

EXPLAINABLE ARTIFICIAL INTELLIGENCE IN ONLINE LEARNING: A REVIEW OF CURRENT APPLICATIONS AND FUTURE RESEARCH DIRECTION

Ghisna Madina

State Islamic University of K.H. Abdurrahman Wahid Pekalongan, Indonesia

E-mail: madinaghisna@gmail.com

Copyright © 2026 The Author



This is an open access article

Under the Creative Commons Attribution Share Alike 4.0 International License

Abstract

Artificial Intelligence (AI) has become increasingly embedded in online learning environments, supporting personalized learning, learning analytics, automated feedback, and early identification of at-risk students. Many AI models are like a "black box," which means it's hard to see how they make decisions. This can make it difficult to understand them and trust them. Because of this, there is an increasing interest in Explainable Artificial Intelligence (XAI). This study reviews existing research about explainable artificial intelligence (XAI) in online learning using the PRISMA 2020 method. A total of 2,136 records were retrieved from six major academic databases, of which 114 studies published between 2018 and 2026 met the inclusion criteria and were analyzed thematically across three domains: current applications, persistent challenges, and future research directions. Findings show that post-hoc explanation techniques, particularly SHAP and LIME, alongside transformer-based architectures and interactive explainable dashboards, are the dominant approaches for supporting student performance prediction, engagement analysis, and adaptive learning. Despite this progress, the review identifies recurring challenges: pedagogical usability, the lack of human-centered evaluation standards, limited model generalizability across educational contexts, and unresolved ethical concerns regarding fairness and accountability. The review argues that future XAI research must move beyond technical explainability toward human-centered design that integrates pedagogical theory, personalized learning analytics, and standardized evaluation frameworks. This synthesis provides a timely, evidence-based foundation for researchers and practitioners seeking to design more transparent, trustworthy, and pedagogically meaningful AI systems for online education.

Keywords: Explainable Artificial Intelligence; Online Learning; Learning Analytics; Human-Centered AI; Systematic Literature Review

1. Introduction

The way we learn has changed a lot because many people are using online learning platforms (Türkmen, 2024). The creation of Learning Management Systems (LMS), Intelligent Tutoring Systems (ITS), and Massive Open Online Courses (MOOCs) has led schools and colleges to use Artificial Intelligence (AI) more often. This helps them give personalized learning, keep track of how students are involved, and offer automatic feedback (Lin et al., 2023; Ogata et al., 2024). This adoption has accelerated further as AI-driven analytics become embedded in everyday online learning infrastructure, positioning AI not merely as a supplementary tool but as a core component of how institutions monitor, predict, and respond to student needs (Lin et al., 2023). Furthermore, the growing role of AI predictive models in making impactful decisions, such as identifying students at risk of learning difficulties or recommending learning interventions, also presents new challenges. One key challenge is the low transparency of AI models, which hinders user trust and wider adoption in educational settings (Nagaraj et al., 2023; Seo et al., 2021). In many cases, both students and educators receive predictions without understanding how the system makes them. This situation limits their ability to leverage AI as a partner that truly supports the learning process (Er, 2023).

Empirical evidence further reveals that the "black box" nature of deep learning models and ensemble algorithms makes it difficult to interpret prediction results (Meylani, 2025; Saqr & López-Pernas, 2024). As a result, educators often receive warnings about student learning risks without adequate pedagogical explanations. This situation diminishes the practical value of the recommendations and raises questions about the reliability, fairness, and accountability of AI systems in educational contexts (Kumawat & Shaktawat, 2025; Seo et al., 2021). Therefore, XAI has emerged as an increasingly

important research area. XAI seeks to bridge the gap between the technical capabilities of AI models and users' need for easily understandable explanations, enabling decisions that support learner-centered learning practices (Nagaraj et al., 2023).

Several recent reviews have begun to map the landscape of XAI in educational settings. For instance, Gunasekara and Saarela (2025) synthesized explainable learning analytics and educational data mining techniques, while Altukhi and Pradhan (2026) reviewed the broader challenges and advancements of integrating XAI into educational systems. While these contributions have advanced understanding of specific dimensions of XAI in education, few have specifically focused on the online and virtual learning ecosystem encompassing LMS, ITS, and MOOC environments or systematically organized findings across the three interconnected dimensions of current applications, implementation challenges, and future research directions using a formal PRISMA-based protocol. This gap is particularly pressing given the rapid growth in XAI-related publications between 2018 and 2026, which calls for an updated, methodologically rigorous synthesis to guide both researchers and practitioners.

To address this gap, the present review is guided by three research questions concerning the XAI techniques currently applied in online learning, the technical and pedagogical challenges encountered in their implementation, and the extent to which existing approaches meet educators' practical needs. Through this framework, the review argues that challenges in implementing XAI, such as maintaining explanation quality across large and heterogeneous learner datasets, ensuring algorithmic fairness across diverse student populations, and reducing the cognitive load imposed by AI-generated explanations, are not merely technical issues but foundational concerns for the effective, transparent, and user-oriented integration of AI into online learning ecosystems. By answering these questions, this review contributes a timely, this is a summary based on facts that explains both the latest technology in XAI (Explainable Artificial Intelligence) and the teaching and ethical issues that are important for using XAI responsibly. The rest of this article is structured like this: the Methodology section explains the PRISMA review method; the Results and Discussion section shows the findings in three main areas; and the Conclusion wraps up the main points and suggests ideas for future research.

2. Literature Review

Explainable Artificial Intelligence (XAI) is a way to make it easier for people to understand how AI systems make decisions. It helps to clear up the confusion that often comes from complicated AI models. Barredo Arrieta and others In 2020, researchers sorted XAI (explainable artificial intelligence) methods into two main categories: those that work with specific models and those that can be used with any model. They also divided them based on whether the explanations are built into the model itself or created after the model has already been trained. Among methods used after a model is built, Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are the most popular ways to explain how models make decisions, owing to their ability to quantify feature-level contributions to a model's predictions without requiring access to its internal architecture (Barredo Arrieta et al., 2020).

In parallel, online learning environments have evolved into complex, technology-mediated ecosystems encompassing LMS, Intelligent Tutoring Systems (ITS), and Massive Open Online Courses (MOOCs), each generating substantial behavioral and performance data suited to AI-driven analysis (Lin et al., 2023; Ogata et al., 2024; Türkmen, 2024). Within these environments, AI models are increasingly tasked with high-stakes functions such as predicting learning outcomes and identifying at-risk students, functions where the absence of transparent reasoning can directly undermine both user trust and pedagogical utility (Nagaraj et al., 2023). This intersection between the technical demands of explainability and the practical, human-centered needs of educators and learners forms the conceptual foundation upon which the present systematic review is built.

These concepts establish the theoretical foundation for examining how explainability has been implemented across online learning environments. Guided by this conceptual background, the present study systematically reviews the recent literature to identify current applications, implementation challenges, and future research directions for XAI in online learning.

3. Research Methods

3.1. Research Design

To get a complete understanding of how XAI has grown in online learning, this study used a careful review of existing research based on PRISMA guidelines (Crompton & Burke, 2023; Rahmani et al., 2024) This method was chosen because it creates a clear and consistent way to find, choose, and combine studies. This helps reduce bias and personal opinions in the review.

3.2. Research Question

This review was structured to answer the following three research questions:

RQ1. What XAI techniques and machine learning architectures are currently implemented in online learning environments?

RQ2. What technical, conceptual, and pedagogical challenges arise in the implementation and evaluation of XAI in online learning?

RQ3. To what extent are existing XAI approaches able to meet the practical needs of educators, and how can future research bridge the gap between the predictive accuracy of AI models and their usefulness in pedagogical decision-making?

3.3. Data Sources and Search Strategy

We searched for information in six important academic websites: IEEE Xplore, ACM Digital Library, SpringerLink, Web of Science, ScienceDirect, and Scopus. To ensure no important research was missed, the search process was also validated using Google Scholar and ERIC (Rahmani et al., 2024). Literature published between 2018 and 2026 was selected to ensure this review represents the latest developments in XAI, including the emergence of modern post-hoc explanation techniques and transformer-based models.

The same Boolean search string was consistently applied across all six databases, with only minor syntax adjustments to accommodate each platform's search operators. The search strategy used a combination of keywords related to XAI and online learning as follows:

(XAI OR "explainable AI" OR "interpretable machine learning" OR "SHAP" OR "LIME") AND ("online learning" OR "MOOC" OR "LMS" OR "intelligent tutoring systems" OR "learning analytics")

This uniform search strategy was verified using bibliometric metadata to ensure broad, consistent coverage of the literature, particularly in reputable journals and conference proceedings in the fields of educational technology and AI. This step was taken to reduce the possibility of selection bias during the article identification process (Bond et al., 2020; Lagisz et al., 2025).

Table 1. Strategic Database Search Matrix and Primary Article Retrieval Yields

Academic Source/Database	Primary Strategy / Filter Adjusted Focus	Initial Yield	Results	Final Retained Papers	Retention Rate
IEEE Xplore	Focus on hardware-integrated LMS, embedded intelligence, and local XAI.	314		18	5.73%
ACM Digital Library	Focus on human-computer interaction (HCI) and student-centered dashboards.	245		14	5.71%
SpringerLink	Focus on textbook theory extensions and automated ITS frameworks.	412		22	5.34%
Web of Science	Focus on interdisciplinary educational policy and systematic scaling trials.	289		16	5.54%
ScienceDirect	Focus on high-fidelity deep sequence models and operational platforms.	356		19	5.34%
Scopus	Broad cross-disciplinary analytics validation indices.	520		25	4.81%
Total	—	2,136		114	5.34%

Source: Synthesized by the author(s) from reviewed literature (2026).

The systematic search identified 2,136 records across six major academic databases. Scopus returned the largest number of initial records (n = 520), followed by SpringerLink (n = 412) and

ScienceDirect (n = 356). After applying the predefined screening criteria, 114 studies were retained for further qualitative assessment, representing an overall retention rate of approximately 5.3%.

3.4. Inclusion and Exclusion Criteria

The collected articles were checked to see if they met certain rules for including or excluding them. A study was included only if it met all of these requirements:

- 1) Discusses online or hybrid learning environments, such as MOOCs, LMS, ITS, personalized online learning, or virtual classrooms.
- 2) Explicitly implements, develops, or evaluates XAI methods.
- 3) Is published in respected international journals or important conference papers.

Conversely, a study was excluded if it: (a) focused solely on conventional face-to-face learning, (b) discussed the general application of AI without the context of learning analytics, or (c) only offered theoretical AI concepts without implementation or evaluation in educational contexts (Rahmani et al., 2024). Empirical research linking the use of XAI to personalized learning, student learning outcomes, or user trust in AI systems was prioritized in the selection process (Suhail & Rajakumari, 2025; Türkmen, 2024).

3.5. Data Extraction and Coding Scheme

Data from each article that met the criteria were then extracted using a thematic coding approach (Crompton & Burke, 2023). The coding process was carried out systematically into three main themes. The first theme was Current Applications, which encompassed algorithms, machine learning models, XAI techniques, and their implementation contexts. The second theme is Challenges, which identifies various technical, pedagogical, ethical, and regulatory obstacles encountered in the implementation of XAI. The third theme is Future Research Directions, which summarizes research recommendations and opportunities for future XAI development (Ilić et al., 2023).

This coding approach enables a consistent, structured literature synthesis, yielding review results that not only describe the current state of XAI in online learning but also provide a deeper understanding of its strengths, limitations, and future development directions (Bond et al., 2024; Yang et al., 2023).

All screening, selection, and data extraction procedures in this review were conducted by a single reviewer using the predetermined criteria described above. The way we chose studies followed the PRISMA 2020 guidelines (Page et al., 2021) to make sure everything was clear and done correctly in the review. Figure 1 shows the entire process of finding databases, removing duplicates, checking for eligibility, reviewing full texts, and finally including the studies that meet the criteria.

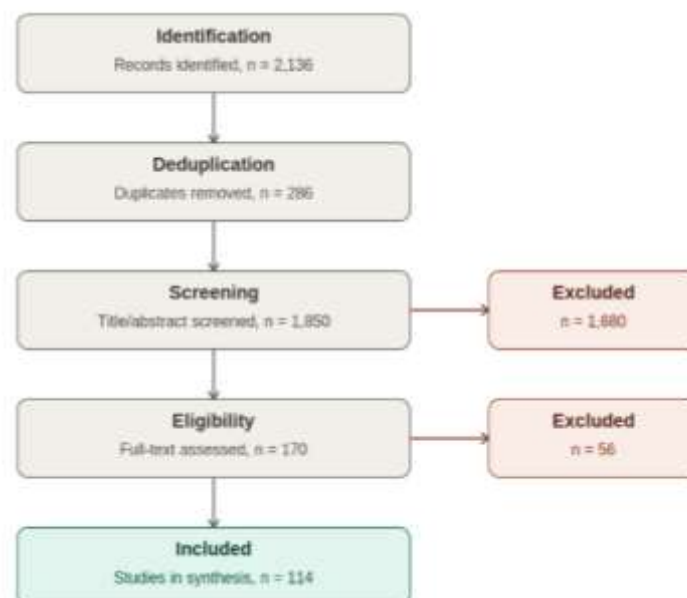


Figure 1. PRISMA 2020 Flow Diagram

Source: Synthesized by the author(s) from reviewed literature (2026).

4. Results and Discussion

4.1. Current Applications of XAI in Online Learning

To address RQ1, the reviewed studies were analyzed by application domains, AI models, explainability techniques, and educational purposes reported in the literature. The findings indicate that

XAI has been implemented across a wide range of online learning environments, particularly LMS, MOOCs, and ITS.

Recent literature shows that research on XAI in online learning is growing, with a focus on increasing the transparency of predictive models without sacrificing predictive performance (Türkmen, 2024). Among the various approaches used, model-agnostic post-hoc methods like Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) are popular tools. They can explain how complex models make decisions without needing to change the model itself (Er, 2023; Türkmen, 2024).

At the prediction stage, tree-based ensemble models, particularly Random Forest, are among the most frequently used algorithms in virtual learning environments and MOOCs. These models are used to support early warning systems by linking the likelihood of students experiencing risk, such as declining performance or dropping out, with various indicators of learning engagement, including video viewing frequency, clickstream patterns, and interaction activity during the learning process (Er, 2023; Murtaza et al., 2026).

Subsequent developments have shown the increasing use of models capable of explaining changes in learning behavior longitudinally. Models based on recurrent neural networks (RNNs) and transformer architectures are increasingly being utilized because they better capture the dynamics of student engagement throughout the learning period than conventional approaches (Akello et al., 2026). Furthermore, hybrid models combining Long Short-Term Memory (LSTM), Transformers, and attention mechanisms have demonstrated superior ability to identify factors that change over time, such as fluctuations in student participation and study time patterns related to academic risk (Ofori & Dake, 2025).

On the implementation side, the explanatory capabilities generated by XAI models are increasingly being integrated into various learning platforms through analytics dashboards and adaptive learning systems. XAI-based dashboards present information on the contribution of each feature, such as clickstream data and learning activity, so that students and educators can understand the factors that influence prediction results and use them as a basis for reflection and self-regulation in the learning process (Afzaal et al., 2021). Furthermore, the principle of transparency is also beginning to be applied to conversational AI, ITS, and various AI-based learning assistants to increase user trust, clarify the rationale behind recommendations, and reduce concerns about the accountability of automated decisions (Kim et al., 2020; Lin et al., 2023; Seo et al., 2021).

Table 2 summarizes these findings across six major application areas, highlighting the consistent dominance of SHAP and LIME as explanation techniques regardless of the underlying model architecture, while also showing a clear progression from static prediction tasks toward temporally aware and interactive explanation formats.

Table 2. Synthesis of Current Applications of XAI in Online Learning

Application Area	AI Models	XAI Techniques	Educational Purpose	Representative Studies
Student Performance Prediction	Random Forest, XGBoost, Ensemble Tree	SHAP, LIME	Predict academic achievement and identify influential learning behaviors	Adnan et al. (2022); Vultureanu-Albiși & Bădică (2021); Shastri et al. (2025)
Dropout Prediction	Random Forest, Ensemble Learning	SHAP, LIME	Early identification of at-risk learners	Er (2023); Ajayi (2025); Nti & Ramanayake (2026)
Student Engagement Analysis	Transformer, LSTM-Transformer	SHAP, Attention Mechanism	Analyze engagement trajectories and temporal learning behaviors	Hidayati et al. (2024); Akello et al. (2026); Ofori & Dake (2025)
Personalized Learning	ITS, Recommendation Models	Explainable Recommendation, SHAP	Deliver adaptive learning support	Ogata et al. (2024); Takami et al. (2023); Suhail &

				Rajakumari (2025)
Learning Analytics Dashboard	Machine Learning	SHAP, Counterfactual Explanation	Support self- regulated learning and instructor decision making	Afzaal et al. (2021); Sušnjak et al. (2022); Ramaswami et al. (2023)
AI Teaching Assistant	Conversational AI	Explainable Feedback	Improve learner trust and instructor interaction	Kim et al. (2020); Seo et al. (2021)

Source: Synthesized by the author(s) from reviewed literature (2026).

Overall, these findings indicate that current XAI applications primarily emphasize prediction accuracy and post-hoc explainability, providing the foundation for examining the conceptual and practical challenges discussed in the next section.

4.2. Conceptual and Technical Challenges

To answer RQ2, the selected studies were synthesized to identify recurring conceptual, technical, pedagogical, and ethical challenges associated with implementing XAI in online learning. Despite the rapid development of XAI in learning, its implementation still faces various conceptual, technical, and methodological challenges. Various studies have shown that the success of XAI is determined not only by the model's ability to generate explanations but also by the extent to which these explanations can be understood, trusted, and utilized in learning decision-making.

4.2.1 The Gap Between Transparency and Pedagogical Usefulness

One major challenge is the gap between algorithm transparency and its usefulness in learning practice. While various XAI methods can explain how the model generates a prediction, these explanations often do not provide information that truly helps educators determine appropriate learning steps (Saqr & López-Pernas, 2024). Most XAI methods still present results in the form of feature importance or statistical weights, which are difficult for non-technical users, including teachers and students, to understand (Meylani, 2025). Thus, algorithm transparency does not automatically translate into pedagogical value. For example, a SHAP visualization can explain why a student is considered at risk, but it does not necessarily provide recommendations that educators can implement to support that student's learning process.

4.2.2 Methodology and Evaluation Challenges

Another issue frequently encountered in the literature is the lack of an evaluation approach capable of comprehensively measuring the effectiveness of XAI. Most research still focuses on evaluations oriented toward technical model performance, such as prediction accuracy or the quality of algorithmic explanations, while evaluations involving real users are still very limited (Brasse et al., 2023; Nauta et al., 2023). As a result, it remains unclear whether the resulting explanations actually help students and educators understand prediction results or improve the quality of academic decision-making. To date, there is no widely agreed-upon evaluation standard for measuring the benefits of XAI explanations from the perspective of human users (Gunning et al., 2019).

4.2.3 Model Technical Limitations and Generalization

From a technical perspective, various XAI methods also face challenges in maintaining model stability and generalization. For instance, the LIME method can give different explanations for similar data when it is used on complicated cases, high-dimensional learning analytics datasets (Hooshyar et al., 2025; Saqr & López-Pernas, 2024). Furthermore, predictive models developed for a specific course or learner group often experience decreased performance when applied to different learning contexts. Changes in learner characteristics, learning design, and interaction patterns within the environment limit the model's generalizability (Er, 2023; Mathrani et al., 2021).

4.2.4 Gap between Population-Based and Individual Needs

The next challenge relates to the discrepancy between how AI models are built and the characteristics of individual learning processes. Most XAI models are developed using large-scale datasets to identify general patterns or average characteristics of a learner population (Saqr & López-Pernas, 2024). However, each student's learning process is dynamic and influenced by various unique factors. As a result, the resulting explanations often reflect group tendencies rather than individual circumstances. This gap limits XAI's ability to provide truly personalized and relevant recommendations for each student, resulting in support that does not fully reflect each individual's learning needs (Saqr & López-Pernas, 2024).

Table 3 synthesizes these four challenges into six recurring categories identified across the reviewed literature, along with their corresponding impact on online learning practice and representative supporting studies.

Table 3. Major Challenges in Implementing XAI for Online Learning

Challenge Category	Description	Impact on Online Learning	Supporting Studies
Pedagogical Explainability	Explanations remain difficult for educators to translate into instructional actions	Low instructional usefulness	Meylani (2025); Saqr & López-Pernas (2024)
Human-centered Evaluation	Evaluation focuses on algorithm performance rather than user understanding	Limited classroom validation	Brasse et al. (2023); Nauta et al. (2023); Kim et al. (2024)
Model Stability	Local explanation methods may generate inconsistent explanations	Reduced trustworthiness	Hooshyar et al. (2025); Saqr & López-Pernas (2024)
Generalizability	Models trained in one context often fail in another	Limited scalability	Er (2023); Mathrani et al. (2021)
Personalization	Population-based prediction overlooks individual learning trajectories	Weak individualized support	Saqr et al. (2024); Ito et al. (2026)
Ethics and Fairness	Privacy, bias, transparency, and accountability remain unresolved	Responsible AI concerns	Eden et al. (2024); Idowu et al. (2024); Nguyen et al. (2023); Niewint-Gori (2025)

Source: Synthesized by the author(s) from reviewed literature (2026).

Collectively, these recurring challenges demonstrate that improving explainability alone is insufficient. Future research must also address pedagogical usability, stakeholder-centered evaluation, and responsible AI governance, which are summarized in Table 4.

Based on the review, several research directions warrant future attention. First, XAI development needs to move toward longitudinal learning analytics that track how student engagement and motivation evolve. Second, causal XAI approaches need to be strengthened so that models identify the factors that genuinely explain learning outcomes rather than mere statistical correlations. Third, pedagogical theory needs to be integrated into AI model development through neural-symbolic architectures, making the resulting explanations more relevant to educators' needs. Fourth, research needs to develop idiographic learning analytics, allowing explanations to be tailored to each student's characteristics rather than based solely on population-level patterns. Fifth, future XAI systems need to support closer human-AI collaboration through prescriptive learning analytics that recommend concrete interventions to educators. Sixth, human-centered evaluation standards must be developed to measure the practical usefulness of explanations from the perspective of real users. Seventh, privacy-preserving approaches such as federated learning and differential privacy are needed to ensure responsible data governance. These seven directions are summarized in Table 4 and discussed in greater depth in the following section.

Table 4. Future Research Directions

Research Direction	Current Limitation	Proposed Future Focus	Expected Contribution
Longitudinal Learning Analytics	Static prediction models	Time-aware learning analytics	Better monitoring of learning trajectories
Causal XAI	Correlation-based explanation	Causal inference	Stronger educational reasoning
Neural-Symbolic AI	Purely data-driven explanation	Integration with pedagogical theory	More meaningful instructional explanations
Idiographic Learning Analytics	Population-level prediction	N-of-1 personalized analytics	Individualized recommendations

Research Direction	Current Limitation	Proposed Future Focus	Expected Contribution
Human-AI Collaboration	Passive explanation	Prescriptive learning analytics	Better teacher-AI collaboration
Human-centered Evaluation	Technical only	evaluation User-centered evaluation metrics	Improved usability and trust
Privacy-preserving XAI	Centralized learning data	Federated learning and differential privacy	Responsible educational AI

Source: Synthesized by the author(s) from reviewed literature (2026).

While these future directions identify key research priorities, translating them into practical XAI systems requires a broader conceptual perspective. Accordingly, the following discussion elaborates these priorities and integrates them into a unified human-focused research agenda.

Overall, the review results indicate that XAI research in online learning has evolved from simply improving model interpretation to developing AI systems that are more transparent, trustworthy, and beneficial to the learning process. Three main themes emerged from the reviewed studies: (1) the growing application of XAI in learning analytics and personalized learning, (2) the persistence of technical and pedagogical challenges, and (3) increasing attention to developing XAI that is human-centered, ethically responsible, and grounded in educational principles. These findings form the basis for the discussion in the following section. In general, the literature also emphasizes the importance of developing human-interpretable XAI standards to ensure their application in online learning is more consistent, transparent, and trustworthy.

4.3. Future Research Directions

To address RQ3, this section synthesizes how future research can bridge the gap between technically accurate AI models and their practical usefulness for educators and learners. Building on the findings from RQ1 and RQ2, the discussion proposes several strategic directions for developing more transparent, trustworthy, and human-centered XAI systems in online learning. Building on the seven priority directions identified in the previous section (Table 4), this section moves beyond summarizing those findings to examine their conceptual implications and to propose concrete paradigm shifts for future research. The findings in this review indicate that the development of XAI in online learning has made rapid progress. However, most research still focuses on improving the technical capabilities of models in generating predictions and explanations. Moving forward, XAI development needs to shift toward a more people-centered approach, aligned with pedagogical needs, and capable of supporting more meaningful decision-making for both educators and students. This shift requires a transformation from simply presenting prediction results to a system capable of generating understandable and actionable learning recommendations.

Table 5 maps each of the six core conceptual challenges discussed below onto its corresponding architectural limitation and the strategic research paradigm proposed to address it, following the same order as subsections (a) through (f).

Table 5. Strategic Alignment Matrix: Core Conceptual Challenges and Proposed Future Research Paradigms

Identified Challenge	Core Current Limitation	Architectural Strategic Future Research Direction / Paradigm Shift
(a) Static, Single-Point-in-Time Prediction	Models capture learning conditions at a single point in time, ignoring how engagement and motivation evolve.	Development of longitudinal and causal-based models to capture dynamic, evolving learning processes.
(b) Explainability-without-Utility Paradox	Surrogate reports (SHAP/LIME charts) display pure statistics without instructional guidance.	Integration of domain-specific pedagogical theory via hybrid neural-symbolic systems.
(c) Aggregate-Instance Mismatch	Models built on cross-sectional average profiles fail to capture individualized learning paths.	Transition to person-specific (N-of-1) learning analytics.

Identified Challenge	Core	Current Limitation	Architectural	Strategic Future Research Direction / Paradigm Shift
(d) Passive, Descriptive Explanation		XAI systems explain predictions but stop short of supporting actionable human decision-making.		Human-in-the-loop, prescriptive learning analytics that recommend interventions to educators.
(e) Functionally-Grounded Evaluation Gaps		Over-reliance on offline engineering proxy metrics rather than real user testing.		Formulation of multi-stakeholder, standardized human-centered evaluation frameworks.
(f) Data Governance and Bias Propagation		Centralized learning data and unmonitored recommenders risk privacy breaches and profiling inequities.		Enforcement of federated learning, differential privacy, and multidisciplinary ethical oversight.

Source: Synthesized by the author(s) from reviewed literature (2026).

4.3.1 Development of Longitudinal and Causal-Based Models

Most current XAI systems still generate predictions based on learning conditions at a single point in time. However, the learning process is a dynamic and constantly changing phenomenon. Therefore, future studies should develop longitudinal models capable of capturing changes in student engagement, motivation, and learning progress throughout the learning process (Murtaza et al., 2026; Nagaraj et al., 2023). Furthermore, the use of causal inference approaches needs to be expanded so that models not only identify statistical relationships between factors but also explain the factors that truly contribute to students' learning success or difficulties.

4.3.2 Integration with Pedagogical Theory

To make the resulting explanations more meaningful to users, XAI development cannot simply rely on patterns discovered from data. AI models need to be integrated with proven learning theories so that the rationale behind each prediction can be understood in an educational context (Sušnjak, 2024). One promising approach is the development of neural-symbolic architectures, which combine machine learning capabilities with knowledge representation, logical rules, and pedagogical theory (Hooshyar et al., 2024). This integration allows the resulting explanations to be more relevant, easier to understand, and more useful for educators in designing learning strategies.

4.3.3 Personalization Through Idiographic Analysis

Most XAI models are still built using general patterns obtained from large populations of learners. However, each student has different characteristics, needs, and learning development. Therefore, further research needs to focus on an idiographic analysis (N-of-1) approach that utilizes longitudinal data from each individual on an ongoing basis (Ito et al., 2026; Saqr & López-Pernas, 2024; Saqr & López-Pernas, 2021). This approach has the potential to produce more personalized explanations and recommendations, including through the use of counterfactual explanations that demonstrate how changes in specific learning behaviors can impact student learning outcomes (Ramaswami et al., 2023).

4.3.4 Human-AI Collaboration in Prescriptive Analytics

The future role of XAI will not only be limited to explaining prediction results but also supporting human decision-making. Therefore, the development of XAI systems needs to adopt a human-in-the-loop approach, placing educators as primary decision-makers with analytical support from AI (Saqr & López-Pernas, 2024). In this context, XAI is expected to evolve toward prescriptive learning analytics, namely systems that not only identify problems but also provide intervention recommendations tailored to the student's condition (Suhail & Rajakumari, 2025; Sušnjak, 2024). The resulting explanations should be presented through an intuitive dashboard so that complex information can be translated into learning recommendations that are easy for educators to understand and use (Nagaraj et al., 2023).

4.3.5 Developing Human-Centered Evaluation Standards

One of the main challenges in XAI research is the lack of evaluation standards capable of measuring the usefulness of explanations from a user perspective. Most studies still look at how good models are using technical measures like accuracy or F1 score. However, they rarely check how explanations affect users' understanding and decision-making (Kim et al., 2024; Nauta et al., 2023). Therefore, further investigation is required to develop a multidimensional evaluation framework that combines technical and human aspects, such as user understanding, cognitive load, trust in the system, decision-making quality, and changes in learning behavior in various educational contexts (Afzaal et al.,

2021; Albreiki et al., 2022).

4.3.6 Data Governance, Ethics, and Personalized Explanations

As the use of personalized learning data increases, privacy and data governance will become increasingly important concerns. An important priority is to integrate approaches that maintain data security, so that models can be developed without compromising student data confidentiality and while still complying with data protection regulations, including GDPR and FERPA (Liu et al., 2025; Sajja et al., 2025; Zhan et al., 2024).

Furthermore, efforts to mitigate algorithmic bias require oversight mechanisms involving various stakeholders. Establishing a multidisciplinary committee consisting of educational researchers, technology developers, educators, and student representatives can be one step to ensure that XAI systems are developed in a fair, transparent, and accountable manner (Niewint-Gori, 2025). Through an ongoing evaluation process, personalized explanations are expected to not only improve the quality of learning but also minimize potential bias and unfairness toward certain groups of students (Eden et al., 2024; Idowu et al., 2024; Nguyen et al., 2023).

Based on the synthesis of current applications, implementation challenges, and future research directions, this review proposes a conceptual framework that illustrates the relationships among XAI techniques, pedagogical interpretation, learner-centered decision-making, and responsible AI governance in online learning environments.



Figure 2. Proposed Human-Centered XAI Framework for Online Learning

Source: Author's own conceptual framework (2026).

Collectively, the proposed framework (Figure 2) synthesizes the principal findings of this review by illustrating how explainability, pedagogy, user-oriented evaluation, and ethical governance interact to support responsible AI adoption in online learning.

5. Limitations

This review has some limitations that should be recognized. First, one person did the checking, choosing, and gathering of data, which may introduce subjectivity despite the use of clearly defined inclusion and exclusion criteria; future replications would benefit from independent dual-reviewer screening with inter-rater reliability assessment. Second, this review did not apply a formal quality appraisal instrument (e.g., MMAT or CASP) to assess the methodological rigor of individual included studies, meaning that studies of varying quality may have contributed unequally to the thematic synthesis. Third, the review was restricted to studies indexed in six major databases and validated through Google Scholar and ERIC; relevant grey literature, unpublished works, or studies outside these sources may not have been captured. These limitations suggest that findings should be interpreted as a comprehensive but not fully exhaustive synthesis of the current XAI-in-online-learning literature. Despite these limitations, the review offers a comprehensive and up-to-date synthesis of recent developments in XAI for online learning. It establishes a foundation for future empirical and interdisciplinary research.

6. Conclusion

This systematic literature review synthesized recent developments in XAI for online learning by examining current applications, implementation challenges, and future research directions. The findings

indicate that XAI has progressed beyond improving model interpretability toward supporting transparent, trustworthy, and human-centered educational decision-making. Nevertheless, important challenges remain regarding pedagogical usefulness, evaluation standards, model generalizability, personalization, and ethical governance.

By integrating evidence from recent studies, this review provides a comprehensive reference for researchers, educators, and AI developers seeking to design more explainable educational AI systems. The proposed people-centered framework highlights the importance of combining explainability with pedagogical theory, user-centered evaluation, and responsible AI practices, offering a conceptual foundation for future research and implementation in online learning.

Bibliography

- Adnan, M., Uddin, M. I., Khan, E., Alharithi, F. S., Amin, S., & Alzahrani, A. A. (2022). Earliest possible global and local interpretation of students' performance in virtual learning environment by leveraging explainable AI. *IEEE Access*, 10, 129843–129864. <https://doi.org/10.1109/ACCESS.2022.3227072>
- Afzaal, M., Nouri, J., Zia, A., Papapetrou, P., Fors, U., Wu, Y., Li, X., & Weegar, R. (2021). Explainable AI for data-driven feedback and intelligent action recommendations to support students' self-regulation. *Frontiers in Artificial Intelligence*, 4, Article 723447. <https://doi.org/10.3389/frai.2021.723447>
- Ajayi, O. O. (2025). Predicting student dropout risk in online learning using stacked ensemble machine learning and explainable AI techniques. *International Journal of Computer Applications*, 187(40), 26–29. <https://doi.org/10.5120/ijca2025925707>
- Akello, E. F., Ijiga, O. M., & Idoko, I. P. (2026). Sequence-aware learning analytics for early identification of at-risk academic trajectories in higher education using transformer models. *International Journal of Innovative Science and Research Technology*, 11(1). <https://doi.org/10.38124/IJSRT/26JAN563>
- Albreiki, B., Habuza, T., & Zaki, N. (2022). Framework for automatically suggesting remedial actions to help students at risk based on explainable ML and rule-based models. *International Journal of Educational Technology in Higher Education*, 19, Article 49. <https://doi.org/10.1186/s41239-022-00354-6>
- Altukhi, Z. M., & Pradhan, S. (2026). Integrating explainable AI in education: Challenges and advancements - A systematic review. *Pacific Asia Journal of the Association for Information Systems*, 18(1), Article 1. https://aisel.aisnet.org/pajais_preprints/57
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & 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>
- Bond, M., Buntins, K., Bedenlier, S., Zawacki-Richter, O., & Kerres, M. (2020). Mapping research in student engagement and educational technology in higher education: A systematic evidence map. *International Journal of Educational Technology in Higher Education*, 17, Article 2. <https://doi.org/10.1186/s41239-019-0176-8>
- Bond, M., Khosravi, H., de Laat, M., Bergdahl, N., Negrea, V., Oxley, E., Pham, P., Chong, S. W., & Siemens, G. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, 21, Article 8. <https://doi.org/10.1186/s41239-023-00436-z>
- Brasse, J., Broder, H. R., Förster, M., Klier, M., & Sigler, I. (2023). Explainable artificial intelligence in information systems: A review of the status quo and future research directions. *Electronic Markets*, 33, Article 26. <https://doi.org/10.1007/s12525-023-00644-5>
- Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, 20, Article 22. <https://doi.org/10.1186/s41239-023-00392-8>
- Eden, C. A., Chisom, O. N., & Adeniyi, I. S. (2024). Integrating AI in education: Opportunities, challenges, and ethical considerations. *Magna Scientia Advanced Research and Reviews*, 10(2), 6–13. <https://doi.org/10.30574/msarr.2024.10.2.0039>
- Er, E. (2023). An explainable machine learning approach to predicting and understanding dropouts in MOOCs. *Kastamonu Education Journal*, 31(1), 143–154. <https://doi.org/10.24106/kefdergi.1246458>
- Gunasekara, S., & Saarela, M. (2025). Systematic literature review on explainable learning analytics and educational data mining. *IEEE Access*, 13, 214387–214408.

- <https://doi.org/10.1109/ACCESS.2025.3643645>
Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), Article eaay7120. <https://doi.org/10.1126/scirobotics.aay7120>
- Hidayati, H., Suyanto, Nurjanah, D., & Laksitowening, K. A. (2024). Leveraging explainable AI to analyze and improve student engagement in virtual learning environment. In *2024 7th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)* (pp. 491–497). IEEE. <https://doi.org/10.1109/ISRITI64779.2024.10963394>
- Hooshyar, D., Šir, G., Yang, Y., Kikas, E., Hämäläinen, R., Kärkkäinen, T., Gašević, D., & Azevedo, R. (2025). *Towards responsible AI for education: Hybrid human-AI to confront the elephant in the room*. arXiv. <https://doi.org/10.48550/arXiv.2504.16148>
- Idowu, J. A., Koshiyama, A. S., & Treleaven, P. (2024). Investigating algorithmic bias in student progress monitoring. *Computers and Education: Artificial Intelligence*, 7, Article 100267. <https://doi.org/10.1016/j.caeai.2024.100267>
- Ilić, M., Mikić, V., Kopanja, L., & Vesin, B. (2023). Intelligent techniques in e-learning: A literature review. *Artificial Intelligence Review*, 56(12), 14907–14953. <https://doi.org/10.1007/s10462-023-10508-1>
- Ito, H., López-Pernas, S., & Saqr, M. (2026). *A systematic review of idiographic research in education: Trends, gaps and future directions*. Hamburg University Press. <https://doi.org/10.57813/eleed.v1ise2026.266>
- Kim, J., Maathuis, H., & Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. *Frontiers in Artificial Intelligence*, 7, Article 1456486. <https://doi.org/10.3389/frai.2024.1456486>
- Kim, J., Merrill, K., Xu, K., & Sellnow, D. D. (2020). My teacher is a machine: Understanding students' perceptions of AI teaching assistants in online education. *International Journal of Human-Computer Interaction*, 36(20), 1902–1911. <https://doi.org/10.1080/10447318.2020.1801227>
- Kumawat, P., & Shaktawat, P. S. (2025). The role of explainable AI (XAI) in enhancing transparency and trust in NLP-powered educational systems. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, 11(4), 432–438. <https://doi.org/10.32628/CSEIT251116172>
- Lagisz, M., Yang, Y., Young, S., & Nakagawa, S. (2025). A practical guide to evaluating sensitivity of literature search strings for systematic reviews using relative recall. *Research Synthesis Methods*, 16(1), 1–14. <https://doi.org/10.1017/rsm.2024.6>
- Lin, C.-C., Huang, A. Y. Q., & Lu, O. H. T. (2023). Artificial intelligence in intelligent tutoring systems toward sustainable education: A systematic review. *Smart Learning Environments*, 10, Article 41. <https://doi.org/10.1186/s40561-023-00260-y>
- Liu, Q., Shakya, R., Khalil, M., & Jovanović, J. (2025). Advancing privacy in learning analytics using differential privacy. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 181–191). Association for Computing Machinery. <https://doi.org/10.1145/3706468.3706493>
- Mathrani, A., Sušnjak, T., Ramaswami, G., & Barczak, A. (2021). Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. *Computers and Education Open*, 2, Article 100060. <https://doi.org/10.1016/j.caeo.2021.100060>
- 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>
- Murtaza, K. G., Cai, L., & Hassan, S. G. (2026). Interpretable time-aware AI framework for early detection and assessment of at-risk learners. *IEEE Access*, 14, 48248–48261. <https://doi.org/10.1109/ACCESS.2026.3675365>
- Nagaraj, B. K., Kalaivani, A., Suraj Begum, R., Akila, S., Sachdev, H. K., & Senthil Kumar, N. (2023). The emerging role of artificial intelligence in STEM higher education: A critical review. *International Research Journal of Multidisciplinary Technovation*, 5(5), 1–19. <https://doi.org/10.54392/irjmt2351>
- Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Computing Surveys*, 55(13), Article 295. <https://doi.org/10.1145/3583558>
- Nguyen, A., Ngo, H. N., Hong, Y., Dang, B., & Nguyen, B. T. (2023). Ethical principles for artificial intelligence in education. *Education and Information Technologies*, 28(4), 4221–4241.

- <https://doi.org/10.1007/s10639-022-11316-w>
- Niewint-Gori, J. (2025). Explainable AI in education: A multi-stakeholder approach to transparency and ethical practice. *IUL Research*, 6(12), 117–132. <https://doi.org/10.57568/iulresearch.v6i12.781>
- Nti, I. K., & Ramanayake, S. (2026). Explainable machine learning for student dropout prediction and tailored interventions in online personalized education. *Discover Artificial Intelligence*, 6, Article 350. <https://doi.org/10.1007/s44163-026-01016-6>
- Ofori, E., & Dake, D. K. (2025). Explainable artificial intelligence in LSTM transformer models for student performance analysis. *Discover Computing*, 28(1). <https://doi.org/10.1007/s10791-025-09814-9>
- Ogata, H., Flanagan, B., Takami, K., Dai, Y., Nakamoto, R., & Takii, K. (2024). EXAIT: Educational explainable artificial intelligent tools for personalized learning. *Research and Practice in Technology Enhanced Learning*, 19, Article 19. <https://doi.org/10.58459/rptel.2024.19019>
- 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>
- Rahmani, A. M., Groot, W., & Rahmani, H. (2024). Dropout in online higher education: A systematic literature review. *International Journal of Educational Technology in Higher Education*, 21, Article 51. <https://doi.org/10.1186/s41239-024-00450-9>
- Ramaswami, G., Sušnjak, T., Mathrani, A., & Umer, R. (2023). Use of predictive analytics within learning analytics dashboards: A review of case studies. *Technology, Knowledge and Learning*, 28(3), 959–980. <https://doi.org/10.1007/s10758-022-09613-x>
- Sajja, R., Sermet, Y., Cwiertny, D. M., & Demir, İ. (2025). Integrating AI and learning analytics for data-driven pedagogical decisions and personalized interventions in education. *Technology, Knowledge and Learning*. <https://doi.org/10.1007/s10758-025-09897-9>
- Saqr, M., & López-Pernas, S. (2021). Idiographic learning analytics: A definition and a case study. In *2021 IEEE 21st International Conference on Advanced Learning Technologies (ICALT)* (pp. 163–165). IEEE. <https://doi.org/10.1109/ICALT52272.2021.00056>
- Saqr, M., & López-Pernas, S. (2024). Why explainable AI may not be enough: Predictions and mispredictions in decision making in education. *Smart Learning Environments*, 11, Article 52. <https://doi.org/10.1186/s40561-024-00343-4>
- Seo, K., Tang, J., Roll, I., Fels, S., & Yoon, D. (2021). The impact of artificial intelligence on learner–instructor interaction in online learning. *International Journal of Educational Technology in Higher Education*, 18(1), Article 54. <https://doi.org/10.1186/s41239-021-00292-9>
- Shastri, K. A., Dinesh, G., Gupta, D., & Nair, A. R. (2025). Predicting student success in online learning using machine learning and explainable AI. In *2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT)* (pp. 566–572). IEEE. <https://doi.org/10.1109/InCACCT65424.2025.11011419>
- Suhail, T. A., & Rajakumari, K. (2025). A human-centric AI tutoring system: Integrating explainability and personalization for improved learning outcomes. In *2025 5th Asian Conference on Innovation in Technology (ASIANCON)* (pp. 1–6). IEEE. <https://doi.org/10.1109/ASIANCON66527.2025.11281019>
- Sušnjak, T. (2024). Beyond predictive learning analytics modelling and onto explainable artificial intelligence with prescriptive analytics and ChatGPT. *International Journal of Artificial Intelligence in Education*, 34(2), 452–482. <https://doi.org/10.1007/s40593-023-00336-3>
- Sušnjak, T., Ramaswami, G. S., & Mathrani, A. (2022). Learning analytics dashboard: A tool for providing actionable insights to learners. *International Journal of Educational Technology in Higher Education*, 19, Article 12. <https://doi.org/10.1186/s41239-021-00313-7>
- Takami, K., Flanagan, B., Dai, Y., & Ogata, H. (2023). Personality-based tailored explainable recommendation for trustworthy smart learning system in the age of artificial intelligence. *Smart Learning Environments*, 10, Article 65. <https://doi.org/10.1186/s40561-023-00282-6>
- Türkmen, G. (2024). 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>
- Vultureanu-Albiş, A., & Bădică, C. (2021). Improving students' performance by interpretable explanations using ensemble tree-based approaches. In *2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)* (pp. 215–220). IEEE. <https://doi.org/10.1109/SACI51354.2021.9465558>

- Yang, W., Wei, Y.-C., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., Yao, N., Wang, X., Gu, X., Amin, M. B., & Kang, B. H. (2023). Survey on explainable AI: From approaches, limitations and applications aspects. *Human-Centric Intelligent Systems*, 3(3), 161–188. <https://doi.org/10.1007/s44230-023-00038-y>
- Zhan, C., Joksimović, S., Ladjal, D., Rakotoarivelo, T., Marshall, R., & Pardo, A. (2024). Preserving both privacy and utility in learning analytics. *IEEE Transactions on Learning Technologies*, 17, 1615–1627. <https://doi.org/10.1109/TLT.2024.3393766>