Applications of Explainable AI (XAI) in Education

(preprint)

Qianhui Liu
University of Illinois Urbana-Champaign
ql29@illinois.edu

Juan D. Pinto
University of Illinois Urbana-Champaign
jdpinto2@illinois.edu

Luc Paquette
University of Illinois Urbana-Champaign
lpaq@illinois.edu

Abstract

As one of the emerging methods for increasing trust in AI systems, explainable AI promotes the use of methods that produce transparent explanations and reasons for decisions made by AI. In this chapter, we present an overview of applications of explainable AI in education with examples of empirical studies. More specifically we discuss the following questions. What are the main categories of explainable AI approaches used to interpret and explain AI models? What are unique characteristics of the use of explainable AI in education? How can explainable AI models be used by teachers, students, and researchers to inform their decisions? What are the challenges and potentials for explainable AI applications in education?

1 Introduction

With rapidly advancing technology and the increasing availability of educational data, building models that leverage this wealth of information has become a compelling avenue to investigate the learning process in order to enhance educational systems. The fields of Educational Data Mining and Learning Analytics have emerged in recent years, focusing on extracting valuable insights and patterns from educational data to improve teaching and learning practices (Baker & Siemens, 2022). Building models with educational data involves the application of data science and machine learning methods to analyze student performance, instructional materials, and various contextual factors. By harnessing the power of educational data, researchers and instructors can gain valuable insights into student behaviors, academic outcomes, and instructional effectiveness, ultimately leading to data-informed decision-making and personalized learning experiences (Ouyang & Jiao, 2021).

Obtaining meaningful insights from models based on educational data holds immense importance in shaping the future of AI applications in education. Humans, not machines, are the main actors in putting the power of data-driven models into effect. Learners and teachers need to gain valuable insights into learning patterns, performance factors, instructional effectiveness, and curriculum design that support targeted interventions for inclusive and personalized learning environments (Bewersdorff et al., 2023). Moreover, insights derived from educational data models aid in the development of policies related to curriculum enhancements and the optimization of educational resources allocation (Lee, 2021). However, taking full advantage of insights derived from data-driven models requires humans to be able to inspect the model to better understand how it makes its decision. For this reason, some researchers have argued for the need for an additional explainability criteria to evaluate how useful a data-driven model is for education (Cardona et al., 2023).

Explainable Artificial Intelligence (XAI) has been an emerging area of research and application for the development of models from educational data. It is one of the methods for increasing trust in AI systems, and promotes the use of methods that produce transparent explanations and reasons for the decisions AI makes (Khosravi et al., 2022). By providing explanations for how models arrive at specific outcomes through visualizations or texts, XAI enables researchers who build the XAI method, instructors and students who are involved in the educational context to understand the underlying features, rules, or other factors that influence the decision-making process. This transparency empowers researchers and instructors to detect areas for improvement (Lu et al., 2020; Zhu et al., 2020), identify potential biases (Tyler et al., 2021), and make informed instructional decisions (Swamy et al., 2023). Additionally, XAI in education can help students understand their own learning progress, receive personalized feedback (Mu et al., 2020) and interventions (Hur et al., 2022), and gain insights into their strengths and weaknesses Scheers & De Laet (2021). By integrating XAI into educational systems, we can foster a more inclusive, accountable, and ethically responsible approach to leveraging AI in education.

The adoption of XAI in education is still in its early stages and researchers are still exploring the potential of XAI methods. Recently, Khosravi et al. (2022) proposed a framework for XAI in education, XAI-ED, with six key aspects and questions to guide the study design, opportunities, challenges, and future research needs for advancing XAI in education. In this book chapter, we survey a sample of empirical studies on XAI to reveal potential benefits of explainability methods in educational contexts. While the business models that dictate uses of AI in education (proprietary vs. open-source) can have clear implications for the ability to scrutinize real-world AI models, we sidestep that discussion here and focus instead on technical interpretability (assuming full access to the model parameters) and its implications. The rest of the chapter is organized as follows. First, we introduce key terms in the community of XAI, including explainability vs interpretability, as well as two ways to design XAI: ante-hoc vs post-hoc. Then, we discuss applications of XAI in education, organized based on which target group benefits from the explanation while presenting existing empirical studies that belong to each category. We then identify emerging challenges for XAI in education and conclude by proposing four areas of work to further advance XAI applications in education.

2 XAI for Education

While the use of the term XAI has seen a recent increase in popularity in education research, some topics surrounding explainability have been explored in educational contexts with other names. ‘Human-in-the-loop’ is such a term: it has been used to describe situations where humans have been involved in the model development process by providing meaningful input and expert tagging. According to Rosé et al. (2019), the inclusion of a human-in-the-loop component is exactly what leads the model results to be interpretable. The term ‘human-centered AI (HAI)’ is another broader term related to XAI. Yang et al. (2021) describe two perspectives in HAI: 1) AI under human control, and 2) AI on the human condition. The later one is closely related with XAI as it requires the explanation and interpretation of the computations, judgements, and adjustment processes of AI algorithms through human context.

2.1 Explainability vs Interpretability

Two terms that are more central to the field of XAI are explainability and interpretability. Many researchers agree that there are some basic conceptual differences between explainability and interpretability. In general, the term explainability has been used to refer to the act of explaining the decisions made by machine models in a human-understandable form (Haque et al., 2023). In contrast, the term interpretability has mostly been used to refer to the intrinsic capability of a model to allow a human to understand its inner logic (Adadi & Berrada, 2018; Haque et al., 2023). However, some researchers seem to hold the opposite opinion about the definition of the two terms (Linardatos et al., 2021; Minh et al., 2022). As there is no commonly recognized distinction between the two terms (Alamri & Alharbi, 2021), throughout this chapter we use ‘explainability’ and ‘interpretability’ interchangeably to refer to all attempts at understanding a model’s decision-making process. Within this overarching category, there exist two types of explainability approaches: ante-hoc and post-hoc methods.

2.2 Ante-hoc vs Post-hoc

Ante-hoc explainability (Alamri & Alharbi, 2021) refers to a model that a human can directly inspect and understand without the need of external analysis tools or proxy models. These are terms that define the nature of the model itself. Ante-hoc explainability is a spectrum that places simpler algorithms (sometimes referred to as white-box/glass-box models; see Mathrani et al., 2021), such as linear regression or small decision trees, on the interpretable end and more complex ones, such as neural networks, on the inscrutable end.

In contrast, post-hoc explainability does not characterize a model but rather a method or set of methods aimed at understanding existing models. These post-hoc methods are applied after a model has been created (hence the name), acting as a wrapper around the model to extract insights from indirect observations of model predictions. While some post-hoc methods also use the internal parameters of models for their explanations, the most common methods are model agnostic and rely exclusively on a model’s inputs and outputs. These methods are most often used to better understand black-box models that are not intrinsically interpretable (though model-agnostic methods can also be used on traditional white-box models; eg. Vultureanu-Albişi & Bădică, 2021).

At face value, the majority of models used in education have traditionally been based on intrinsically interpretable algorithms and thus fall under the ante-hoc explainability category. These are typically linear or logistic regression models, but also including more complex approaches such as structural equation modeling or hierarchical linear modeling that nevertheless contain interpretable linear models at their core. Part of the reason for this is that black-box models are a more recent phenomenon. Artificial neural networks, for example, have only become popular in recent years with the advent of big data and powerful graphics cards for parallel processing—despite the groundwork having been laid many decades ago (Tappert, 2019).

Despite the early popularity of AI methods with more ante-hoc explainability, it is post-hoc explainability that has received the most attention in recent years. Many different post-hoc explainability methods have been developed, and research into these methods make up the bulk of publications in the field of XAI. This may be in part due to the increased popularity of advanced AI methods, such as deep neural networks, which often outperform traditional approaches. Even models that use less-complex algorithms can quickly become too large to realistically understand (e.g. a model with a very large number of input features or a decision tree with a large maximum depth and hundreds of leaves), rendering them likewise opaque. In such contexts, post-hoc methods have been used to explain these black-box models.

One important distinction among different post-hoc explainability approaches is that between global vs. local explainability. Global explainability is concerned with explaining how the model works at the top level—that is, which input features are most important in how a model makes decisions or affect the model's decisions overall. Local explainability, on the other hand, deals with questions at the instance level, such as the impact of input features on a specific decision. While it is possible to obtain global measures by aggregating local explainability results, this approach can lead to very different results. Global explainability is important when scientific understanding is the goal, whereas local explainability is better when justification for a specific decision is required (Doshi-Velez & Kim, 2017).

3 To whom and for what Purpose can XAI be Applied in Education?

XAI techniques enable humans to gain insights into the underlying patterns and factors driving a model's decisions. By providing explanations and justifications for the discovered knowledge, XAI enhances our understanding and trust in the discovered insights. This transparency helps researchers, analysts, and domain experts validate their findings, identify potential biases or errors, and further refine the knowledge discovery process. XAI allows for a more collaborative and interactive approach, enabling human experts to work with AI models to uncover valuable insights and make informed decisions based on the discovered knowledge.

In the context of education, XAI can benefit different stakeholders, including researchers, instructors, and students. As such, it is important to keep in mind who will make use of the explanation and how this explanation will be used, as this may inform the choice of which XAI method is best suited for a given task. Indeed, an explanation that is easy to understand for an educational researcher may not be accessible to instructors or students, and the usefulness of an explanation will depend on its target audience.

In this section, we introduce how XAI has been used in education to inform researchers, instructors and students. We found that XAI in education has mostly been targeted at researchers, while a more limited number of studies directly informed instructors and students. However, we believe that instructors and students could be important beneficiaries of XAI, and further work could focus in this direction.

3.1 Researchers

3.1.1 Informing our Understanding of Learning

One of the primary applications of data-driven models by educational researchers is to reveal new insights about the learning process. Generally, these new insights can be informed based on both a model's outputs, its input features and the relationship between them. However, the nature of these insights will vary based on how easy a model and its prediction can be explained. The use of non-explainable methods may limit researchers to insights that can be obtained by examining only the model's output, while the use of explainable input features and explainable methods may allow researchers to obtain richer insights about the relationship between input features and outputs.

Educational researchers have incorporated explainable methods to discover new insights about learning beyond what may be revealed by non-explainable methods. For example, Effenberger & Pelánek (2021) used an interpretable clustering method to identify different groups, or clusters, containing similar solutions used by students when solving programming problems in an introductory computer science course. To achieve this, the authors proposed an interpretable data mining pipeline that emphasizes the development of interpretable input features, patterns, and visualizations. These principles were applied throughout the different stages of the clustering process to maximize the knowledge that can be gained from the resulting clusters: feature selection, pattern mining, pattern selection, and clustering summarization. In the context of their study, Effenberger & Pelánek (2021) focused on selecting a list of input features that captured comprehensive and interpretable keywords in the students' programs. Different from traditional data-driven clustering methods, the proposed model clustered the students' solutions around explainable programming keywords, which allowed the researchers to gain more meaningful insights about the nature of students' solutions and whether these solutions align with pedagogical goals.

Effenberger and Pelánek's work represents a branch of explainable methods in education where, instead of using a single algorithm or model to improve explainability, an interpretable data mining pipeline is implemented. In such a pipeline, explainability is considered at all stages of the data mining process by selecting interpretable features and patterns, and by providing visualization results to researchers. Such an interpretable approach can be generalized to other contexts and problems.

Another approach to using XAI to better understand the learning process is through the inspection of the relationships captured by the data-driven models themselves. However, this can prove challenging for black-box models created using advanced AI approaches such as deep learning. The black box nature of deep learning models has been hindering researchers and users from knowing the mechanism behind their high-accuracy predictions. However, some researchers have explored how to reveal the hidden states of deep learning models and through intrinsically explainable deep learning models. For example, Pinto et al. (2023) created a gaming-the-system detector using an interpretable convolutional neural network (CNN) based on an approach originally described by Jiang & Bosch (2021). They constrained the kernel weights in the convolutional layer to be either close to 0 or 1 instead of the standard continuous values. With this binary representation, the model can extract patterns that closely approximate the way humans understand the world. Both studies found that the patterns extracted from the interpretable CNN model achieved higher prediction performance while remaining interpretable. Pinto et al. (2023) were able to compare patterns to those identified by domain experts, providing a method to inform researchers' understanding of gaming-the-system behavior with insights from the model itself. In general, explainability methods have shown great potential in opening the black box of deep learning models. Both studies mentioned above leveraged a novel deep learning approach as a way to mimic experts' understanding of learning patterns. They showed the possibility of creating interpretable deep learning models that could be applied to diverse educational contexts with the goal of improving our understanding of the learning process.

3.1.2 Improving Model Quality

Researchers use diverse metrics, such as goodness of fit indicators and measures of predictive accuracy, as measures of a data-driven model's quality. Researchers will attempt to iteratively improve their models by maximizing these measures. However, while these measures provide information about a model's overall quality, they do not provide rich information about why a model may be performing poorly. As such, improving a model's quality will either involve a process of trial and error, where different modeling and different data pre-processing approaches are compared to each other to select the best model, or will depend on the use of additional tools to identify areas of potential improvements. Explanations provided by XAI methods can be a useful tool to support the iterative improvement of models.

Studies have used explainable methods to improve models, especially during the training of deep knowledge tracing (DKT) models. Zhu et al. (2020) used a Finite State Automaton (FSA) to interpret the hidden state transition of DKT when receiving inputs. The FSA accepts an input item and evaluates whether this item has a positive effect on the final prediction outputs of the model, and vice versa. Based on the results of the FSA, the authors found that the traditional DKT approach was not able to handle the long sequence of input for this study. They proposed using an attention-based model as an alternative. Similarly, Lu et al. (2020) also adopted a post-hoc XAI method to interpret Recurrent Neural Network-based DKT models. The layer-wise relevance propagation (LRP) was applied to calculate the relevance from the model's output layer to its input layer. By relating the input and output of DKT models, the XAI methods mentioned above helped researchers understand why their DKT models did not perform well in certain cases. Based on the explainable results, researchers were then able to improve their models according to the revealed weakness instead of through trial and error.

Explainability methods also have the potential to be combined with other methods as a way to reduce model bias. For example, Tyler et al. (2021) combined nonnegative matrix factorization (NMF) and bias mitigation to predict STEM degree enrollment. This study manipulated a raw dataset containing a high quantity (4000) of input variables related to students' STEM enrollment. This high number of variables can significantly impact the explainability of a data-driven model developed using this data. NMF was used to identify top factors for predicting STEM enrollment where each factor consists of variables that align with meaningful and interpretable educational phenotypes, such as academic performance and student identity in science. They measured and compared gender biases of models created using NMF to that of a baseline model created using the raw data. Their result showed that models created using a smaller number of interpretable factors (identified through NMF) showed less bias without significantly reducing predictive accuracy of the model. This study showed another possible goal of XAI in mitigating biases toward historically marginalized groups that might arise in data-driven models.

3.1.3 Selecting Data-Driven Interventions

Data-driven models have been used with the goal of deploying automated interventions to support students during their learning activities. In this context, models are used to detect situations of interest that are used to trigger student-facing interventions. By allowing researchers to better understand the factors leading a model to detect a given situation of interest, XAI methods provide a useful tool to improve the design of data-driven interventions that more directly attempt to address those factors.

Studies have helped researchers design interventions based on the important features in prediction models. Mu et al. (2020) calculated SHAP values—quantifying the contribution of features to each prediction—for each input feature in their model to select interventions to support students who are ‘wheel-spinning’ (when a student repeatedly fails to complete an educational task). In this study, the SHAP values were used to identify which actionable input feature contributed most to the model’s prediction of wheel-spinning behavior. This feature could then be used to select an appropriate intervention that directly addresses a potential explanation for the student’s wheel spinning behavior, for example if the feature identifies a possible gap in the students’ prerequisite knowledge. The authors compared the intervention selected using XAI to those selected using logistic regression, XGBoost feature importance, and a domain expert’s prescription. The results showed that the suggestions based on the proposed method aligned the most to the expert’s suggestion. Although this study proved the feasibility of using XAI to provide better interventions, it heavily relied on the model being built using actionable features that were themselves associated with potential interventions. Thus, this study highlights the importance of combining explainability metrics such as SHAP values with meaningful and interpretable features.

Similarly, Hur et al. (2022) used SHAP values to provide interventions encouraging the use of self-regulated learning behavior as students learned introductory statistics topics. The SHAP-driven recommendations were applied for a small curriculum of 12 introductory statistics topics in a self-guided, self-paced online learning system. A random forest regressor was trained using pretest score, quiz score (12 topics), and reading time (12 topics) features to predict posttest score. SHAP values were calculated to reveal features that most negatively influenced predicted learning outcomes, and interventions were selected based on these SHAP values. A randomized controlled trial of 73 participants was conducted to compare SHAP recommended actions and expert recommendations based on predefined rules. Students in the two conditions performed similar learning and topic-choosing behaviors, suggesting that XAI-informed interventions could facilitate statistics learning to a similar degree as expert interventions.

3.2 Instructors

3.2.1 Modifying Curriculum

AI models can uncover meaningful patterns and correlations in educational data that might not be immediately apparent to instructors. By analyzing student performance data, engagement levels, or learning progress, models can identify patterns that indicate the effectiveness of specific curriculum elements or teaching strategies. However, it may be challenging for instructors who lack advanced knowledge of AI techniques to interpret these patterns with the goal of obtaining actionable insights. XAI methods can be used to make the results of data-driven models more accessible to instructors. Instructors can then leverage these insights to tailor curriculum design, instructional materials, and learning activities to better meet the needs of their students.

Swamy et al. (2023) compared the explanations produced by two post-hoc XAI methods, LIME and SHAP, by applying them on models trained to predict student success in a series of Massive Open Online Courses (MOOCs) by presenting them to course instructors. Their goal was to investigate the instructors’ trust in the explanations, and how they may interpret and make use of the generated explanations. While the results showed that instructors did not consistently prefer explanations generated by either of the two methods and just as often selected a confounder explanation for which inaccuracies were purposefully added, the study also showed how explanations could be used to engage instructors in reflection about the course, helping provide actionable course design decisions.

3.2.2 Identifying Students in Need

Data-driven models have been used to provide actionable information to instructors, for example through learning analytics dashboards (Schwendimann et al., 2017). The use of such dashboard supports instructors in responding to the academic needs of each student—for example, by providing descriptive information about students' learning behaviors or early warnings for at-risk students. However, many dashboards have low impact on instructors due to the lack of (1) contextualization, (2) grounding to learning theories, and (3) explanation for black-box predictive models (Shabaninejad et al., 2022). XAI methods can present contextual information to instructors, providing them with additional explanations to facilitate their interpretation of the information presented on the dashboard, with the goal of better supporting students in need.

Shabaninejad et al. (2022) developed an instructor-facing dashboard, called Students Inspection Facilitator (SIF), to help identify students in need of attention. Specifically, this interface takes students' online learning log data as input and presents a list of students, color-coded and ordered by their priority to be inspected, based on five risk factors related to performance, engagement, and deviation from the class. To provide explainability and interpretability, the dashboard not only includes a radar chart showing each student's weekly performance and engagement, but this chart is also augmented with comparisons to the class's first, second and thirds quartiles and with textual flags such as "low performance" and "inconsistent engagement". In addition, a "help interface" was designed to provide information to instructors about the process used to rank students and to provide guidelines on how to interpret the dashboard's visualizations. For each student in need, the instructors can scan through the holistic view of their data and read the interpretation of their risk factors to make decisions on proactive support. Although SIF has not been evaluated by instructors in authentic settings, the authors proposed insightful design objectives and principles for XAI dashboards for instructors.

3.3 Students

3.3.1 Promoting Understanding and Trust

Students are another user group that can benefit from XAI. The use of XAI methods can serve as an important tool to promote students' understanding of their learning process by providing them with explanations of how their learning performance may be related to their patterns of engagement with learning activities. In addition, the use of explanations may also be beneficial in increasing students' trust in the AI-driven systems that impact their learning experience—for example, by providing them with an explanation of why they received a given intervention. Similar to the use of XAI for instructors, student-facing explanations will need to be carefully designed to be easily accessible to students with limited understanding of AI systems and how they use data to make decisions.

In their work, Conati et al. (2021) explored how to increase student uptake and the effectiveness of AI-driven hints provided by an intelligent tutoring system through an explanation interface. The authors first used a machine learning framework called FUMA to discover classes of exploratory behaviors exhibited by students when using the tutor and groups of students based on their learning needs, which were then used to generate adaptive hints for students. They then developed an explanation interface to convey to students the motivations (why) and processes used (how) by the intelligent tutor when selecting which hints students received. This explanation interface aimed to promote the students' understanding on two levels: 1) an understanding of the AI driving hints in the learning platform, and 2) an understanding of the specific hints that they received. The results indicated that providing explanations increased the students' trust in the hints, their perceived usefulness of the hints, and their intention to use hints again in the future.

4 What are the Challenges for XAI Applications in Education?

While there are many opportunities for improving the utility of data-driven educational models using XAI, there remain open challenges that limit their wider adoption. In this section, we specifically discuss three important challenges: *1) the lack of standardized evaluation methods to assess the quality of explanations generated through XAI, 2) the tradeoff between accuracy and explainability and 3) the limits of post-hoc explanation methods.*

4.1 Lack of Standardized Evaluations

Standard metrics to evaluate either specific explanations or the intrinsic interpretability of a model are not currently used but could go a long way in motivating greater emphasis on explainability and transparency. None of the studies included in Alamri & Alharbi (2021) review of explainable student performance prediction models evaluated explainability, but the authors argue that reporting such metrics should become as commonplace as reporting model accuracy. However, this type of evaluation is not a straight-forward task, depending heavily on contextual factors such as the type of model, the XAI method used, as well as the intended end-users and overall purpose of the model and explanation.

As a good starting point in this direction, Doshi-Velez & Kim (2017) proposed a taxonomy of domain-agnostic evaluation methods that can be used by education researchers interested in explainability. From most specific and costly to less robust and resource-intensive, they organized these approaches into application-grounded, human-grounded, and functionally grounded evaluations.

Application-grounded evaluation involves real humans and real tasks. This involves testing a model in the real-world target application for which it was developed. For example, an instructor-facing model's explanation of its predictions can be evaluated by how well instructors are able to use it to address the needs of students. This evaluation approach can be costly and time consuming, but it has high fidelity with the needs of end-users.

Human-grounded evaluation involves real humans but simplified tasks. It involves measuring how well humans can accurately answer questions about a model based on its interpretation. Examples of the types of experiments in this category include (1) binary forced choice, where participants must pick which explanation they consider best when presented with multiple options (e.g. Swamy et al., 2023), (2) forward simulation/prediction, where participants must correctly predict the model's output given specific inputs, and (3) counterfactual simulation, where participants must correctly identify how a specific input needs to be changed in order to alter the model's given output. The evaluation approach taken by Mu et al. (2020), in which they compared interventions based on SHAP explanations to what an expert would prescribe, would also fall into this category. Human-grounded evaluation provides a general measure of a model's interpretability or explanation for real humans, but it may not capture the specific needs of the end-user.

Functionally grounded evaluation does not involve humans and uses proxy tasks. This involves measuring some specific aspects of a model or its explanation that capture constructs related to interpretability. The kinds of metrics described by Alamri & Alharbi (2021), as well as the real-world consistency metric used by Lu et al. (2020), would fit into this category. Examples of specific quantifiable properties include model sparsity, simplicity, or stability. These metrics act as a proxy for interpretability and are the most cost-effective option—but the least valid—since no humans are involved. They are an especially appropriate option when the particular class of models or explainability methods have been previously validated via application- or human-grounded experiments, or for use in a preliminary study in a domain without much prior research.

It is important to note that the evaluation approaches described here are concerned with the overall interpretability of a model, the accuracy of an explanation with regards to the model in question, and the usability of interpretations/explanations in real-world settings. They do not attempt to measure the validity of a model or the fidelity of an explanation in relation to the real-world construct involved.

4.2 Tradeoff between Accuracy and Explainability

As XAI methods introduce a new criterion to AI model applications in education, some studies have shifted their focus to discuss the tradeoff between model accuracy and explainability. To put this in a perhaps oversimplified way, more complex models generally lead to higher accuracy, but the heightened complexity simultaneously makes it difficult for humans to understand the models' inner mechanisms. For example, Oliveira et al. (2023) compared the performance of two models in essay cohesion prediction, one is a semantic feature-based model and the other is a BERT-based (one of the pre-trained deep neural networks) model. The authors concluded that, although BERT-based model can reach higher prediction accuracy, this model is based on word-level information and is thus not as explainable as the semantic feature-level model. This raises a question for XAI applications in education: how to choose between a model accuracy and its explainability. Despite various ante-hoc and post-hoc methods, studies have mentioned the lowering of model accuracy as the consequence of applying XAI in education (Jiang & Bosch, 2021; Oliveira et al., 2023), though some researchers argue that this tradeoff is an inaccurate assumption (Rudin, 2019). When implementing and choosing XAI methods in research and practice, the possibility of a decrease in accuracy should be taken into consideration.

4.3 Limits of Post-hoc Methods

Some researchers—mainly outside of education—have discussed various problems with post-hoc explainability methods as a whole. These critiques often come with calls for a re-conceptualization of the problem and a shift toward greater ante-hoc interpretability for models used in high-stakes domains.

One of the most conspicuous issues with post-hoc methods is the disagreement problem. This refers to the finding that such explanation generated using different post-hoc methods often disagree on their explanations of the same model. In an attempt to formalize certain measures of disagreement, Krishna et al. (2022) conducted interviews with data scientists. They found that explanation disagreement occurs often and that data scientists are aware of this problem but do not know what to do about it. When confronted with this problem, many of those interviewed often used arbitrary heuristics, such as trusting the explanations of their ‘favorite method’. The authors propose rethinking the field of XAI, starting from an agreed-upon set of guiding principles to develop more consistent post-hoc explainability methods.

Taking a similar approach to quantify the disagreement problem but in the education domain, Swamy et al. (2022) compared five different post-hoc methods on models trained to predict student success in a series of MOOCs. They demonstrated important disagreement using quantitative methods, including Spearman’s rank correlation, Jensen Shannon Distance, and Principal Component Analysis. The same team conducted a follow-up study, again finding significant disagreement between LIME and SHAP explanations for a different series of models (Swamy et al., 2023). They also surveyed educators on their trust and preference between these and a fake confounder explanation created for the sake of the study. The educators did not consistently prefer any one explanation and just as often selected the confounder explanation as the one they believed to be the most accurate, demonstrating the magnitude of this problem. Most studies that use post-hoc methods rely on a single approach, and even when using multiple methods are used, they often fail to address or clearly report disagreements (eg. Baranyi et al., 2020).

Aside from the disagreement problem, some have questioned the very validity of post-hoc explanations. Laugel et al. (2019) demonstrated that counterfactual explanations have a high risk of generating unjustified examples which can lead to dubious explanations when compared with ground-truth data. They went as far as to argue that all current model-agnostic state-of-the-art post-hoc explainability methods are liable to errors in how they understand a model, precisely because they treat the model as a literal black box.

Rudin (2019) takes this point further to suggest that the use of the word ‘explainable’ or ‘explanations’ in reference to post-hoc methods is inaccurate and misleading, and that we should instead refer to such attempts as ‘summaries of predictions’, ‘summary statistics’, or ‘trends’. Providing copious examples, they identify five key issues with post-hoc explainability: (1) the accuracy-interpretability tradeoff is a myth, (2) post-hoc explanations are not faithful to the original model, (3) explanations are often non-sensical or unclear, (4) explanations alone are often not enough for high-stakes decisions and yet are difficult to properly use in combination with other parts of the decision-making process, and (5) adding explanations on top of black-box models simply makes the system more complicated and thus increases the chance of human error. Ultimately, Rudin (2019) concludes that post-hoc explainability is too problematic to be used in high-stakes scenarios and that we should instead be focused on trying to create intrinsically interpretable models.

In the domain of education, discussion of these issues has only just begun. Based on the methods and assumptions in much of the work dealing with interpretability, it seems that researchers are largely unaware of the problems with post-hoc explainability. By way of example, Scheers & De Laet (2021) investigated students’ reactions when presented with a LIME explanation of a model’s predictions on their own future performance. They observed that counterintuitive explanations caused students to re-evaluate their mental models, ultimately increasing trust in the system. However, the accuracy of the explanations is never questioned, the accuracy of the model is never questioned, and the model is assumed to capture causality in the real world. Awareness of the issues identified with current post-hoc methods, along with a greater critical understanding of and agreement on proper XAI methodology, may be important first steps in addressing this challenge.

5 Conclusion

In this chapter, we provided a brief overview into Explainable AI (XAI) and its applications in education. Although the formal study of methods for the explanation of data-driven AI models in education is still in their early stage, we present examples of studies exploring the potential of XAI to support researchers, instructors, and students in different educational contexts. These examples illustrate the diverse potential benefits of XAI for education. We note that most applications of XAI in educational settings presented in this chapter focused on providing explanations to researchers in order to inform our understanding of the learning process, improve model quality, and better select data-driven interventions. However, there existing meaningful opportunities to use XAI to support both instructors, to improve their curriculum design and help identify struggling students, and students, for example to promote their understanding of the learning material and to improve their trust in AI systems.

As the use of XAI methods in education continues to grow, we propose four areas of work to further advance XAI applications in education. First, pay more attention to implementing explainability methods with instructors and students in mind. Currently, limited studies have focused on these two user groups of XAI in education. Compared with researchers, instructors and students have less background knowledge about AI models, thus require more support to understand, trust, and accept AI results in their teaching and learning. Second, develop evaluation methods of explainability. As discussed in the previous section, current research is still exploring the feasibility of XAI in education, and very limited studies have considered how to evaluate these methods. This is closely related to the questions of for whom and for what purpose XAI is applied in education. We expect that researchers, instructors, and students would each take advantage of explainability from different perspectives. Therefore, evaluation methods should also differ accordingly. Third, it may be beneficial to prioritize models that provide intrinsic interpretability. Some traditional models such as regression and decision trees have been widely used in education, but without a specific focus on interpretability. Maximizing the interpretability of models produced using these approaches can be a worthwhile endeavor which can further inform the application of other models with similar interpretability design or criteria. Fourth, be mindful when using post-hoc explainability methods. Various post-hoc methods are found to be inconsistent when calculating feature importance and there is yet no sound conclusion on which method is more helpful than others. In many cases, it may be best to combine the results of multiple post-hoc methods to reach conclusions or consider other explainability methods.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: A systematic review. *IEEE Access*, *9*, 33132–33143. https://doi.org/10.1109/ACCESS.2021.3061368

Baker, R. S., & Siemens, G. (2022). Learning analytics and educational data mining. In R. K. Sawyer (Ed.), *The Cambridge Handbook of the Learning Sciences* (3rd ed., pp. 259–278). Cambridge University Press. https://doi.org/10.1017/9781108888295.016

Baranyi, M., Nagy, M., & Molontay, R. (2020). Interpretable deep learning for university dropout prediction. *Proceedings of the 21st Annual Conference on Information Technology Education*, 13–19. https://doi.org/10.1145/3368308.3415382

Bewersdorff, A., Zhai, X., Roberts, J., & Nerdel, C. (2023). Myths, mis- and preconceptions of artificial intelligence: A review of the literature. *Computers and Education: Artificial Intelligence*, *4*, 100143. https://doi.org/10.1016/j.caeai.2023.100143

Cardona, M. A., Rodríguez, R. J., & Ishmael, K. (2023). *Artificial intelligence and the future of teaching and learning*. U.S. Department of Education, Office of Educational Technology.

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence*, *298*, 103503. https://doi.org/10.1016/j.artint.2021.103503

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning* (No. arXiv:1702.08608). arXiv. https://arxiv.org/abs/1702.08608

Effenberger, T., & Pelánek, R. (2021). Interpretable clustering of students' solutions in introductory programming. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 101–112). Springer International Publishing. https://doi.org/10.1007/978-3-030-78292-4_9

Haque, A. B., Islam, A. K. M. N., & Mikalef, P. (2023). Explainable Artificial Intelligence (XAI) from a user perspective: A synthesis of prior literature and problematizing avenues for future research. *Technological Forecasting and Social Change*, *186*, 122120. https://doi.org/10.1016/j.techfore.2022.122120

Hur, P., Lee, H., Bhat, S., & Bosch, N. (2022). Using machine learning explainability methods to personalize interventions for students. *Proceedings of the 15th International Conference on Educational Data Mining*, 438–445. https://doi.org/10.5281/ZENODO.6853181

Jiang, L., & Bosch, N. (2021). Predictive sequential pattern mining via interpretable convolutional neural networks. *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, 761–766.

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y.-S., Kay, J., Knight, S., Martinez-Maldonado, R., Sadiq, S., & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022). *The disagreement problem in explainable machine learning: A practitioner's perspective* (No. arXiv:2202.01602). arXiv. https://doi.org/10.48550/arXiv.2202.01602

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., & Detyniecki, M. (2019). *The dangers of post-hoc interpretability: Unjustified counterfactual explanations* (No. arXiv:1907.09294). arXiv. https://arxiv.org/abs/1907.09294

Lee, Y. (2021). Applying explainable artificial intelligence to develop a model for predicting the supply and demand of teachers by region. *Journal of Education and e-Learning Research*, *8*(2), 198–205. https://doi.org/10.20448/journal.509.2021.82.198.205

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy*, *23*(1), 18. https://doi.org/10.3390/e23010018

Lu, Y., Wang, D., Meng, Q., & Chen, P. (2020). Towards interpretable deep learning models for knowledge tracing. *Artificial Intelligence in Education*, *12164*, 185–190. https://doi.org/10.1007/978-3-030-52240-7_34

Mathrani, A., Susnjak, T., Ramaswami, G., & Barczak, A. (2021). Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. *Computers and Education Open*, *2*, 100060. https://doi.org/10.1016/j.caeo.2021.100060

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: A comprehensive review. *Artificial Intelligence Review*, *55*(5), 3503–3568. https://doi.org/10.1007/s10462-021-10088-y

Mu, T., Jetten, A., & Brunskill, E. (2020). Towards suggesting actionable interventions for wheel-spinning students. *Proceedings of The 13th International Conference on Educational Data Mining (EDM 2020)*, 183–193.

Oliveira, H., Ferreira Mello, R., Barreiros Rosa, B. A., Rakovic, M., Miranda, P., Cordeiro, T., Isotani, S., Bittencourt, I., & Gasevic, D. (2023). Towards explainable prediction of essay cohesion in Portuguese and English. *LAK23: 13th International Learning Analytics and Knowledge Conference*, 509–519. https://doi.org/10.1145/3576050.3576152

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, *2*, 100020. https://doi.org/10.1016/j.caeai.2021.100020

Pinto, J. D., Paquette, L., & Bosch, N. (2023). Interpretable neural networks vs. Expert-defined models for learner behavior detection. *Companion Proceedings of the 13th International Conference on Learning Analytics & Knowledge Conference (LAK23)*, 105–107.

Rosé, C. P., McLaughlin, E. A., Liu, R., & Koedinger, K. R. (2019). Explanatory learner models: Why machine learning (alone) is not the answer. *British Journal of Educational Technology*, *50*(6), 2943–2958. https://doi.org/10.1111/bjet.12858

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Scheers, H., & De Laet, T. (2021). Interactive and explainable advising dashboard opens the black box of student success prediction. In T. De Laet, R. Klemke, C. Alario-Hoyos, I. Hilliger, & A. Ortega-Arranz (Eds.), *Technology-Enhanced Learning for a Free, Safe, and Sustainable World* (Vol. 12884, pp. 52–66). Springer International Publishing. https://doi.org/10.1007/978-3-030-86436-1_5

Schwendimann, B. A., Rodríguez-Triana, M. J., Vozniuk, A., Prieto, L. P., Boroujeni, M. S., Holzer, A., Gillet, D., & Dillenbourg, P. (2017). Perceiving learning at a glance: A systematic literature review of learning dashboard research. *IEEE Transactions on Learning Technologies*, *10*(1), 30–41. https://doi.org/10.1109/TLT.2016.2599522

Shabaninejad, S., Khosravi, H., Abdi, S., Indulska, M., & Sadiq, S. (2022). Incorporating explainable learning analytics to assist educators with identifying students in need of attention. *Proceedings of the Ninth ACM Conference on Learning @ Scale*, 384–388. https://doi.org/10.1145/3491140.3528292

Swamy, V., Du, S., Marras, M., & Kaser, T. (2023). Trusting the explainers: Teacher validation of explainable artificial intelligence for course design. *LAK23: 13th International Learning Analytics and Knowledge Conference*, 345–356. https://doi.org/10.1145/3576050.3576147

Swamy, V., Radmehr, B., Krco, N., Marras, M., & Käser, T. (2022). Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs. *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*. https://doi.org/10.5281/ZENODO.6852964

Tappert, C. C. (2019). Who is the father of deep learning? *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 343–348. https://doi.org/10.1109/CSCI49370.2019.00067

Tyler, M., Liu, A., & Srinivasan, R. (2021). Behavioral phenotyping for predictive model equity and interpretability in STEM education. In I. Roll, D. McNamara, S. Sosnovsky, R. Luckin, & V. Dimitrova (Eds.), *Artificial Intelligence in Education* (pp. 361–366). Springer International Publishing. https://doi.org/10.1007/978-3-030-78270-2_64

Vultureanu-Albişi, A., & Bădică, C. (2021). Improving students' performance by interpretable explanations using ensemble tree-based approaches. *2021 IEEE 15th International Symposium on Applied Computational Intelligence and Informatics (SACI)*, 215–220. https://doi.org/10.1109/SACI51354.2021.9465558

Yang, S. J. H., Ogata, H., Matsui, T., & Chen, N.-S. (2021). Human-centered artificial intelligence in education: Seeing the invisible through the visible. *Computers and Education: Artificial Intelligence*, *2*, 100008. https://doi.org/10.1016/j.caeai.2021.100008

Zhu, J., Yu, W., Zheng, Z., Huang, C., Tang, Y., & Fung, G. P. C. (2020). Learning from interpretable analysis: Attention-based knowledge tracing. In I. I. Bittencourt, M. Cukurova, K. Muldner, R. Luckin, & E. Millán (Eds.), *Artificial Intelligence in Education* (pp. 364–368). Springer International Publishing. https://doi.org/10.1007/978-3-030-52240-7_66