ethical and epistemological considerations

The integration of Explainable Artificial Intelligence (XAI) in education constitutes an ethical and epistemological enterprise that shapes knowledge construction, power relations, and the contours of human agency. While studies acknowledge these dimensions, they often relegate them to cursory or abstract commentary without practical application. This thematic synthesis underscores the implications of such omissions and proposes a redirection toward ethically and epistemically grounded design in educational XAI.

Transparency and Accountability

Transparency is frequently identified as a cornerstone of XAI, often associated with user trust or regulatory compliance (Chaudhry et al., 2022; Gunning & Aha, 2019). However, few studies define transparency in relation to its audience—whether learners, teachers, or developers—or distinguish among its various types: algorithmic, procedural,

or epistemic. As a result, the term risks becoming a vague placeholder that obscures rather than clarifies explanatory value (Meske et al., 2021).

Accountability is even more underdeveloped. The reviewed literature rarely interrogates who bears responsibility when explanations mislead or reinforce inequities. In rare cases, scholars raise concerns about overreliance on predictive dashboards or the ethical risks of partial transparency—such as providing technically accurate but pedagogically misleading explanations (Morandín-Ahuerma, 2024). When accountability is framed solely as legal liability, it undermines efforts to foster responsibility and critical agency within educational systems.

Fairness, Bias, and Inclusion

Despite increasing attention to algorithmic fairness, most studies fail to explore how XAI mitigates or reinforces bias in educational decision-making (Agarwal, 2023; Hamida et al., 2024). Fairness is often treated as a computational issue, abstracted from its sociopolitical roots, with little attention to participatory design or inclusive validation (Maity & Deroy, 2024). The concept of epistemic injustice, which highlights the marginalization of individuals' knowledge and interpretive capacity (Fricker, 2007), is largely absent from discussions—even though it is highly relevant in educational contexts that serve historically disadvantaged populations.

Several systems are deployed without involving the communities they affect, resulting in epistemologies that reflect dominant values about what counts as intelligible or useful knowledge. Models tailored to monolingual, digitally literate, or STEM-oriented users risk alienating those from different cultural or disciplinary backgrounds (Gumabay & Gumabay, 2024). Without inclusive epistemological grounding, XAI tools contribute to a technocratic model of education that privileges efficiency over equity.

Epistemological Positioning of Explanations

A recurring gap in the literature is the limited reflection on the nature of knowledge conveyed through XAI explanations. Most systems reduce explanations to factual or causal accounts, privileging a positivist stance that ignores interpretative, dialogic, or contextual ways of knowing (Gaur et al., 2024; Liu et al., 2024). This framing is particularly problematic in domains such as ethics, humanities, or civic reasoning, where ambiguity and contestation are central.

Moreover, few systems are designed to cultivate epistemic agency—the learner's ability to interrogate, reinterpret, and critique presented knowledge. Explanations

typically position users as passive recipients rather than active participants in meaning-making. This dynamic undermines educational goals centered on critical thinking and reinforces what scholars call "automated epistemologies," wherein certain knowledge forms are codified and others excluded (Feldman-Maggor et al., 2024; Maity & Deroy, 2024).

Critical Synthesis

Ethical and epistemological engagement within educational XAI remains shallow, fragmented, and inconsistently operationalized (Table 6). While transparency and fairness are widely invoked, they are rarely contextualized in relation to learners' needs or instructional environments. Likewise, the epistemic content and implications of AI-generated explanations receive insufficient theorization. As a result, many systems inform without empowering and clarify without inviting challenge.

A new orientation is required—one that treats ethics not as compliance but as care and critique, and explanation not as output but as a dialogic process. Educational XAI should foster reflection, discourse, and inquiry, not merely deliver justifications. Realizing this vision demands participatory design, shared accountability, and theoretical grounding that affirms human dignity and justice.

Table 6

Summary of Ethical and Epistemological Considerations in Educational XAI

Focus Area	Common Observations	Critical Reflections
Transparency and Accountability	Transparency often invoked without clarifying type or audience; accountability rarely addressed	Transparency becomes superficial if not linked to pedagogical or epistemic clarity; liability emphasized over shared responsibility
Fairness, Bias, and Inclusion	Bias mitigation framed in technical terms; no engagement with epistemic injustice or participatory design	Equity overlooked; dominant epistemologies shape system logic and may exclude marginalized perspectives
Epistemological Positioning of Explanations	Explanations treated as factual outputs; little attention to dialogic, situated, or co-constructed understanding	Users positioned as passive recipients; epistemic agency and critique are largely absent in system design

5 DISCUSSION

This systematic review reveals a dynamic yet conceptually fragmented field. While Explainable AI (XAI) is frequently presented as a remedy to the opacity of educational AI systems, its pedagogical potential remains inadequately realized,

inconsistently theorized, and variably implemented across studies (Liu et al., 2024; Türkmen, 2025). The evidence underscores the need for a critical reorientation of XAI in education, one that centers educational theory, ethical rigor, and user agency.

5.1 Incoherence in definitions and purposes of explainability

A persistent issue is the definitional vagueness surrounding "explainability." Studies often celebrate explainability as an inherent good, yet fail to specify its intended audience, cognitive purpose, or evaluative criteria (Kristiawan et al., 2024). This lack of epistemological clarity leads to divergent implementations—from algorithmic feature attributions devoid of educational scaffolding to visualizations that assume high technical fluency among learners (Feldman-Maggor et al., 2024). Without anchoring explainability in pedagogical objectives, design efforts risk superficiality.

Designers frequently rely on vague constructs such as “building trust” or “enhancing transparency” without interrogating their educational validity. As a result, many systems appear explainable but do not substantively support learner understanding, critical engagement, or knowledge construction. Reframing explainability as a relational and context-sensitive process—rather than a static system feature—offers a more productive direction (Chaudhry et al., 2022; Huang, 2023).

5.2 Explainability as instrumental rather than transformative

XAI is often deployed to optimize usability or foster user compliance rather than to deepen learning. Learners are treated as passive recipients of algorithmic output, while teachers receive visual summaries that prompt action rather than inquiry (Alrayes et al., 2024; Gomez et al., 2025). Few systems support dialogic interaction with explanations or enable users to critique underlying assumptions. This instrumental framing aligns XAI with surveillance and performance management rather than epistemic development (Ren & Wu, 2025).

Such implementations conflict with the educational aim of cultivating critical thinkers. Systems focused on algorithmic justification tend to discourage independent reasoning, reinforcing automation over autonomy (Liefooghe & van Maanen, 2022).

Without opportunities for users to interrogate or revise explanations, XAI tools risk becoming didactic rather than reflective.

5.3 The missing pedagogical and ethical foundations

Many XAI applications lack grounding in learning theory. Explanations are rarely designed with consideration for cognitive development, prior knowledge, or disciplinary epistemologies (Abu Bakar & Ismail, 2020; G et al., 2024). Moreover, the field treats fairness, transparency, and accountability as abstract ideals rather than operational design goals. Ethical principles are cited, but seldom translated into interface features, data governance structures, or user rights (Abbas, 2024; Morandín-Ahuerma, 2024).

This technical formalism privileges engineering concerns over human learning needs. It sidelines the diversity of learners and contexts, leading to systems that presume uniformity in comprehension and values (Hamida et al., 2024). To counteract this, XAI design must be informed by educational justice, participatory methods, and sensitivity to epistemic difference (Gumabay & Gumabay, 2024; Maity & Deroy, 2024).

5.4 A reimagined framework for educational explainability

This review proposes a tripartite framework for educational XAI:

- ***Pedagogical Explainability:*** Explanations should support cognitive engagement, conceptual understanding, and reflective inquiry. Designs must be grounded in learning sciences and tailored to developmental stages (Chan & Zary, 2019; Song, 2024).
- ***Epistemic Explainability:*** Explanations should clarify how knowledge is constructed, revealing assumptions and enabling critique. Systems must foster epistemic agency by positioning users as co-constructors of meaning (Gaur et al., 2024; Zhang et al., 2022).
- ***Ethical Explainability:*** Explanations should address the broader implications of AI decisions. This includes transparency about intentions, avenues for contestation, and accountability mechanisms (Choi et al., 2024; Nobles, 2025).

These dimensions are mutually constitutive. A system that is technically transparent yet pedagogically opaque is inadequate. Likewise, ethically informed systems

that lack epistemic depth fall short of educational goals. A holistic approach must integrate these dimensions to create XAI systems that empower rather than marginalize.

5.5 Limitations

While this review offers a comprehensive and critical synthesis of research on explainable artificial intelligence (XAI) in educational technologies, several limitations must be acknowledged. These limitations pertain to the scope of the literature included, the methodological diversity of the reviewed studies, and the evolving nature of the field.

5.6 Scope and selection bias

First, the review was restricted to peer-reviewed empirical studies published in English between 2015 and 2025. While this timeframe captures the most recent and relevant developments in XAI, it may exclude foundational work from earlier years or important insights from non-English-speaking contexts. The exclusion of grey literature, dissertations, and technical reports may have further limited the inclusion of cutting-edge prototypes, practical implementations, or negative findings that are not yet represented in formal academic outlets.

The reliance on six major academic databases (Scopus, Web of Science, IEEE Xplore, ACM Digital Library, ERIC, and SpringerLink) provided a broad but not exhaustive search. Despite efforts to use inclusive search terms and Boolean logic, it is possible that relevant studies—especially those published in interdisciplinary or domain-specific venues (e.g., special education, civic education)—were inadvertently omitted.

5.7 Methodological heterogeneity and evaluation gaps

Second, the methodological heterogeneity of the included studies posed challenges for synthesis. The review included qualitative, quantitative, and mixed-methods research with varying levels of rigor, transparency, and theoretical coherence. Many studies lacked detail on their evaluation metrics for explainability, offered inconsistent definitions of core constructs, or employed limited user engagement methods. These disparities made it difficult to conduct comparative analysis across

studies or to derive generalizable conclusions about the effectiveness of different XAI approaches.

Additionally, many studies relied on short-term experimental designs, pilot tests, or self-reported user metrics, such as satisfaction or perceived trust. The lack of longitudinal designs, learning outcome data, or qualitative insight into user reasoning limited the ability to evaluate the long-term educational impact of XAI systems. Consequently, some of the claims made by the reviewed studies—particularly those regarding improved engagement or understanding—should be interpreted with caution.

5.8 Evolving terminologies and conceptual boundaries

Third, the field of explainable AI in education is still rapidly evolving, with terminologies and conceptual boundaries in flux. Terms such as "interpretability," "transparency," and "explainability" are used interchangeably or with inconsistent meanings, both within and across studies. While this review attempts to clarify and categorize these usages, the lack of consensus in the literature may affect the internal coherence of the thematic synthesis. Furthermore, novel approaches to explainability, including affective or culturally adaptive explanations, may not yet be captured in the published corpus but are likely to emerge in the near future.

5.9 Reflexivity and interpretive framing

Finally, as with all systematic reviews, the process of inclusion, coding, and interpretation was mediated by researcher judgment. Although rigorous procedures were employed for screening and data extraction—including interrater agreement and critical appraisal protocols—the synthesis inevitably reflects the interpretive lens of the review authors. The decision to adopt a critical perspective, emphasizing pedagogical, epistemic, and ethical dimensions, may foreground certain issues while placing less emphasis on technical advancements or system performance metrics. This positioning is intentional, but it also shapes the contours of the review.

6 CONCLUSION AND RECOMMENDATIONS

This systematic review has examined the current state of research on Explainable Artificial Intelligence (XAI) in Educational Technologies through a critical and multidisciplinary lens. The analysis of empirical studies published between 2015 and 2025 reveals that while XAI has become a central concern in the development of AI-powered learning systems, its implementation remains fragmented, under-theorized, and often misaligned with core educational values.

Many reviewed studies adopt narrow or vague definitions of explainability, emphasizing technical transparency or user trust without addressing the cognitive, pedagogical, or ethical functions of explanation. Although some progress has been made in developing user-friendly interfaces and integrating interpretable models into intelligent tutoring systems and dashboards, these innovations frequently lack grounding in learning theory, overlook issues of equity and inclusion, and fail to support learners' and teachers' epistemic agency.

Explainability, in its current operationalization, is too often treated as a feature to be appended post hoc rather than as a pedagogical process to be co-designed with users. Learners are positioned as recipients rather than participants in meaning-making, and teachers are rarely involved in shaping or evaluating the explanations provided. As a result, educational XAI systems risk becoming tools of compliance rather than vehicles for inquiry, critical thinking, or equitable learning.

6.1 Recommendations for research, design, and policy

In light of these findings, the following recommendations are proposed to guide future research, development, and implementation of XAI in educational contexts:

- ***Theorize Explainability as a Pedagogical Construct:*** Future studies must move beyond surface-level interpretations of explainability and engage with educational theories of learning, cognition, and knowledge construction. Explanations should not only clarify system behavior but also scaffold student understanding, promote metacognitive reflection, and support concept transfer. Research should evaluate explanation quality based on educational outcomes, not just user trust or satisfaction.

- ***Design for Epistemic and Ethical Engagement:*** Explainable systems must be designed to support epistemic agency—the ability of users to question, critique, and co-construct explanations. This requires participatory design processes that include diverse stakeholders, particularly marginalized learners and educators from underrepresented settings. Ethical principles such as fairness, transparency, and accountability should be embedded not only in the model logic but also in how explanations are framed, delivered, and used.
- ***Shift from Explanation as Output to Explanation as Interaction:*** Explanations should not be static disclosures but dynamic interactions. Systems should allow for follow-up questions, customization of explanation depth or modality, and integration of learner feedback. This interactive orientation aligns with dialogic pedagogy and fosters a culture of inquiry rather than passive acceptance.
- ***Evaluate Learning-Centered Outcomes:*** Future empirical research should prioritize longitudinal, mixed-method, and contextually rich evaluations of XAI systems. Metrics should include conceptual understanding, reasoning quality, reflective engagement, and learning gains—not only behavioral proxies or affective reactions. Evaluation frameworks should also assess the interpretability of explanations across user profiles, disciplines, and cultural-linguistic settings.
- ***Develop Cross-Disciplinary Frameworks:*** Progress in educational XAI requires collaboration across fields: AI, education, learning sciences, ethics, human-computer interaction, and critical data studies. Researchers must work to build integrative frameworks that can guide design decisions, empirical inquiry, and policy development in ways that respect both technological capacity and human dignity.

In conclusion, the promise of XAI in education lies not in its capacity to explain algorithms, but in its potential to empower learners and educators to engage more deeply, critically, and equitably with knowledge. Realizing this promise demands a fundamental rethinking of what explainability means, what it is for, and for whom it is designed.

REFERENCES

Abbas, H. (2024). Transforming education: The role of artificial intelligence. *Studies in Engineering and Exact Sciences*, *5*(3), Article e12579. https://doi.org/10.54021/seesv5n3-041

Abu Bakar, M. A., & Ismail, N. (2020). Exploring metacognitive regulation and students' interaction in mathematics learning: An analysis of needs to enhance students' mastery. *Humanities & Social Sciences Reviews*, *8*(2), 67–74. https://doi.org/10.18510/hssr.2020.82e07

Agarwal, A. (2023). Exploring the landscape of explainable artificial intelligence: Benefits, challenges, and future perspectives. *International Journal of Advanced Research*, *11*(12), 1042–1046. https://doi.org/10.21474/IJAR01/18074

Allahverdiyeva, N. (2024). The role, importance and application of artificial intelligence in the creation of digital education. *SCIENTIFIC WORK*, *18*(2), 120–125. https://doi.org/10.36719/2663-4619/99/120-125

Alrayes, A., Henari, T. F., & Ahmed, D. A. (2024). ChatGPT in education – Understanding the Bahraini academics perspective. *Electronic Journal of e-Learning*, *22*(2), 112–134. https://doi.org/10.34190/ejel.22.2.3250

Arianto, F. (2021). Metacognitive strategy and science problem-solving abilities in elementary school students. *International Journal of Social Science and Human Research*, *4*(9). https://doi.org/10.47191/ijsshr/v4-i9-42

Boppiniti, S. (2020). A survey on explainable AI: Techniques and challenges. *International Journal of Innovative Engineering Research and Technology*, *7*(3), 57–66. https://doi.org/10.26662/ijiert.v7i3.pp57-66

Braaten, M., & Windschitl, M. (2011). Working toward a stronger conceptualization of scientific explanation for science education. *Science Education*, *95*(4), 639–669. https://doi.org/10.1002/sce.20449

Chan, K. S., & Zary, N. (2019). Applications and challenges of implementing artificial intelligence in medical education: Integrative review. *JMIR Medical Education*, *5*(1), Article e13930. https://doi.org/10.2196/13930

Chaudhry, M. A., Cukurova, M., & Luckin, R. (2022). A transparency index framework for AI in education. In M. M. Rodrigo, N. Matsuda, A. I. Cristea, & V. Dimitrova (Eds.). *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners' and Doctoral Consortium* (pp. 195–198). Springer International Publishing. https://doi.org/10.1007/978-3-031-11647-6_33

Choi, J.-I., Yang, E., & Goo, E.-H. (2024). The effects of an ethics education program on artificial intelligence among middle school students: Analysis of perception and attitude changes. *Applied Sciences*, *14*(4), 1588. https://doi.org/10.3390/app14041588

Darwish, A. (2022). Explainable artificial intelligence: A new era of artificial intelligence. *Digital Technologies Research and Applications*, *1*(1), 1. https://doi.org/10.54963/dtra.v1i1.29

Darwish, D. (2025). Classroom assessment and evaluation with artificial intelligence. In *Deep Science Publishing* (pp. 34–43). Deep Science Publishing. https://doi.org/10.70593/978-81-984306-7-0_3

Davis, R. O. (2024). Korean in-service teachers' perceptions of implementing artificial intelligence (AI) education for teaching in schools and their AI teacher training

programs. *International Journal of Information and Education Technology*, *14*(2), 214–219. https://doi.org/10.18178/ijiet.2024.14.2.2042

Ewals, L. J. S., Heesterbeek, L. J. J., Yu, B., van der Wulp, K., Mavroeidis, D., Funk, M., Snijders, C. C. P., Jacobs, I., Nederend, J., Pluyter, J. R., & e/MTIC Oncology group. (2024). The impact of expectation management and model transparency on radiologists' trust and utilization of AI recommendations for lung nodule assessment on computed tomography: Simulated use study. *JMIR AI*, *3*, Article e52211. https://doi.org/10.2196/52211

Feldman-Maggor, Y., Nazaretsky, T., & Alexandron, G. (2024). Explainable AI for unsupervised machine learning: A proposed scheme applied to a case study with science teachers. In Proceedings of the 16th International Conference on Computer Supported Education (pp. 436–444). SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0012687000003693

Fricker, M. (2007). *Epistemic injustice: Power and the ethics of knowing*. Oxford University Press.

Gaur, A. S., Sharan, H. O., & Kumar, R. (2024). AI in education. In R. Kumar, A. Joshi, H. O. Sharan, S.-L. Peng, & C. R. Dudhagara (Eds.), *Artificial intelligence applications* (pp. 39–54). IGI Global. https://doi.org/10.4018/979-8-3693-2964-1.ch003

Göçen, A., & Aydemir, F. (2020). Artificial intelligence in education and schools. *Research on Education and Media*, *12*(1), 13–21. https://doi.org/10.2478/rem-2020-0003

Gomez, M. J., Armada Sánchez, Á., Albaladejo-González, M., García Clemente, F. J., & Ruipérez-Valiente, J. A. (2025). Utilising explainable AI to enhance real-time student performance prediction in educational serious games. *Expert Systems*, *42*(3). https://doi.org/10.1111/exsy.70008

Gorantla, B., & Devineni, S. (2024). Evaluation of explainable artificial intelligence using TOPSIS method. *Computer Science, Engineering and Technology*, *2*(2), 10–20. https://doi.org/10.46632/cset/2/2/2

Guan, H. (2023). Advantages and challenges of using artificial intelligence in primary and secondary school education. *Journal of Education, Humanities and Social Sciences*, *22*, 377–383. https://doi.org/10.54097/ehss.v22i.12469

Gumabay, D. C. A. N., & Gumabay, D. M. V. N. (2024). Opportunities and challenges for information technology and business educators in implementing generative artificial intelligence in instruction. *International Journal of Management and Humanities*, *11*(4), 1–7. https://doi.org/10.35940/ijmh.D1769.11041224

Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Hamida, S. U., Chowdhury, M. J. M., Chakraborty, N. R., Biswas, K., & Sami, S. K. (2024). Exploring the landscape of explainable artificial intelligence (XAI): A systematic review of techniques and applications. *Big Data and Cognitive Computing*, *8*(11), 149. https://doi.org/10.3390/bdcc8110149

Hoya, F., Mah, D., Prilop, C., Jacobsen, L., & Weber, K. (2024). Preservice teachers' AI usage: The effects of perceived usefulness, subjective norm, behavioral intention, and self-efficacy. *OSF Preprints*. https://doi.org/10.31219/osf.io/284gk

Huang, J. (2023). Engineering ChatGPT prompts for EFL writing classes. *International Journal of TESOL Studies*. https://doi.org/10.58304/ijts.20230405

Kaharuddin, K., Ahmad, D., Mardiana, M., Latif, I., Arafah, B., & Suryadi, R. (2024). Defining the role of artificial intelligence in improving English writing skills among Indonesian students. *Journal of Language Teaching and Research, 15*(2), 568–678. https://doi.org/10.17507/jltr.1502.25

Karpouzis, K. (2024). Explainable AI for intelligent tutoring systems. In M. Farmanbar, M. Tzamtzi, A. K. Verma, & A. Chakravorty (Eds.), *Frontiers of artificial intelligence, ethics, and multidisciplinary applications* (pp. 59–70). Springer Nature. https://doi.org/10.1007/978-981-99-9836-4_6

Kim, J., Ham, Y., & Lee, S.-S. (2024). Differences in student-AI interaction process on a drawing task: Focusing on students' attitude towards AI and the level of drawing skills. *Australasian Journal of Educational Technology*. https://doi.org/10.14742/ajet.8859

Kristiawan, D., Bashar, K., & Pradana, D. A. (2024). Artificial intelligence in English language learning: A systematic review of AI tools, applications, and pedagogical outcomes. *The Art of Teaching English as a Foreign Language, 5*(2), 207–218. https://doi.org/10.36663/tatefl.v5i2.912

Li, Z. (2023). The significance of educational application of artificial intelligence and its current state in China. *Science Insights Education Frontiers, 16*(2), 2589–2597. https://doi.org/10.15354/sief.23.re215

Liefooghe, B., & van Maanen, L. (2022). Three levels at which the user's cognition can be represented in artificial intelligence. *Frontiers in Artificial Intelligence, 5*, Article 1092053. https://doi.org/10.3389/frai.2022.1092053

Lin, M. P.-C., Liu, A. L., Poitras, E., Chang, M., & Chang, D. H. (2024). An exploratory study on the efficacy and inclusivity of AI technologies in diverse learning environments. *Sustainability, 16*(20), 8992. https://doi.org/10.3390/su16208992

Liu, Q., Pinto, J. D., & Paquette, L. (2024). Applications of explainable AI (XAI) in education. In D. Kourkoulou, A.-O. Tzirides, B. Cope, & M. Kalantzis (Eds.), *Trust and inclusion in AI-mediated education* (pp. 93–109). Springer. https://doi.org/10.1007/978-3-031-64487-0_5

Liulka, V., Savenkova, O., & Dedukhno, A. (2024). The peculiarities of using artificial intelligence in teaching foreign languages in higher education institutions in Ukraine. *Humanities Science Current Issues, 2*(73), 195–201. https://doi.org/10.24919/2308-4863/73-2-30

Madi, I. A. E., Redjdal, A., Bouaud, J., & Séroussi, B. (2024). Exploring explainable AI techniques for text classification in healthcare: A scoping review. *Studies in Health Technology and Informatics, 316*, 846–850. https://doi.org/10.3233/SHTI240544

Maity, S., & Deroy, A. (2024). *Human-centric explainable AI in education*. https://doi.org/10.35542/osf.io/k5u9b

Marrone, R., Zamecnik, A., Joksimović, S., Johnson, J., & De Laat, M. (2024). Understanding student perceptions of artificial intelligence as a teammate. *Technology, Knowledge and Learning*. https://doi.org/10.1007/s10758-024-09780-z

Mehendale, P. (2022). Bridging the gap: Enhancing trust and transparency in machine learning with explainable AI. *Journal of Artificial Intelligence & Cloud Computing*, 1–4. https://doi.org/10.47363/JAICC/2022(1)E123

Meske, C., Abedin, B., Junglas, I., & Rabhi, F. (2021). Introduction to the Minitrack on explainable artificial intelligence (XAI). *Proceedings of the Annual Hawaii International Conference on System Sciences*. https://doi.org/10.24251/HICSS.2021.153

Milad, A., & Whiba, M. (2024). Exploring explainable artificial intelligence technologies: Approaches, challenges, and applications. *International Science and Technology Journal*, *34*(1), 1–21. https://doi.org/10.62341/amia8430

Morandín-Ahuerma, F. (2024). A critical analysis of the European Union's considerations on the ethical use of artificial intelligence in education. *OSF Preprints*. https://doi.org/10.31219/osf.io/escqj

Najdawi, M. H. A., Shwedeh, F., Abdelmoghies, M. M., Kitana, A., & Ali, A. (2024). Applying artificial intelligence in predicting educational excellence in higher education institutions: A case study in Jordanian universities. *Edelweiss Applied Science and Technology*, *8*(6), 7273–7289. https://doi.org/10.55214/25768484.v8i6.3579

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. *Proceedings of the Association for Information Science and Technology*, *58*(1), 504–509. https://doi.org/10.1002/pra2.487

Nobles, C. (2025). Human factors engineering in explainable AI: Putting people first. *International Conference on Cyber Warfare and Security*, *20*(1), 313–322. https://doi.org/10.34190/iccws.20.1.3348

Nurjanah, A., Salsabila, I. N., Azzahra, A., Rahayu, R., & Marlina, N. (2024). Artificial intelligence (AI) usage in today's teaching and learning process: A review. *Syntax Idea*, *6*(3), 1517–1523. https://doi.org/10.46799/syntax-idea.v6i3.3126

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., . . . Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71:contentReference[oaicite:1]{index=1}

Ren, X., & Wu, M. L. (2025). Examining teaching competencies and challenges while integrating artificial intelligence in higher education. *TechTrends*. https://doi.org/10.1007/s11528-025-01055-3

Robinson, C. L., D'Souza, R. S., Yazdi, C., Diejomaoh, E. M., Schatman, M. E., Emerick, T., & Orhurhu, V. (2024). Reviewing the potential role of artificial intelligence in delivering personalized and interactive pain medicine education for chronic pain

patients. *Journal of Pain Research, 17*, 923–929. https://doi.org/10.2147/JPR.S439452

Sadvakassova, A., Kydyrbekova, A., & Chetin, O. (2024). Using of virtual reality and artificial intelligence in education: Literature review. *"Bilim" Scientific and Pedagogical Jornal, 110*(3), 10–18. https://doi.org/10.59941/2960-0642-2024-3-10-18

Sandoval, W. (2014). Science Education's Need for a Theory of Epistemological Development. *Science Education, 98*(3), 383–387. https://doi.org/10.1002/sce.21107

Song, D. (2024). Artificial intelligence for human learning: A review of machine learning techniques used in education research and a suggestion of a learning design model. *American Journal of Education and Learning, 9*(1), 1–21. https://doi.org/10.55284/ajel.v9i1.1024

Susantini, E., Puspitawati, R. P., Raharjo, R., & Suaidah, H. L. (2021). E-book of metacognitive learning strategies: Design and implementation to activate student's self-regulation. *Research and Practice in Technology Enhanced Learning, 16*(1). https://doi.org/10.1186/s41039-021-00161-z

Türkmen, G. (2025). The review of studies on explainable artificial intelligence in educational research. *Journal of Educational Computing Research, 63*(2), 277–310. https://doi.org/10.1177/07356331241310915

Yildirim, Y., & Celepcikay, A. (2021). Artificial intelligence and machine learning applications in education. *Eurasian Journal of Higher Education, 2*(4), 1–11. https://doi.org/10.31039/ejohe.2021.4.49

Yildiz Durak, H., Eğin, F., & Onan, A. (2025). A comparison of human-written versus AI -Generated text in discussions at educational settings: Investigating features for ChatGPT, Gemini and BingAI. *European Journal of Education, 60*(1). https://doi.org/10.1111/ejed.70014

Zhai, G., Guo, K., & Li, S. (2023). Research on innovation of ideological and political education in universities based on artificial intelligence. In K. M. Salleh, M. I. R. M. Mohammad Ismail, & L. Yuan (Eds.). *Proceedings of the 2023 International Conference on Applied Psychology and Modern Education (ICAPME 2023)* (pp. 27–32). Atlantis Press sarl. https://doi.org/10.2991/978-2-38476-158-6_5

Zhang, H., Lee, I., Ali, S., DiPaola, D., Cheng, Y., & Breazeal, C. (2022). Integrating ethics and career futures with technical learning to promote AI literacy for middle school students: An exploratory study. *International Journal of Artificial Intelligence in Education, 33*(2), 1–35. https://doi.org/10.1007/s40593-022-00293-3

Authors' Contribution

Both authors contributed equally to the development of this article.

Data availability

All datasets relevant to this study's findings are fully available within the article.

How to cite this article (APA):

Meylani, R. (2025). RETHINKING EXPLAINABILITY IN EDUCATIONAL ARTIFICIAL INTELLIGENCE: A CRITICAL SYSTEMATIC REVIEW OF MODELS, APPLICATIONS, AND ETHICAL DIMENSIONS. *Veredas Do Direito*, 22(3), e223553. https://doi.org/10.18623/rvd.v22.n3.3553