Learner Modeling Interpretability and Explainability in Intelligent Adaptive Systems

Diego Zapata-Rivera[1] and Burcu Arslan[2]

Abstract. Learner models are used to support the implementation of personalization features in Adaptive Instructional Systems (AISs; e.g., adaptive sequencing of activities, adaptive feedback), which are important aspects of Intelligent Adaptive Systems. With the increased computational power, more advanced methodologies, and more available data, learner models include a variety of Artificial Intelligence techniques. These techniques have different levels of complexity, which influence interpretability and explainability of learner models. Interpretable and explainable learner models can facilitate appropriate use of the learner modeling information in AISs, their adoption, and scalability. This chapter elaborates on the definitions of interpretability and explainability, describes interpretability and explainability levels of different models, elaborates on the levels of explainability to produce needed information for teachers and learners, and discusses implications and future work in this aera.

1 Introduction

Learner models are representations of learners' knowledge, skills and other attributes that are used by Adaptive Instructional Systems (AIS) to support personalization. Personalization features may involve adaptive selection of activities and adaptive feedback. Learner models can include information about learners' cognitive and non-cognitive skills [28, 62]. Learner models can be produced based on information gathered before learner's interactions with the AIS and refined as more information about the learner is collected as learners interact with the AIS [62]. Learners and other educational stakeholders can have access to learner model information [15, 76]. Types of interaction with the learner model and purposes for accessing have been studied in the area of Open Learner Modeling (OLM) [77]. OLM approaches include the design and use of interfaces to facilitate interaction with learner model information. These interactions can include guidance mechanisms (e.g., interaction scripts, negotiation approaches) and guidance provided by pedagogical agents or other humans.

Communicating learning model information to teachers and learners requires knowing their needs for assessment information, and the evidence available to support learner model claims. A learner model that supports the generation of explanations for various types of end users can facilitate this process. Different types of learner modeling approaches (i.e., top-down, bottom-up, hybrid) exhibit different affordances and challenges for the generation of explanations for various types of end users [78].

Interpretability and explainability of learner models are key concerns in learning and assessment contexts, especially considering that making inferences from learning and assessment data is challenging due to different sources of noise in learning and assessment data [37]. For example, teachers and students may want to know how information maintained by the system is used to support student learning. Also, students may be interested in knowing more about how the system calculates their progress. Interpretability and explainability can facilitate the adoption of AISs since trust in these systems may

[1] dzapata@ets.org, Educational Testing Service, Princeton NJ 08541, USA
[2]Educational Testing Service Global, 1077 XX, Amsterdam, The Netherlands

increase as teachers and learners better understand why recommendations and decisions by the system are made [19, 37, 39, 71, 79]. For example, a learner model that supports interpretability and explainability can be deployed to justify the suggested adaptive sequencing of activities and/or the information presented in dashboards for teachers and learners as well as to explain the underlying mechanisms of pedagogical agents' interactions (e.g., triggering interactions based on the status of cognitive and non-cognitive skills).

In this chapter, we define interpretability and explainability with respect to the transparency of the models and the explanation generated to the end users (e.g., learners and teachers), classify the models in terms of their interpretability and explainability, and discuss explainability to teachers and learners. We conclude with discussing the implications and future directions.

2 Interpretability and Explainability

With the increased computational power, more advanced methodologies, and more available data, more Adaptive Instructional Systems (AISs) can now make use of an AI technique called Machine Learning (ML) to make predictions, decision-making, and personalization in addition to the symbolic, rule-based Artificial Intelligence (AI) techniques [17, 40, 42, 43, 50, 59, 63, 72]. ML algorithms can help with that creation of models learned from (big) data and use these models to support decision making by making predictions and identifying hidden relationships and patterns in the data. In general, creating ML models can be efficient in terms of human-labor and the relative high accuracy that ML models may have. However, ML models have a couple of high-risk drawbacks that require careful and thorough processes to ensure that their applications do not harm end users.

The first drawback is related to the quality of the data that are used to train ML models. Because the ML models are learning from the data, inaccurate, incomplete, or incompatible datasets (i.e., data biases) give rise to biased decisions and predictions [60]. Therefore, to lower the risks to the end users, it is important to: a) assess the quality of the data, b) collect data from diverse groups, and d) be transparent about the content and characteristics of training data [2, 37, 60].

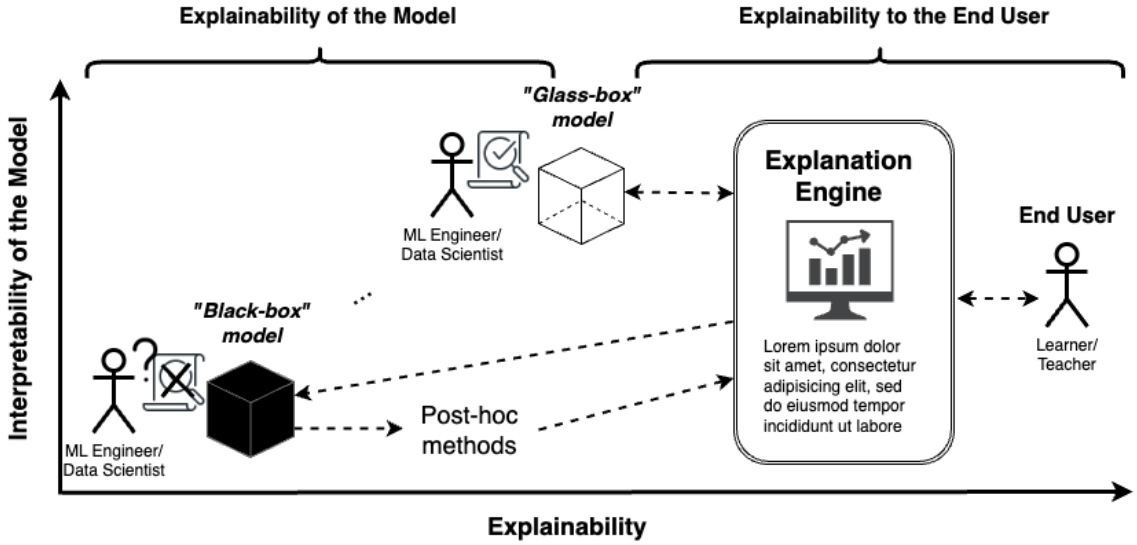
The second high-risk drawback is related to the complexity of the algorithms, which is the focus of this chapter. A group of ML models that leverages more complex ML algorithms such as deep neural networks or deep learning (DL) and large language models (LLMs). DL utilizes artificial neural networks, which are algorithms inspired by the structure and function of the human brain at a very high level. The input data is processed through multiple layers, where each layer extracts and amplifies specific features of the data. Different from other ML algorithms, DL can better handle unstructured data and can perform feature extraction automatically with minimal domain knowledge and human effort and with high predictive accuracy [4]. However, even if the complex ML models' prediction accuracy is high, these models have a potential to make their decisions based on the correlations between irrelevant features and the outcome variable (e.g., see [12] for an ML model that classifies husky *vs* wolf images based on the pixels related to snow rather than the feature of the canine). LLMs work by analyzing and processing vast amounts of text data and use DL to understand and generate human-like text predicting the most likely next word or phrase based on their training data without necessarily understanding the meaning of the text and without necessarily generating output based on facts (i.e., hallucinations) [32]. These models are trained on diverse datasets from books, websites, and other written materials to learn language patterns, grammar, and context. As a result, LLMs can perform tasks like having dialogs with humans, translating languages, summarizing texts, and creating content. LLMs are being employed in different educational contexts, such as creating (conversational) intelligent tutoring systems [17, 59], having personalized educational dialogs with students [66], classification of algebra errors [43]. Despite their advantages, to perform these complex tasks, DL and LLM models include hundreds to billions of parameters and involve complex computations, which makes it harder to *interpret* the models' inner decision-making process. The biases included in the training data combined with a lack of understanding how the AISs make their decisions may lead to unreliable, thus untrustworthy systems. Finally, there is work on the use of neural-symbolic approaches [11, 65, 73] aimed at leveraging the advantages and mitigating the disadvantages of both rule-based symbolic and sub-symbolic ML approaches.

Although the decisions are not as high stake as in the use case of AI in medical decision-making, in the context of education, it is important that learners and teachers have adequate and valid explanations about the AIS's decisions so that they: a) trust the system's decisions, b) have agency to take an appropriate action when they detect an inaccurate or biased prediction, decision, or recommendation (see

also [37]). The *interpretability* and *explainability* of AI systems are the central focus of the Explainable AI (XAI) field (for a historical perspective see [21]). These two terms are closely related and there is no consensus on how they are defined [1, 2, 6, 9]. Most ML researchers use these terms interchangeably to refer to the degree to which an AI algorithm's output can be understood by humans (e.g., [1, 38, 48]) although there are differences between these two terms as different psychological constructs from the perspective of cognitive science (see [13]).

In the scope of this chapter, we use *interpretability of a model* as a notion attached to the model's inner decision-making transparency in relation to its expert user (e.g., ML engineers, data scientists). If a model's inner decision-making processes are transparent in a way that experts (e.g., ML engineers, data scientists) can understand *how* the model works, the model's interpretability is considered as high (see Figure 1). These types of ML models are classified as inherently interpretable, "glass-box", or "white-box" models. On the other hand, if a model's inner decision-making processes are hard to comprehend by the experts because of their complexity, it is referred as a "black-box" model. The ML models that fall in between these two categories are called "grey-box" models [2, 3, 10].

Fig. 1. A figure depicting the notions of interpretability, explainability, and explanation in relation to the model transparency and end user*



*The three dots between "glass-box" and "black-box" models represent "grey-box" models. The bidirectional arrow between the end user and explanation depicts that end user can reject or modify the explanation depending on the context, which in turn feed into to the model.

On the other hand, we use *explainability* as a term that encompasses two notions: explainability of a model and explainability to the end user. *Explainability of a model*, which is closely related to the interpretability of a model, is a process to apply methods to understand *why* the ML models make their decisions [2]. While interpretability of a model is a more static construct, explainability of a model can be a more dynamic construct. For example, "glass-box" models do not necessarily require researchers to apply an additional method to understand *why* they make their decision since *how* they work is transparent. Thus, these types of models have both high model interpretability and explainability (see Figure 1). However, "black-box" models are initially have low interpretability and they require applying additional methods (i.e., post-hoc methods, see [48]) to be able to increase their model explainability [2, 48, 55].

Taking a human-centered approach, in addition to the explainability of a model, in this chapter, we discuss another notion under explainability, which is *explainability to the end user*. Unlike explainability of the model which is related to an ML model's inner workings and the methods to understand its decision-making processes, explainability to the end user is a process to generate *explanations* about the model's decision-making process through external representations (e.g., a graph) and/or natural language for end users who are experts or non-experts (e.g., teachers or learners) based on their needs so that they can comprehend the explanations and take action to ensure agency. These explanations can support human-machine interactions in AISs, better accountability to increase trust in these systems (see [37] for a comprehensive XAI-ED framework; see also [15] for an open learner

modeling framework). Explainability to the end user includes two-way interactions with the Explanation Engine in which end users (e.g., learner or teacher) can interact with it to elicit more information, provide context, or reject the explanation provided. With the help of the Explanation Engine, the information provided by the end user used as feedback to the model (see two-way arrow between the end user and the Explanation Engine and the arrows from the Explanation Engine to the model in Figure 1). Explanation Engine is also responsible for not only presenting explanations but also presenting them at the right time (see also [18] for lessons learned from ITSs for XAI). Moreover, it may allow end users to have an option to turn-on and turn-off the explanations based on their needs.

In the next section, we provide more information on the interpretability and explainability of different types of learner models in AISs.

3 Interpretability and Explainability of Different Types of Models

Historically, a variety of approaches have been implemented for modeling learners (see [79]). In addition to differences in variables chosen to depict the learner's knowledge and skills and the context in which they were applied, these approaches may include different types of models that have different levels of interpretability and explainability. We first present some types of models and their levels of interpretability (see Table 1). Subsequently, we provide a brief overview about the methods related to increasing the explainability of the "black-box" and "glass-box" models.

Table 1. Different types of models and their levels of interpretability

Model Types	Model Subtypes	Interpretability Level	Model Label
Deep Learning	Deep Knowledge Tracing (e.g., [23, 51]), Graph Neural Networks (e.g., [70]), Large Language Models (e.g., [50])	Low	"Black-box"
Machine Learning I (Ensemble Methods)	Random Forest Decision Tree (e.g., [72]), XGBoost (e.g., [64]), AdaBoost (e.g., [26])	Low	"Black-box"
Neural-Symbolic Learning	Knowledge Enhanced Graph Neural Networks (e.g., [52]), Temporal Learner Modeling (e.g., [31])	Medium	"Grey-box"
Machine Learning II (Fuzzy and Probabilistic Methods)	Fuzzy Logic (e.g., [27]), Bayesian Knowledge Tracing (e.g., [74]), Naïve Bayes (e.g., [44]), Bayesian Learner Models (e.g., [20, 53, 84]); Knowledge Spaces (e.g., [24, 25])	Medium	"Grey-box"
Machine Learning III	Linear and Logistic Regression (e.g., [67]), Generalized Additive Models (e.g., [22]), Decision Trees (e.g.,[75])	High	"Glass-box"
Symbolic	Cognitive modeling (e.g., [5, 7, 8]), Rule-based systems (e.g., [33]), Constrained-Based Learner Models* (e.g., [46])	High	"Glass-box"

* Constrained-Based Learner Models have other versions that can be classified as probabilistic or Deep Learning (e.g.,[47]).

As we mentioned above, interpretability and explainability of learner models is essential for teachers and learners to better understand why recommendations and decisions are made by the AIS. While rule-based, symbolic AI approaches make decisions in a transparent way, the level of human effort and content knowledge required to infuse knowledge into these models is extensive. On the other hand, although ML models can learn from the data without human involvement in the learning process, human

effort and expertise required is also extensive to develop ML algorithms, to provide labels for the training data in supervised ML models, and to make sure the ML models learn the correct representation (see [49] for human-in-the-loop ML).

For the "glass-box" ML models, the Explanation Engine can generate global explanations about how different features or variables contribute to model's decisions (see the row Machine Learning III in Table 1) without necessarily applying post hoc methods [37]. Therefore, the fidelity of explanations is considered high. Although these type of "glass-box" models are considered as inherently interpretable, when the number of features is high, it might get harder for humans to understand the decision process. On the other hand, the models that have low (i.e., "black-box") or medium (i.e., "grey-box") interpretability require applying additional methods to move on the explainability axis from low to medium or high (see Figure 1). These different types of methods generate different types of explanations (see [48] for a comprehensive list of methods). Model-agnostic global methods such as surrogate models generate explanations about how different features or variables affect model's overall behavior by learning another interpretable ML model to approximate the outcome of the "black-box" or "grey-box" model resulting in low fidelity explanations [2, 37, 48]. In contrast to model-agnostic global methods, model-agnostic local methods, such as LIME (Local Interpretable Model-agnostic Explanations [54]) or SHAP (SHapley Additive exPlanations [41]), focus on explaining individual predictions, which might be particularly useful in AISs where learner-level explanations are necessary and the AISs does not include a "glass-box" model. In addition to these types of methods, there are example-based explainability methods such as counterfactual explanations [68].

Although different explainability methods have been introduced to make the "black-box" models more explainable, it is important to emphasize that most of these approaches are based on an approximation of the model's behavior; thus, they do not offer high fidelity explanations as inherently interpretable models do, and should be used with caution (see also [56–58] for critiques of using post hoc methods to make explanations for "black-box" models).

In the next section, we describe the types of information needs of teachers and learners and elaborate on the explainability required to meet those needs.

4 Explainability to Teachers and Learners

Teachers and learners interacting with AISs have different types of assessment information needs. Learner models can provide the information needed to support learning and teaching processes. As we discussed above, Explainability Engines, for example, can be used to generate explanations required to respond to teachers' and learners' questions (see Figure 1). A variety of external representations can be used to provide users with responses to their questions. Researchers in the area of Open Learner Modeling (OLM) have explored various types of external representations such as graphical representations, interactive reports, dashboards, and the use of pedagogical agents that make use of learner model information to provide guidance to users in the exploration of learner models [15, 34, 80, 83]. Table 2 summarizes some of the most common assessment information needs of teachers and students and identifies the level of explainability required by the learner model to provide such information.

Table 2. Assessment information needs of teachers and learners and required level of model explainability *

End User	Assessment Information Needs	Required Level of Model Explainability
Teachers	*Student performance at the individual, sub-group, and class levels.* • What are my students' strengths and weaknesses? • How did the class perform on a task or a group of tasks? • How does a student's performance compare to that of other students? • Progress information at the individual, subgroup, and class levels • How much progress have my students made towards mastery?	Low-Medium
	Information that can help inform future teaching. • How difficult were the tasks for my students? • What were the most frequent errors and misconceptions?	Low-Medium
	Information that helps understand current performance. • Were my students engaged in the task(s)? • Did my students try to game the system? • How reliable are the knowledge and engagement estimates calculated by the system?	Medium-High
	Instructional recommendations. • What should I do next to help an individual student or the class as a whole?	Medium-High
Learners	*Actionable feedback that they can use to guide their learning.* • What are my strengths and weaknesses? • How can I improve?	Low-Medium
	Progress and performance information. • How much progress have I made towards mastery? • How does my performance compare to that of other students?	Low-Medium
	Evidence supporting assessment claims. • What type of information was used to calculate my knowledge levels? • Can I provide additional evidence to update my knowledge levels in the system?	Medium-High

*First two columns are adapted from "Supporting Human Inspection of Adaptive Instructional Systems", by [76].

Teacher questions related to student performance at the individual, sub-group, and class levels, such as "What are my students' strengths and weaknesses?", or related to information that can help inform future teaching, such as "How difficult were the tasks for my students?" may require low-medium levels of learner model explainability depending on whether the teacher is interested in digging deeper into the evidence used to answer these questions. Other questions related to information that helps teachers to understand current learner performance, such as "Were the students engaged in the task?" or "How reliable are the knowledge and engagement estimates calculated by the system?" may require learner models that support medium-high explainability levels since there could be a variety of aspects influencing these estimates. Questions that are related to instructional recommendations, such as "What should I do next to help an individual student or the class as a whole?" may require predictive models

that make use of evidence from various sources. Explanations generated using these types of predictive models may require additional user support to help users understand how the data are used to make predictions and the limitations of these models. Similarly, in the case of learners, questions related to receiving actionable feedback that learners can use to guide their learning, such as "What are my strengths and weaknesses?", or related to their progress and performance, such as "How much progress have I made towards mastery?" may require learner models that support low-medium explainability levels based on the amount of supporting evidence required by learners. However, questions related to evidence supporting assessment claims, such as "What type of information was used to calculate my knowledge levels?" or "Can I provide additional evidence to update my knowledge levels in the system?" require medium-high explainability levels since they require additional evidence and more sophisticated explanations.

Learner model explainability can benefit from a clear structure connecting claims to supporting evidence which may include process and response data. The implementation of an *evidence layer* can facilitate the generation of explanations through external representations, and interaction mechanisms that make use of learner model information to support learning and teaching [80]. The implementation of such evidence layer can be facilitated by using top-down and hybrid approaches that combine top-down and bottom-up approaches to designing learner models with different levels of interpretability and explainability (e.g., models that have been created by leveraging Evidence-Centered Design principles [45] together with several psychometric models and "big data" processes [82]). However, different techniques are explored to improve the explainability of "black-box" learner models resulting from the application of bottom-up approaches (e.g., Chain-of-Thought prompt engineering; [69]) and neural-symbolic or neuro-symbolic approaches for AI models; [29, 30]). These approaches may require a considerable amount of human effort to both creating the evidence managing mechanisms and validating the results produced by the model [61, 69]. This evidence layer can support the implementation of the Explanation Engine, which can offer explainability services in an instructional ecosystem, making a positive impact in terms of scalability. Finally, the evidence layer could be conceptually placed to the right of the Explanation Engine, between the output of the models feed and the Explanation Engine in Figure 1.

5 Conclusions and Discussion

As new advances in ML become available and applications of these technologies extend, it is important to emphasize the need for interpretable and explainable models in education settings. Below we conclude by discussing the implications of improving interpretability and explainability of learner models in the context of AISs.

- An appropriate level of learner model interpretability and explainability is required to support trust and adoption of AISs. Understanding how AISs support teaching and learning is an important first step in making sure that teachers' and learners' expectations are met. A general understanding of how adaptive components of the AIS are implemented and how they are intended to support teaching and learning may have a positive effect in adoption of these systems. Different levels of interaction with learner models should be supported to answer different types of user questions [35, 36]. These levels of interaction require models that support the generation of appropriate explanations.
- The amount of human effort required to create explainable learner models that can respond to the needs for information of educational stakeholders can vary. The amount of data required to support learner model claims and the mechanisms for evidence identification and aggregation can also vary depending on the type of learner modeling approach used [79] and the data available to create those models. We expect that as new advances in AI become available, learner models will become more useful in supporting human decision making. Privacy, data security, and evaluation of learner models in supporting appropriate decision making will continue to be areas of interest.

Modeling approaches should support the generation of explanations that consider various levels of uncertainty associated with different types of evidence sources and the nature of evidence aggregation and accumulation processes. AISs should consider maintaining different views of the learner model to capture teachers' and learners' perspectives. These perspectives can contribute to interesting negotiation

and reflection processes that can have positive instructional value (e.g., knowledge awareness, self-reflection and self-regulation [14, 15]). In fact, human-in-the-loop approaches can reduce diagnostic complexity and provide immediate confirmation when levels of uncertainty are high. Teachers value flexibility when interacting with AISs. They appreciate the system handling common cases but be alerted on particular cases that may require their attention, so they have the opportunity to override suggestions made by the AIS based on additional information about the learner and the learning context that they may have [16, 81].

6 Future work

Future work involves continue advancing in the development and evaluation of modeling approaches that support appropriate use of learner modeling information. Improvements in interpretability and explainability of these models contributes to achieving this goal. As more data (e.g., multimodal data) and AI technologies to create innovative learner models become available, additional opportunities for personalization in education contexts will arise (e.g., through the use of AISs). It is paramount that AI systems are designed taking into account the need for user understanding of the benefits and limitations of these technologies. We expect that additional work will be done in areas such as human-centered AI, data privacy, and data security to support the responsible use of AI.

Acknowledgements

This material is based upon work supported by the National Science Foundation and the Institute of Education Sciences under Grant #2229612. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the U.S. Department of Education.

References

1. Adadi A, Berrada M (2018) Peeking inside the black-box: A survey on explainable Artificial Intelligence (XAI). IEEE Access 6:52138–52160

2. Ali S, Abuhmed T, El-Sappagh S, Muhammad K, Alonso-Moral JM, Confalonieri R, Herrera F (2023) Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion 99:101805

3. Alonso JM, Castiello C, Mencar C (2015) Interpretability of fuzzy systems: Current research trends and prospects. Springer Handbook of Computational Intelligence

4. Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, Santamaría J, Fadhel MA, Al-Amidie M, Farhan L (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. Journal of Big Data 8:53. doi: 10.1186/s40537-021-00444-8

5. Anderson JR (2005) Human symbol manipulation within an integrated cognitive architecture. Cognitive Science 29:313–341

6. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, Herrera F (2020) Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion 58:82–115

7. Arslan B, Taatgen NA, Verbrugge R (2017) Five-year-olds' systematic errors in second-order false belief tasks are due to first-order theory of mind strategy selection: A computational modeling study. Frontiers in Psychology 8. doi: 10.3389/fpsyg.2017.00275

8. Arslan B, Verbrugge R, Taatgen N (2017) Cognitive control explains the mutual transfer between dimensional change card sorting and first-order false belief understanding: A computational modeling study on transfer of skills. Biologically Inspired Cognitive Architectures 20:10–20. doi: 10.1016/j.bica.2017.03.001

9. Benchekroun O, Rahimi A, Zhang Q, Kodliuk T (2020) The need for standardized explainability. arXiv:201011273

10. Bennetot A, Franchi G, Del Ser J, Chatila R, Diaz-Rodriguez N (2022) Greybox XAI: A Neural-Symbolic learning framework to produce interpretable predictions for image classification. Knowledge-Based Systems 258:109947

11. Besold TR, Kühnberger KU (2015) Towards integrated neural–symbolic systems for human-level AI: Two research programs helping to bridge the gaps. Biologically Inspired Cognitive Architectures 14:97–110

12. Besse P, Castets-Renard C, Garivier A, Loubes JM (2019) Can everyday AI be ethical? Machine Learning algorithm fairness. Statistiques et Société 6

13. Broniatowski DA (2021) Psychological foundations of explainability and interpretability in artificial intelligence

14. Bull S (2020) There are open learner models about! IEEE Transactions on Learning Technologies 13:425–448

15. Bull S, Kay J (2016) SMILI☺: A framework for interfaces to learning data in open learner models, learning analytics and related fields. International Journal of Artificial Intelligence in Education 26:293–331

16. Cardona MA, Rodríguez RJ, Ishmael K (2023) Artificial Intelligence and Future of Teaching and Learning: Insights and Recommendations. US Department of Education, Office of Educational Technology

17. Chen Y, Ding N, Zheng HT, Liu Z, Sun M, Zhou B (2023) Empowering Private Tutoring by Chaining Large Language Models. arXiv preprint arXiv:230908112

18. Clancey WJ, Hoffman RR (2021) Methods and standards for research on explainable artificial intelligence: Lessons from intelligent tutoring systems. Applied AI Letters 2:53

19. Conati C, Barral O, Putnam V, Rieger L (2021) Toward personalized XAI: A case study in intelligent tutoring systems. Artificial Intelligence 298:10350

20. Conati C, Gertner A, Vanlehn K (2002) Using Bayesian networks to manage uncertainty in student modeling. User Modeling and User-Adapted Interaction 12:371–417

21. Confalonieri R, Coba L, Wagner B, Besold TR (2021) A historical perspective of explainable Artificial Intelligence. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 11:1391

22. Dikaya LA, Avanesian G, Dikiy IS, Kirik VA, Egorova VA (2021) How personality traits are related to the attitudes toward forced remote learning during Covid-19: Predictive analysis using generalized additive modeling. Frontiers in Education 6:108

23. Ding X, Larson EC (2021) On the interpretability of deep learning based models for knowledge tracing. arXiv preprint arXiv:210111335

24. Falmagne JC, Albert D, Doble C, Eppstein D (2013) Knowledge Spaces: Applications in Education. Springer Science & Business Media

25. Falmagne JC, Koppen M, Villano M, Doignon JP, Johannesen L (1990) Introduction to knowledge spaces: How to build, test, and search them. Psychological Review 97:201

26. Forbes-Riley K, Litman D (2004) Predicting emotion in spoken dialogue from multiple knowledge sources. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004. pp 201–208

27. Gagan G, Lalle S, Luengo V (2012) Fuzzy Logic Representation for Student Modelling. In: ITS 2012-11th International Conference on Intelligent Tutoring Systems-Co-adaptation in Learning. Springer-Heidelberg, pp 428–433

28. Greer J, McCalla G (1994) Student Models: The Key to Individualized Educational Systems. Springer Verlag, New York, NY

29. Hammond K, Leake D (2023) Large language models need symbolic AI. In: Proceedings of the 17th International Workshop on Neural-Symbolic Reasoning and Learning, CEUR Workshop Proceedings, Siena, Italy. pp 3–5

30. Hitzler P, Eberhart A, Ebrahimi M, Sarker MK, Zhou L (2022) Neuro-symbolic approaches in artificial intelligence. National Science Review 9:035

31. Hooshyar D (2023) Temporal learner modelling through integration of neural and symbolic architectures. Education and Information Technologies. doi: 10.1007/s10639-023-12334-y

32. Huang L, Yu W, Ma W, Zhong W, Feng Z, Wang H, Liu T (2023) A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions

33. Jaques PA, Seffrin H, Rubi G, Morais F, Ghilardi C, Bittencourt II, Isotani S (2013) Rule-based expert systems to support step-by-step guidance in algebraic problem solving: The case of the tutor PAT2Math. Expert Systems with Applications 40:5456–5465

34. Kay J (2021) Scrutability, Control and Learner Models: Foundations for Learner-Centered Design in AIED. In: Roll I, McNamara D, Sosnovsky S, Luckin R, Dimitrova V (eds) Artificial Intelligence in Education. AIED 2021. Lecture Notes in Computer Science. Springer, Cham

35. Kay J, Kummerfeld B, Conati C, Porayska-Pomsta K, Holstein K (2023) Scrutable AIED. In: Handbook of Artificial Intelligence in Education. p 101

36. Kay J, Zapata-Rivera D, Conati C (2020) The GIFT of Scrutable Learner Models: Why and How. In: M. RA, Sinatra AC, Graesser X, Hu B, Goldberg, J. A, Hampton (eds) – Data Visualization. U.S. Army CCDC - Soldier Center, Orlando, FL, pp 25–40

37. Khosravi H, Shum SB, Chen G, Conati C, Tsai YS, Kay J, Gašević D (2022) Explainable artificial intelligence in education. Computers and Education: Artificial Intelligence 3:100074

38. Koh PW, Liang P (2017) Understanding black-box predictions via influence functions. In: International conference on machine learning. PMLR, pp 1885–1894

39. Leichtmann B, Humer C, Hinterreiter A, Streit M, Mara M (2023) Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. Computers in Human Behavior 139:107539

40. Lin CC, Huang AYQ, Lu OHT (2023) Artificial intelligence in intelligent tutoring systems toward sustainable education: a systematic review. Smart Learn Environ 10:41. doi: 10.1186/s40561-023-00260-y.

41. Lundberg SM, Lee SI (2017) A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems 30

42. McNamara DS, Arner T, Butterfuss R, Fang Y, Watanabe M, Newton N, Roscoe RD (2023) iSTART: Adaptive comprehension strategy training and stealth literacy assessment. International Journal of Human–Computer Interaction 39:2239–2252

43. McNichols H, Zhang M, Lan A (2023) Algebra Error Classification with Large Language Models. In: International Conference on Artificial Intelligence in Education. Springer Nature Switzerland, Cham, pp 365–376

44. McQuiggan SW, Mott BW, Lester JC (2008) Modeling self-efficacy in intelligent tutoring systems: An inductive approach. User modeling and user-adapted interaction 18:81–123

45. Mislevy RJ, Almond RG, Lukas JF (2003) A brief introduction to evidence-centered design. ETS Research Report Series 2003:1–29

46. Mitrovic A, Martin B, Suraweera P (2007) Intelligent tutors for all: Constraint-based modeling methodology, systems and authoring. IEEE Intelligent Systems 22:38–45

47. Mitrovic A, Ohlsson S (2016) Implementing CBM: SQL-Tutor after fifteen years. International Journal of Artificial Intelligence in Education 26:150–159

48. Molnar C (2022) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable, 2nd, ed

49. Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D, Bobes-Bascarán J, Fernández-Leal Á (2023) Human-in-the-loop machine learning: a state of the art. Artificial Intelligence Review 56:3005–3054. doi: 10.1007/s10462-022-10246-w

50. Ouyang F, Wu M, Zheng L, Zhang L, Jiao P (2023) Integration of artificial intelligence performance prediction and learning analytics to improve student learning in online engineering course. International Journal of Educational Technology in Higher Education 20:4

51. Piech C, Spencer J, Huang J, Ganguli S, Sahami M, Guibas L, Sohl-Dickstein J (2015) Deep knowledge tracing. arXiv preprint arXiv:150605908

52. Raj K (2023) A Neuro-symbolic Approach to Enhance Interpretability of Graph Neural Network through the Integration of External Knowledge. In: Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. pp 5177–5180

53. Reye J (2004) Student modelling based on belief networks. International Journal of Artificial Intelligence in Education 14:63–96

54. Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd SIGKDD International Conference on Knowledge Discovery and Data Mining. pp 1135–1144

55. Rizzo M, Veneri A, Albarelli A, Lucchese C, Conati C (2023) A theoretical framework for AI models explainability with application in biomedicine. In: IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB. pp 1–9

56. Rosé CP, McLaughlin EA, Liu R, Koedinger KR (2019) Explanatory learner models: Why machine learning (alone) is not the answer. British Journal of Educational Technology 50:2943–2958

57. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1:206–215

58. Rudin C, Radin J (2019) Why are we using black box models in AI when we don't need to? A lesson from an explainable AI competition. Harvard Data Science Review 1:1–9

59. Schmucker R, Xia M, Azaria A, Mitchell T (2023) Ruffle&Riley: Towards the Automated Induction of Conversational Tutoring Systems. arXiv preprint arXiv:231001420

60. Schramowski P, Turan C, Andersen N, Rothkopf CA, Kersting K (2022) Large pre-trained language models contain human-like biases of what is right and wrong to do. Nature Machine Intelligence 4:258–268

61. Shum K, Diao S, Zhang T (2023) Automatic Prompt Augmentation and Selection with Chain-of-Thought from Labeled Data. arXiv preprint arXiv:230212822. doi: http://arxiv.org/abs/2302.12822

62. Shute VJ, Zapata-Rivera D (2012) Adaptive educational systems. In: Durlach P (ed) Adaptive Technologies for Training and Education. Cambridge University Press, New York, pp 7–27

63. Singh N, Gunjan VK, Mishra AK, Mishra RK, Nawaz N (2022) Seistutor: A custom-tailored intelligent tutoring system and sustainable education. Sustainability (Switzerland) 14:4167

64. Su W, Jiang F, Shi C, Wu D, Liu L, Li S, Shi J (2023) An XGBoost-Based Knowledge Tracing Model. International Journal of Computational Intelligence Systems 16:13

65. Sun R, Bookman LA (1994) Computational Architectures Integrating Neural and Symbolic Processes: A Perspective on the State of the Art. Kluwer, Norwell, MA

66. Tack A, Piech C (2022) The AI teacher test: Measuring the pedagogical ability of blender and GPT-3 in educational dialogues. arXiv preprint arXiv:220507540

67. Vaessen BE, Prins FJ, Jeuring J (2014) University students' achievement goals and help-seeking strategies in an intelligent tutoring system. Computers & Education 72:196–208

68. Wachter S, Mittelstadt B, Russell C (2017) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv JL & Tech 31:841

69. Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, Chi E, Le Q, Zhou D (2022) Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems 35:24824–24837

70. Xia Z, Dong N, Wu J, Ma C (2023) Multi-Variate Knowledge Tracking Based on Graph Neural Network in ASSISTments. IEEE Transactions on Learning Technologies

71. Xu W (2019) Toward human-centered AI: a perspective from human-computer interaction. interactions 26:42–46

72. Yang C, Chiang FK, Cheng Q, Ji J (2021) Machine learning-based student modeling methodology for intelligent tutoring systems. Journal of Educational Computing Research 59:1015–1035

73. Yu D, Yang B, Liu D, Wang H, Pan S (2023) A survey on neural-symbolic learning systems. Neural Networks

74. Yudelson MV, Koedinger KR, Gordon GJ (2013) Individualized Bayesian Knowledge Tracing models. In: Artificial Intelligence in Education: 16th International Conference, AIED 2013. Springer, Memphis, TN, USA, pp 171–180

75. Zacharis NZ (2018) Classification and regression trees (CART) for predictive modeling in blended learning. IJ Intelligent Systems and Applications 3:9

76. Zapata-Rivera D (2019) Supporting human inspection of adaptive instructional systems. In: Adaptive Instructional Systems: First International Conference, AIS 2019, Held as Part of the 21st HCI International Conference, HCII 2019. Springer International Publishing, Orlando, FL, USA, pp 482–490

77. Zapata-Rivera D (2020) Open student modeling research and its connections to educational assessment. Int J Artif Intell Educ. doi: 10.1007/s40593-020-00206-2

78. Zapata-Rivera D, Arslan B (2021) Enhancing personalization by integrating top-down and bottom-up Approaches to Learner Modeling. In: R. S, J S (eds) Adaptive Instructional Systems. Adaptation Strategies and Methods. HCII 2021. Lecture Notes in Computer Science. Springer, Cham, pp 234–246

79. Zapata-Rivera D, Arslan B (2021) Enhancing Personalization by Integrating Top-Down and Bottom-Up Approaches to Learner Modeling BT - Adaptive Instructional Systems. Adaptation Strategies and Methods. In: Sottilare RA, Schwarz J (eds). Springer International Publishing, Cham, pp 234–246

80. Zapata-Rivera D, Brawner K, Jackson GT, Katz IR (2017) Reusing Evidence in Assessment and Intelligent Tutors. In: Sottilare R, Graesser A, Hu X, Goodwin G (eds) - Assessment Methods. U.S. Army Research Laboratory, Orlando, FL, pp 125–136

81. Zapata-Rivera D, Hansen EG, Shute VJ, Underwood JS, Bauer MI (2007) Evidence-based approach to interacting with open student models. International Journal of Artificial Intelligence in Education 17:273–303

82. Zapata-Rivera D, Liu L, Chen L, Hao J, Davier A (2016) Assessing Science Inquiry Skills in Immersive, Conversation-based Systems. In: Daniel BK (ed) Big Data and Learning Analytics in Higher Education. Springer International Publishing, pp 237-252,

83. Zapata-Rivera JD, Greer J (2002) Exploring Various Guidance Mechanisms to Support Interaction with Inspectable Learner Models. In: Proceedings of Intelligent Tutoring Systems ITS 2002. pp 442–452

84. Zapata-Rivera JD, Greer JE (2004) Interacting with inspectable Bayesian student models. International Journal of Artificial Intelligence in Education 14:127–163